Recovering Population Parameters from a Single Gene Genealogy: An Unbiased Estimator of the Growth Rate

Yosef E. Maruvka,*,[1] Nadav M. Shnerb,[1] Yaneer Bar-Yam,[2] and John Wakeley[3]

[1]Department of Physics, Bar-Ilan University, Ramat Gan, Israel
[2]New England Complex Systems Institute
[3]Department of Organismic & Evolutionary Biology, Harvard University

***Corresponding author:** E-mail: yosi.maruvka@gmail.com.

Associate editor: Rasmus Nielsen

Abstract

We show that the number of lineages ancestral to a sample, as a function of time back into the past, which we call the number of lineages as a function of time (NLFT), is a nearly deterministic property of large-sample gene genealogies. We obtain analytic expressions for the NLFT for both constant-sized and exponentially growing populations. The low level of stochastic variation associated with the NLFT of a large sample suggests using the NLFT to make estimates of population parameters. Based on this, we develop a new computational method of inferring the size and growth rate of a population from a large sample of DNA sequences at a single locus. We apply our method first to a sample of 1,212 mitochondrial DNA (mtDNA) sequences from China, confirming a pattern of recent population growth previously identified using other techniques, but with much smaller confidence intervals for past population sizes due to the low variation of the NLFT. We further analyze a set of 63 mtDNA sequences from blue whales (BWs), concluding that the population grew in the past. This calls for reevaluation of previous studies that were based on the assumption that the BW population was fixed.

Key words: coalescent, human population growth, blue whale population, large-sample theory, mitochondrial DNA.

Introduction

Because mutations accumulate over time along genetic lineages within a population, genetic variation in a sample taken today contains a record of past processes and events. We can therefore make inferences about the past from samples of DNA sequences or other genetic data. The current astounding pace of improvement in methods of DNA sequencing and genotyping offers an unprecedented opportunity to achieve the long-standing goal of population genetics, which is to quantify the forces that shape variation within the genomes of humans and other species. The technologies of population-genetic inference have also undergone amazing growth recently, especially in the area of computation, but they cannot be said to have kept pace with methods of sequencing and genotyping. The impending availability of immense data sets—such as those from the 1000 Genomes Project (http://www.1000genomes.org) and even more ambitious ventures which will no doubt be undertaken—provides strong motivation to develop inference methods for data sets that are very large both in the number of individuals sampled and in the number of base pairs sequenced.

We contribute to this endeavor by developing a new method of estimating demographic parameters, especially two of the principal quantities that determine the fate of a population: the effective population size and the population growth rate. Based on the large-sample asymptotic properties of gene genealogies, we show how the number of lineages ancestral to a sample depends on these two quantities. In striking contrast to the properties of small-sample gene genealogies, which are subject to high levels of stochastic variation, we show that the number of ancestral lineages declines nearly deterministically as a function of time in the past if the sample size is large. Our method of estimation is straightforward to implement and relatively cheap computationally. It provides point estimates, which simulations suggest are unbiased, together with confidence intervals obtained by a parametric bootstrap approach. Although it is a computational method, it is based on a coalescent analysis that illustrates how the availability of large samples can alleviate the formidable computational burden of coalescent-based inference.

A coalescent analysis involves modeling the genetic ancestry of a sample back to its most recent common ancestor (MRCA). In all, this ancestry is called the gene genealogy. All members of a sample share the same gene genealogy, and the resulting nonindependence makes population-genetic inference particularly difficult. In order to make quantitative estimates of the processes affecting a population, using genetic data, it is necessary to account for the fact that the true gene genealogy is unknown (or "missing data"). This requires a statistical model for sampling gene genealogies. Most current methods of inference are based on the standard neutral coalescent (Kingman 1982a,b; Hudson 1983; Tajima 1983), as this model has proven surprisingly robust to deviations from its initial assumptions (Möhle 1998; Nordborg and Krone 2002).

© The Author 2010. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Mol. Biol. Evol. 28(5):1617–1631. 2011 doi:10.1093/molbev/msq331 Advance Access publication December 16, 2010

The standard neutral coalescent is a model for a small sample from a large population whose size is constant over time. It is a stochastic process that generates a random-joining tree for the gene genealogy, together with a series of random coalescence times, one for each node of the tree. Time is usually rescaled by twice the "effective" population size. (This is the robustness of the coalescent: It can be applied to many different kinds of populations provided one replaces the actual size of a population with its effective size; Nordborg and Krone 2002; Sjödin et al. 2005.) After this rescaling, neutral mutations occur with rate $\theta/2$ along each branch of the tree, where θ is proportional to the product of the effective population size and the rate of neutral mutation per generation. See Hein et al. (2005) and Wakeley (2008) for reviews of the basic model as well as its extensions to include population structure, changes in population size over time, and selection.

The most exact methods of population genetic inference are based on the likelihood of a full data set, with the addition of prior distributions of parameters in the case of Bayesian methods. The likelihood is the probability of the data under the model (e.g., the standard neutral coalescent) with specific values of parameters (e.g., θ). Formally, the likelihood might be computed by averaging over the unknown gene genealogy of the sample, considering every possible gene genealogy in proportion to its probability under the coalescent model. In practice, this is achieved by Monte Carlo integration—employing simulations to sample a large number of randomly generated gene genealogies under the coalescent model—using a variety of different techniques. Canonical references to these methods are Griffiths and Tavaré (1994a,b) and Kuhner et al. (1995), with reviews by Stephens (2001), Tavaré (2004), and Felsenstein (2007).

The task of averaging over gene genealogies to compute likelihoods represents a serious challenge due to the enormity of the space of gene genealogies and the fact that only a miniscule fraction of them contribute significantly to the likelihood of any particular data set. Further, the computational burden increases explosively with the sample size, as the data become more and more complicated. In response to this, a number of approximate methods have been proposed in which inferences are based on manageably small sets of "summary statistics" extracted from the data (Fu and Li 1997; Tavaré et al. 1997; Weiss and von Haeseler 1998; Pritchard et al. 1999; Beaumont et al. 2002; Leman et al. 2005; Becquet and Przeworski 2007). The success of summary-statistic methods depends on reducing the dimensionality of the data, so that a greater fraction of gene genealogies can contribute to the computation, while preserving the information in the data relevant to the population parameters of interest.

The approach we take here shares some key features with summary statistic methods. Our inferences are based on information extracted from the data rather than on the full data set itself, and our method involves the simulation of gene genealogies. It differs slightly in concept because our summary statistics are indirect estimates of the properties of the gene genealogy instead of direct summaries of the data. However, our approach is primarily distinguished by the fact that it capitalizes on the special properties of gene genealogies of large samples. We demonstrate this with novel analytical results for both stable fixed size and exponentially growing populations. Specifically, we show that the backward-time dynamics of one feature of the tree—the number of lineages as a function of time or NLFT—is essentially deterministic when the number of lineages is large. We obtain expressions for the NLFT that are simple and accurate. The nearly deterministic behavior of the NLFT means that a relatively small sample of gene genealogies may be taken as representative of all gene genealogies.

Watterson (1975) initiated the study of large samples, but subsequent analyses aimed at understanding genetic variation or improving methods of inference have been few (Wakeley and Takahashi 2003; Rauch and Bar-Yam 2005). Our analytical work builds on the analysis in Rauch and Bar-Yam (2005) in which a rescaled version of the NLFT was studied under essentially the same assumptions we make here. In particular, we assume that $1 \ll n_0 \ll N$, where n_0 is the sample size and N is the population size in an idealized well-mixed population model such as the Wright–Fisher model (Fisher 1930; Wright 1931). However, simulations and a heuristic analysis show that our analytical expressions for the NLFT are accurate even when the entire population is sampled ($n_0 = N$) and much deeper into the past than might be expected.

Our work is also similar in spirit to lineages-through-time methods (Nee et al. 1995) and analyses (Stadler 2008), which have been used to study species diversity (Baldwin and Sanderson 1998; Moreau et al. 2006; Bininda-Emonds et al. 2007). The fact that the NLFT contains information about the effective size of a population over time is the basis for the several skyline plot methods that are available (Pybus et al. 2000; Strimmer and Pybus 2001; Drummond et al. 2005; Minin et al. 2008). Briefly, those methods allow population size to change over time with few restrictions and are geared toward relatively small samples in that they involve extensive computations (Drummond et al. 2005; Minin et al. 2008). In contrast, our method estimates the parameters of a specified model of a population and is less computationally intensive because it makes use of the nearly deterministic behavior of the NLFT for large samples. Thus, these approaches are complementary. In the sections Human Growth Rates and Blue Whale Population, we present applications in which these two approaches give similar answers concerning the effective population size over time.

In addition to the novel analysis and the fact that many previous methods may not scale up easily to large data sets, the method we present here is valuable because it provides unbiased estimates of the population growth rate as well as the population size. We demonstrate this using simulations. Previous methods, regardless of whether they use maximum likelihood (Kuhner et al. 1998) or Bayesian techniques (Kuhner and Smith 2007) or the skyline plot (Pybus et al. 2000), yield unbiased estimates of the population size but produce biased estimates of the growth rate.

Methods and Results

Model

We assume a haploid Wright–Fisher model of reproduction (Fisher 1930; Wright 1931), with either constant population size or exponential growth over time. The current population size is N_0. Following Slatkin and Hudson (1991) and Kuhner et al. (1998), we model exponential growth using $N(t) = N_0\,e^{-\gamma t}$ to give the population size in generation t in the past. At the present, time 0, a subsample of the whole population $n_0 \leqslant N_0$ (or simply n and N for a population of constant size) is chosen at random without replacement. We study the gene genealogy of the sample at a single locus under the assumptions that all variation is selectively neutral and there is no intralocus recombination. We use μ to denote the mutation rate per "locus" per generation (and not the common notation of mutation per site) and L to denote the length of the locus, that is, the number of sites sequenced, which can be finite for the finite site model or infinite for the infinite site model.

Simulation Procedures

When necessary, we used simulations to produce pseudo data sets. We did this by first generating the gene genealogy for a sample according to the haploid Wright–Fisher sampling process backward in time. That is, in every generation, the parent of each (haploid) individual is chosen uniformly at random from the whole population as it existed in the previous generation. Note that in growing populations, the previous generation may be smaller than the current one. If two or more individuals have the same parent, they coalesce and the number of ancestral lineages decreases accordingly. This procedure is continued until only a single lineage remains, that is, the MRCA of the sample.

Once the gene genealogy is obtained in this way, mutations are placed randomly on each branch of the tree, starting at the root, or MRCA. For purposes of illustrating the general behavior of the NLFT and of our inference method under finite sequence lengths, we used the symmetric, four-state Jukes–Cantor model of mutation (Jukes and Cantor 1969). Under this model, each branch receives a binomially distributed number of mutations, with a number of trials equal to the length of the branch in generations and a probability of success equal to the per-generation mutation rate, μ. Mutations occur at each site in the locus uniformly at random, and all the descendents of the lineage on which a mutation occurs inherit that mutation (possibly obscured by subsequent mutations). In some of the simulations and in the application to mitochondrial DNA (mtDNA), we used the more appropriate F84+Γ mutation model, which differs from the previous model in that instead of assuming symmetric mutations, it allows for transition bias, and it also allows for rate variation among sites (Felsenstein 2004).

In our examination of the NLFT for data from simulated gene genealogies, we considered three possible ways in which the NLFT might be obtained. In the first case, we obtained the true NLFT directly from the simulated gene genealogy. In the other two cases, we obtained the NLFT indirectly from a matrix of genetic differences between every pair of sequences. In the second case, we assumed $L = \infty$ so that the data conform to the infinite site model of Watterson (1975). For infinite sequence length, we generated a matrix of pairwise distances between sampled sequences by counting the number of mutations separating the two individuals on the gene genealogy. In the third case, particular values of L were assumed and the distance between two sequences was chosen to be the Hamming distance between them for the case of the Jukes–Cantor mutation model, and for the F84+Γ model, the Kimura two-parameter distance was used. Note that also the Hamming distance is not identical to a simple counting of the number of mutations due to the possibility of recurrent mutations.

For both infinite and finite sequence lengths, we obtained the NLFT by reconstructing a rooted ultrametric tree of the sampled sequences, meaning a tree in which the branch length from every leaf (or sample) to the root (or MRCA) is the same (Felsenstein 2004). We used the simple weighted pair group method or WPGM algorithm (Sokal and Michener 1958; Sneath and Sokal 1973). We tried other methods as well, including UPGMA and UPGMC, to verify that these could also be used in our estimation routine (see Simulation-Based Inference Method below) but found no compelling reason to prefer these over WPGM. For a given mutation rate μ, we rescaled the genetic distances by multiplying by the average time to one mutation event, that is, such that $g_{i,j} = d_{i,j}/\mu$ is the estimated number of generations separating sequence i and sequence j given that $d_{i,j}$ is the number of mutations between the two or an estimate of this number.

The NLFT in Populations of Constant Size

We use $n(t)$ to denote the number of ancestral lineages at generation t in the past (i.e., the NLFT) given a present-day sample of size $n_0 \equiv n(0)$. We begin by considering populations of constant size. Figure 1 shows two gene genealogies, simulated using the method described above. It can be seen that for large $n(t)$ and N, gene genealogies may differ from each other microscopically, yet plotting the number of lineages versus time produces nearly identical rather smooth curves. Differences become apparent when $n(t)$ is small, which occurs when t is large, that is, of order N. This suggests that if we examine only this feature of a tree, we can ignore its specific topology. Different trees will behave the same as long as their demographic parameters (n_0 and N) are the same. Thus, figure 1 indicates that fluctuations in $n(t)$ among different realizations of the gene genealogy are weak, so that its behavior appears largely deterministic.

Analytic expressions for $n(t)$ can be obtained using the well-known results of occupancy distributions (David and Barton 1962; Johnson and Kotz 1977). In particular, $n(t+1)$ under the Wright–Fisher model may be viewed as the result of tossing $n(t)$ "balls" randomly into N "boxes" such that each ball has chance $1/N$ of landing in any particular box. Then, $n(t+1)$ is the number of boxes that contain at least one ball. Watterson (1975) used this formulation in his

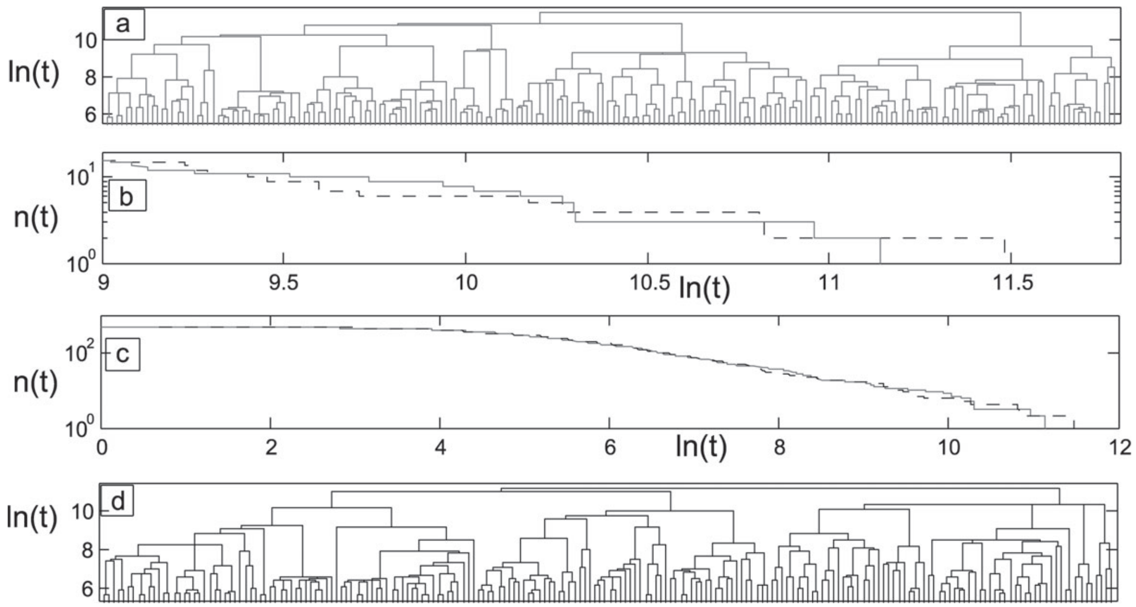


FIG. 1. Panels a and d show two independent gene genealogies, simulated using the procedure described in Methods and Results. In both cases, $N = 5 \times 10^4$ and $n_0 = 500$. Time is measured in generations and is plotted on a log scale for ease of presentation; we also show only the upper 200 nodes of each tree. Clearly, the trees are different but panels b and c show that, with respect to the NLFT, most of the difference is attributable to only the last 10 or 20 coalescent events. Panel c plots the NLFT for each (entire) gene genealogy and panel b redisplays same data but only for the very top portion of each tree. The solid lines in b and c correspond to the $n(t)$ of the tree in a, and the dashed lines in b and c correspond to the $n(t)$ of the tree in d.

pioneering work, making particular use of the expressions

$$E\left[n(t+1)|n(t)=i\right] = N - N\left(1-\frac{1}{N}\right)^{i} \qquad (1)$$

and

$$\mathrm{Var}\left[n(t+1)|n(t)=i\right] = N\left(1-\frac{1}{N}\right)^{i} + N(N-1)\left(1-\frac{2}{N}\right)^{i} - N^2\left(1-\frac{1}{N}\right)^{2i}. \qquad (2)$$

These are exact formulas, which hold for any admissible, that is, positive, values of n_0 and N, up to the strong sampling limit ($n_0 = N$), and even beyond ($n_0 > N$), as may be true for growing populations.

Watterson (1975) considered the case where $n(t)$ is of order N and noted the nearly deterministic behavior of the normalized variable $n(t+1)/N$ given $n(t) = i$ in the limit $N \to \infty$. When $n(t) = O(N)$, we have

$$\begin{aligned} E\left[n(t+1)|n(t)=i\right] &= i - \frac{i(i-1)}{2N} + \frac{i(i-1)(i-2)}{6N^2} - \cdots \qquad (3) \\ &= i - \frac{i^2}{2N}\left(1 + O\left(\frac{i}{N}\right)\right), \qquad (4) \end{aligned}$$

and following a similar expansion for the variance, we have

$$\mathrm{Var}\left[n(t+1)|n(t)=i\right] = \frac{i^2}{2N}\left(1 + O\left(\frac{i}{N}\right)\right). \qquad (5)$$

Rauch and Bar-Yam (2005) studied $p(t) \equiv n(t)/N$ for the case $p(0) = 1$ and obtained an expression for $p(t)$ using a differential equation that is valid when $p(t)$ is small.

Here, based on equations (3) through (5), we seek a deterministic approximation for $n(t)$. As shown in figure 1, simulations suggest deterministic behavior, at least when $n(t)$ is not too small. We treat both $n(t)$ and t as continuous variables, which we justify heuristically by focusing on the case $1 \ll n(t) \ll N$ or $1 \ll i \ll N$ in equations (3) through (5). In this case, the expected value of $n(t+1)$ is large, whereas the variance of $n(t+1)$ is much smaller because $n(t)/N$ is small. For the same reason, the expected change in $n(t)$ is a small fraction of its current large value. Subtracting $n(t) = i$ from both sides of (3), we write

$$\frac{dn(t)}{dt} = -\frac{n(t)(n(t)-1)}{2N}. \qquad (6)$$

The solution of this differential equation, with $n(0) = n_0$, is

$$n(t) = \frac{n_0}{n_0 - (n_0 - 1)e^{-t/2N}}, \qquad (7)$$

which predicts the NLFT for a population of constant size.

Note that the assumption, $1 \ll n(t) \ll N$, which we used to justify our approach implies that we should not keep the $O(n(t)/N)$ term in (6). If we neglect this term, as in Rauch and Bar-Yam (2005), then we would instead obtain

$$n(t) = \frac{n_0}{1 + n_0 t/2N}, \qquad (8)$$

which agrees very closely with our (7) as long as $n(t)$ is large. However, when $n(t)$ becomes small (e.g., for large t), such

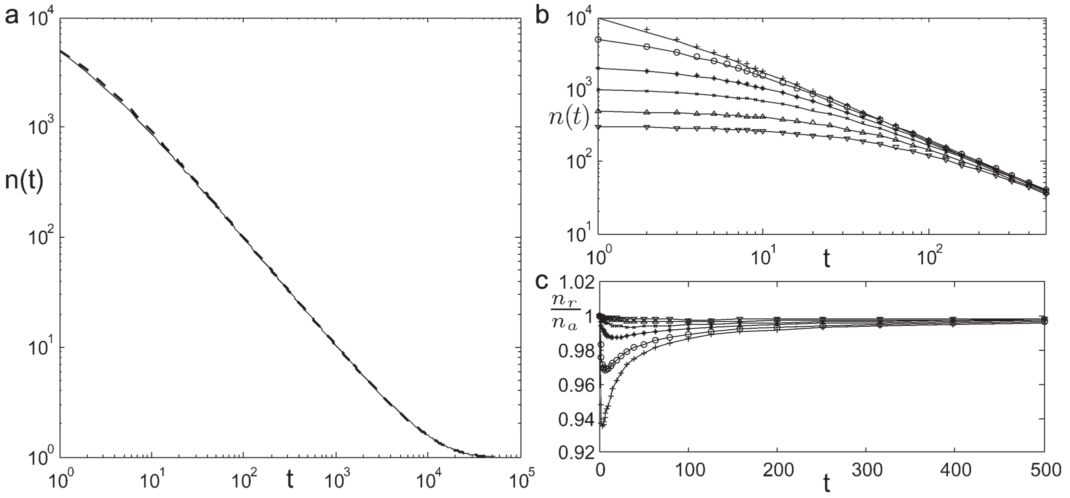


FIG. 2. Panel a shows the expected number of lineages as a function of time in the past for a constant-sized population. The solid line is the exact recursion equation (1) and the dashed line is the continuum approximation (9), for a sample of the whole population, that is, $n_0 = N$ (here, $n_0 = N = 5,000$). Panels b and c show the quality of approximation (9) as the sample size grows. In b, the solid lines show the number of lineages obtained by the recursion equation (1) and the points show the continuum approximation, plotted as a function of time for different sample sizes. In c, the ratio between the two is shown. The maximal deviation is about 7% for full sampling ($n_0 = N$) and the error decreases with time. The population size is $N = 10^4$ and the sample sizes are $n_0 = 300; 500; 1,000; 2,000; 5,000; 10,000$, denoted by triangle-up, triangle-down, ×-mark, asterisk, circle, and plus sign, respectively.

that $n(t)/2N$ is not negligible compared with $n(t)^2/2N$, then (8) and (7) differ considerably. Although the fluctuations in $n(t)$ become substantial in this case (see fig. 1), we prefer (7) over (8) because keeping the $O(n(t)/N)$ term in (6) provides a better description of the "average" behavior of $n(t)$ when it is small. In view of this, we rewrite (7) as

$$E[n(t)] \simeq \frac{n_0}{n_0 - (n_0 - 1)e^{-t/2N}}. \tag{9}$$

The results presented below show that (9) predicts the average value of $n(t)$ with surprising accuracy even for very short and very long times when we do not necessarily expect it to fit.

The following heuristic argument provides a sense of the range of time over which we can expect the behavior of $n(t)$ to be largely deterministic. Let q_t be the chance that an individual who lived t generations ago has at least one descendant in the sample. As long as correlations among such individuals may be neglected—that is, as long as the typical number of sampled descendants of an individual does not constitute a substantial fraction of the sample—we may assume that the probability distribution of $n(t)$ is binomial$(N, 1-q_t)$. Then, the average scales like $(1-q_t)N$, and the variance like $q_t(1-q_t)N$. The quantity $1-q_t$, thus, must fall in time like $1/t$. Accordingly, the standard deviation of $n(t)$ falls, to the leading order, like $1/\sqrt{t}$. Therefore, the ratio between the standard deviation and the mean grows like $\sqrt{t/N}$. For a large N, this quantity is small, approaching unity only for t of order N when the number of ancestors is small and the correlations cannot be neglected in any case.

A comparison of the continuous equation (9) with the recursion equation (1) is presented in figure 2*a*. One can see that the deviations from the exact recursion formula appear very small even for a sample of the whole population $n_0 = N$ and over the entire range of time. In figure 2*b* and *c*, we examine more closely the effect of varying the sample size n_0 on the correspondence between our continuous solution and the exact recursion, focusing particularly on small t. This is important because our differential equation (6) neglects the possibility of multiple coalescent events in a single generation, whereas these are fairly sure to occur for large samples from finite populations under the Wright–Fisher model. However, figure 2*c* shows that even in a full sample, $n_0 = N$, the discrepancy is not larger than about 7%, and it vanishes as time increases into the past.

We have not depicted the way in which $n(t)$ depends on N, but this is straightforward: When N is smaller, $n(t)$ will decay faster per generation as we follow the ancestry of the sample back in time. Because of this and owing to the nearly deterministic behavior of the NLFT, if we knew $n(t)$ for the gene genealogy of a given sample, we could easily retrieve the population size N. The inset of figure 3 shows the results of estimating the population parameter N, using a best fit between (9) and the true $n(t)$ for each single gene genealogy. The distribution of estimates shown in the inset of figure 3 is clustered tightly around the true value of N, with a standard deviation ($\sigma = 68$) of only about 1% of the true value of $N = 5,000$ in this case.

The NLFT in Growing Populations

Analogous results hold for growing populations. As mentioned above, we assume that $N(t) = N_0 e^{-\gamma t}$, where $N_0 \equiv N(0)$ is the current total population size, and γ is the per-generation growth rate of the population. In this case,

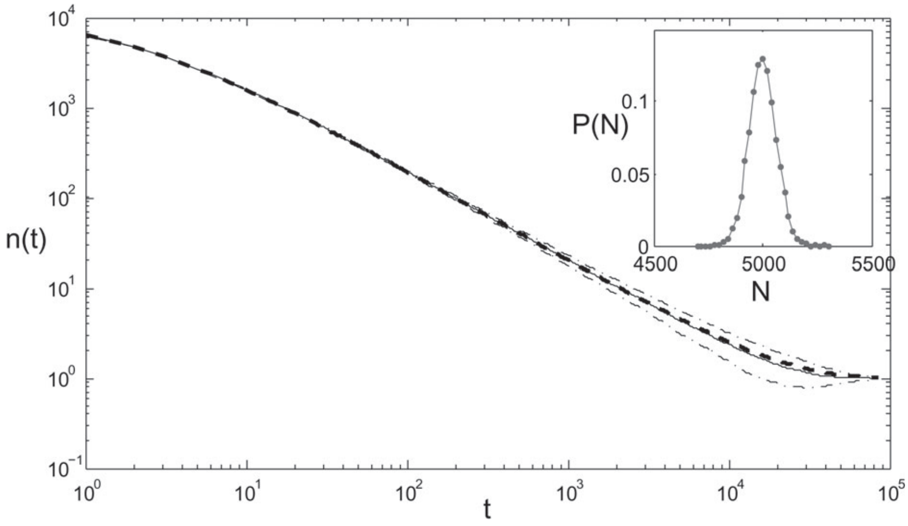


FIG. 3. The number of lineages as a function of time. The middle solid line is the average taken from 100 realizations of the simulation with $N = 5,000$, $n_0 = 5,000$. The dashed line is the expression for $E[n(t)]$, obtained by iterating the single generation recursion (1). The dots (which appear as two thick lines toward the end of the graph) are the average number of lineages in the simulated gene genealogies, ± the standard deviation. The inset shows the distribution of estimated N, obtained by fitting the number of lineages as function of time, which is known exactly in the simulation, to our new expression (9). The distribution was obtained from 5000 simulated gene genealogies.

the gene genealogy is characterized by two demographic parameters: the growth rate γ and the current population size N_0. As before, we start with a sample of size n_0 and, again, the fluctuations of $n(t)$ among different gene genealogies with the same demographic parameters are weak. This is shown in figure 4 and supplementary figure S1, Supplementary Material online, which repeat figures 1 and 3 for the case of a growing population.

Equations (1) and (2) can be easily generalized to growing populations. Note that we assume that the growth of the population is deterministic, that is, $N(t)$ is not a random variable. We have

$$E[n(t+1)|n(t)=i] = N_0\,e^{-\gamma(t+1)} - N_0\,e^{-\gamma(t+1)}\left(1-\frac{1}{N_0\,e^{-\gamma(t+1)}}\right)^{i}, \quad (10)$$

with the formula for $\mathrm{Var}[n(t+1)|n(t) = i]$ (not shown) obtained similarly by replacing N with $N(t+1) = N_0\,e^{-\gamma(t+1)}$ in equation (2).

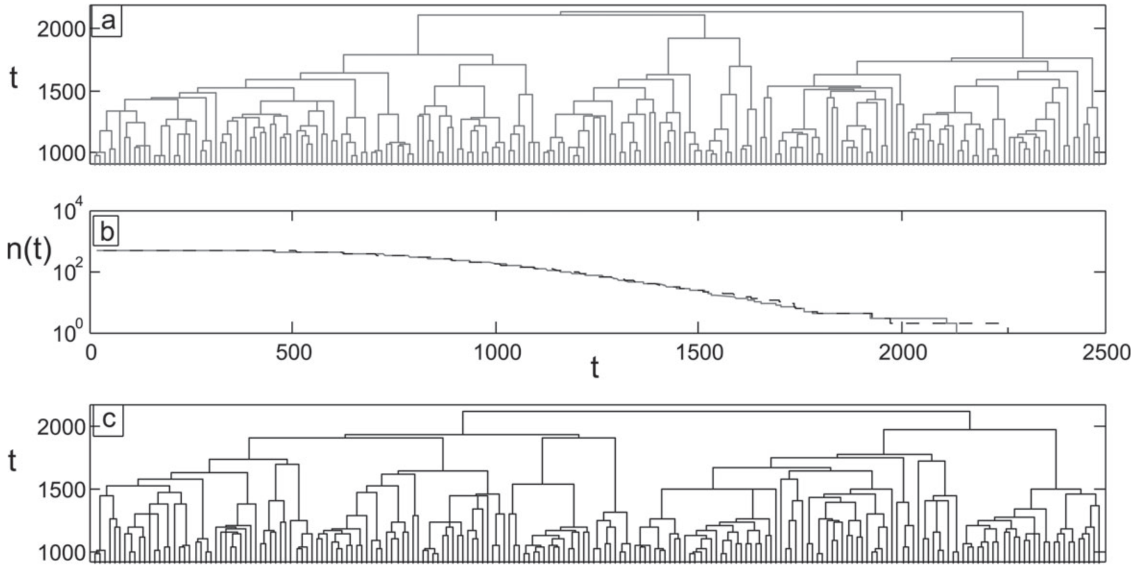


FIG. 4. Similar to figure 1, but for a growing population. The parameters are $N_0 = 5 \times 10^5$, $\gamma = 0.005$, and $n_0 = 500$. Note that here time is given simply in generations rather than on a log scale as in figure 1. The solid and dashed lines in b correspond to the $n(t)$ of the trees in a and c, respectively.

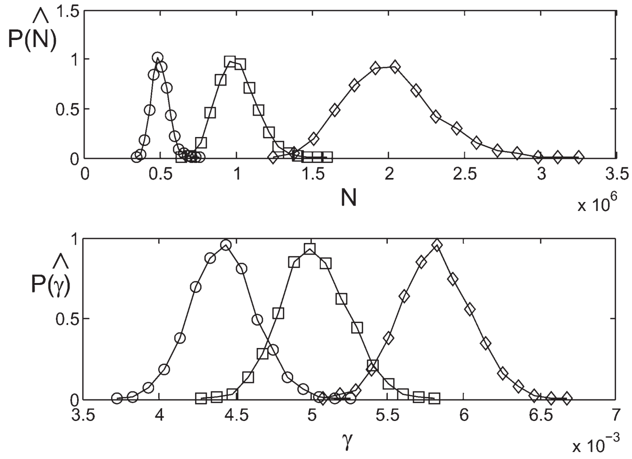


Fig. 5. The distributions of estimated parameters, $\hat{N_0}$ and $\hat{\gamma}$, for three different populations. We simulated 5, 000 independent gene genealogies and estimated N_0 and γ for each one using the true $n(t)$ for three sets of parameters: $N_0 = 5 \times 10^5$ and $\gamma = 0.0044$ (circles), $N_0 = 10^6$ and $\gamma = 0.005$ (squares), and $N_0 = 2 \times 10^6$ and $\gamma = 0.0058$ (diamonds). The sample size was $n_0 = 500$.

As in the case of a constant-sized population, figure 4 and supplementary figure S1, Supplementary Material online, show that $n(t)$ under population growth has an almost deterministic behavior, so that we may again employ the analytical approach described above. Analogous to (6), for a growing population, we have

$$\frac{\mathrm{d}n(t)}{\mathrm{d}t} = -\frac{n(t)(n(t)-1)}{2N_0\,\mathrm{e}^{-\gamma t}}.$$

The solution of this equation, subject to the condition $n(0) = n_0$, is

$$n(t) = \frac{n_0}{n_0 - (n_0 - 1)\,\mathrm{e}^{-\frac{\mathrm{e}^{\gamma t}-1}{2N_0\gamma}}}. \qquad (11)$$

For the replacement of the difference equation by the differential equation to be valid, the condition $n(t) \ll 2N\exp(-\gamma t)$ should hold. As with (9), we rewrite (11) as

$$E[n(t)] \simeq \frac{n_0}{n_0 - (n_0 - 1)\,\mathrm{e}^{-\frac{\mathrm{e}^{\gamma t}-1}{2N_0\gamma}}}. \qquad (12)$$

Supplementary figure S2, Supplementary Material online, shows a comparison of the results of the recursion equation (10) and the continuum approximation (12). As for the case of a constant-sized population (shown in fig. 2), the correspondence under population growth is very good.

Since the fluctuations of the NLFT around this average (12) are weak, we can again very effectively use the values of $n(t)$ from a single gene genealogy to estimate population parameters, in this case N_0 and γ. As in the case of a constant-sized population, the recovered demographic parameters change only slightly between gene genealogies simulated with the same parameters. The quality of the resulting estimates is shown in figure 5, which demonstrates that populations with fairly similar values of N_0 and γ can be distinguished easily. Moreover, it appears from figure 5 that the estimates are unbiased in contrast to the biased estimates for small samples using a Markov chain Monte Carlo likelihood approach even for the case where the real genealogy is known (Kuhner et al. 1998).

Finite Sequence Length—Analyzing Recovered Trees

In the NLFT in Growing Population section, we estimated parameters using the true gene genealogies, which are known in simulations. In reality, any information about the gene genealogy of a sample must be inferred from genetic data, typically DNA sequences. We would only have perfect knowledge of the gene genealogy if the sequence length was infinite, so that every mutation was observable (Watterson 1975), and the mutation rate was infinite, so that the number of mutations on each branch in the tree reflected precisely the length of the branch. Then, we could simply construct a matrix of the number of differences between each pair of sequences and from these infer the gene genealogy and $n(t)$ without error. This suggests the following three-stage routine for estimating population parameters from genetic data:

1. Reconstruct the gene genealogy from the sequence data, for example, using a clustering algorithm.
2. Extract the number of lineages as a function of time, $n(t)$, from the recovered tree.
3. Fit the theoretical prediction, for example, (9) or (12), to the $n(t)$ extracted from the recovered tree.

The parameter estimates are the values that provide the best fit in Step 3. In trying to implement this technique to real data, one encounters two major obstacles.

First, the real mutation rate μ is finite. There is not a one-to-one correspondence between genetic and genealogical distance, and this causes the statistical properties of the tree recovered from the clustering algorithm to differ from those of the true tree. Thus, the "estimated" NLFT function, $n_e(t)$, differs from the true function. Figure 6a shows $\langle n_e(t)\rangle$—the average value of $n_e(t)$ over many simulation replicates with the same population parameters—together with the true $n(t)$, for different mutation rates in a growing population. The deviations become large when μ is small and the data contain less information about the gene genealogy. For a constant-sized population, the same comparison is shown in the upper panel of supplementary figure S3, supplementary Material online, where the effect is less severe but still considerable.

The second problem comes from the fact that the sequence length, L, is necessarily finite, so that there may be multiple mutations at single sites. This also yields a distortion of the recovered tree with respect to the real one and is demonstrated in figure 6b for a growing population and in the lower panel of supplementary figure S3, Supplementary Material online, for a fixed population. For a given mutation rate, shorter sequences experience more recurrent mutation, causing the distortion to be greater. As a result of these two problems, the naive procedure, based on the idea that we might extract the true $n(t)$ from the data, is

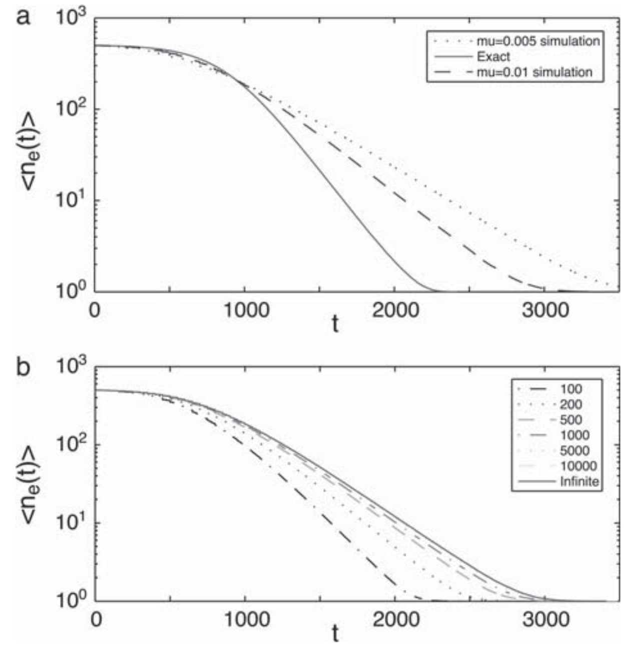


Fig. 6. Panel a: The average estimated NFLT, $\langle n_e(t) \rangle$, plotted against time for different mutation rates. The data were obtained from simulations with parameters $n_0 = 500$, $N_0 = 4 \times 10^6$, and $\gamma = 0.005$. The solid line is the true NFLT (i.e., infinite mutation rate), the dashed line is for $\mu = 0.01$, and the dotted line is for $\mu = 0.005$. Panel b: The average estimated NFLT, $\langle n_e(t) \rangle$, plotted against time for different sequence lengths. The data were obtained from simulations with parameters $N_0 = 4 \times 10^6$, $\gamma = 0.005$, $n_0 = 500$, and $\mu = 0.01$. Sequence lengths are given by shades and styles of lines as in the legend.

not applicable for real data sets such as the one we analyze below.

It should be stressed, though, that the estimated NLFT, $n_e(t)$, is still strongly indicative of the population parameters. As demonstrated in figure 7 for different values of γ and in supplementary figure S4, Supplementary Material online, for different values of N_0, $n_e(t)$ depends on both the size of the population and its growth rate, and the standard deviations are not large. The noise and systematic deviations introduced by finite mutation rate and recurrent mutations are not strong enough to render $n_e(t)$ useless. However, they do distort the tree so that the retrieved NLFT differs substantially from the analytical predictions, (9) or (12). If we had a function—let us call it $\tilde{n}_{\gamma,N}(t)$, with the subscripts γ and N to emphasize its dependence on those parameters—which predicted the outcome of the NLFT measured from the clustering algorithm, then we could return to Step 3 of the routine described above and calculate the value of the parameters by minimizing $\chi^2(n_e, \tilde{n})$ by which we mean the squared deviations between $n_e(t)$ and $\tilde{n}_{\gamma,N}(t)$. It is, perhaps, needless to say that we have not found any analytic expression for $\tilde{n}_{\gamma,N}(t)$.

Simulation-Based Inference Method

The solution we adopt is to use our simulations themselves to predict $\tilde{n}_{\gamma,N}(t)$, as the average function $\langle n_e(t) \rangle$ obtained by applying the clustering algorithm to a large number of simulated data sets with given values of N and γ. The simulations require a mutation model, and unless mentioned otherwise, we use the simple Jukes–Cantor model (Jukes and Cantor 1969) as well as L and μ which we assume are known. We emphasize that we do not take errors in the estimated μ into account. However, for some of the simulations and in the data application, we use the more realistic F84+Γ mutation model as mentioned above. For a clustering algorithm, we used the simple WPGM (Sokal and Michener 1958; Sneath and Sokal 1973). However, we verified the basic properties depicted in figures 6b through 7 for other clustering algorithms, including UPGMA and neighbor joining (Saitou and Nei 1987), indicating that any clustering algorithm could be used in our method.

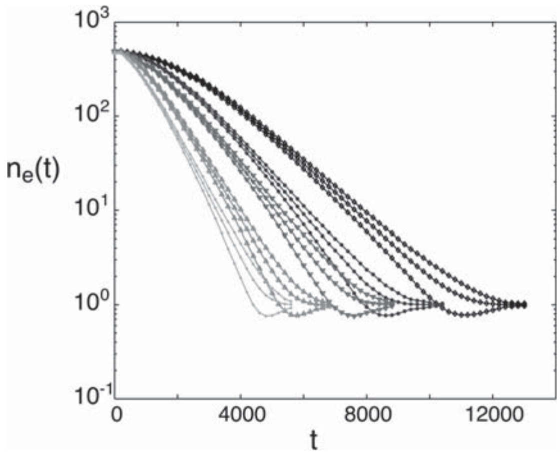


Fig. 7. The average ± one standard deviation of the NLFT, $n_e(t)$, estimated from simulated data, for a series of different growth rates γ. The growth rates are 0.004 (dots), 0.003 (triangles-up), 0.002 (triangles-down), 0.0015 (circles), and 0.001 (diamonds). The other parameters are $N_0 = 4 \times 10^6$, $n_0 = 500$, and $\mu = 0.0023$.

This method—estimating parameters by minimizing the squared deviations between the $n_e(t)$ for a data set and $\tilde{n}_{\gamma,N}(t)$ from simulations—has several advantages. First, it is general and may be adapted to estimate other quantities. Thus, although here the analytical work was instrumental in the development of the method, the method itself is not restricted to functions for which analytic expressions may be obtained. Second, it can accommodate any tree-building algorithm, as the results for $n_e(t)$ and $\tilde{n}_{\gamma,N}(t)$ will be subject to the same distortions as long as $n_e(t)$ and $\tilde{n}_{\gamma,N}(t)$ are produced by the same process. Finally, in the present case, it may be used for large data sets easily because the complexity of the calculation depends only slightly on n_0 as a result of the nearly deterministic behavior of the NLFT. For example, we can obtain a fairly precise picture of the surface of squared deviations we wish to minimize by simulating "tens" of gene genealogies for each pair of values, (N, γ).

Two further, technical details of our method for estimating N_0 and γ are as follows.

First, for simplicity, we obtained parameter estimates by minimizing the squared deviations between simulations and data based on the times of each of the $n_0 - 1$ coalescent events rather than on $n(t)$ itself. This allows us to avoid summing deviations over a very large number of generations, making an arbitrary decision about when to stop this sum and having to store the simulation-obtained average function $\tilde{n}_{\gamma,N}(t)$ for such a large number of time points. The $n_0 - 1$ coalescent events provide a natural set of anchor points, and the time to the ith coalescent event is simply an inversion of the relationship between n and t for which we found useful deterministic formulas in the previous sections. For a constant-sized population, it is well known from coalescent theory that the expected value of this time is, using our notation, given by

$$t(i) = 2N\left(\frac{1}{n_0 - i} - \frac{1}{n_0}\right).$$

We note that this equation may be obtained by inverting (8) but not by inverting (7). However, the equation above applies only to the true gene genealogy and not to the tree recovered from data using a tree-building algorithm.

Second, we note that the common methods of finding global minima, such as the Levenberg–Marquardt algorithm, are difficult to apply here. The reason is that these algorithms require the use of the derivative of the surface of squared deviations with respect to the parameters. Here, we can obtain the derivative only by evaluating the squared deviation at discrete points in the parameter space. This involves averaging over many Monte Carlo realizations at each point considered, where the variance goes like $1/\sqrt{k}$ if k is the number of realizations. A good approximation of the derivative of surface of squared deviations would require a precise evaluation at two very close points in the parameter space. This would necessitate averaging over a very large number of realizations at each step in the algorithm procedure and would make the whole process inefficient.

Therefore, in the search for the best parameter estimates, we used a two-step method that is applicable due to the relatively simple structure of the surface of squared deviations. Looking at supplementary figure S5, Supplementary Material online (as well as fig. 11), one can see that the surface consists of one global minimum located in a single valley. We found this to be true in every case we considered. The shape of this valley reflects the fact that increasing both N and γ results in trees with similar structure. The relevant parameter space can be scanned first with low accuracy, that is, using a small number of realizations k for each (γ, N) to locate the region containing the global minimum. Next, the minimum can be found more precisely by searching over a grid of (γ, N) values, more closely spaced and with a larger number of realizations at each point. To save time, we begin by fitting $n_e(t)$ to (12) to find initial parameters around which to scan. In the simulations to illustrate the method, we used one gene genealogy per point in the first pass, then 20 gene genealogies per point to find the minimum. In the application to the mtDNA data, we increased this from 20 to 40.

We used a parametric bootstrap approach to determine the error of the estimates. That is, using the estimated values of N_0 and γ, we generated a large number of replicate data sets by running our simulation with different seeds for the random number generator. For each replicate, we estimated N_0 and γ using our simulation-based technique. Figure 8 shows the results obtained from fitting 300 different realizations under the same true parameters $N_0 = 10^5$, and $\gamma = 0.003$. The average estimates are $N_0 = 1.008 \times 10^5 \pm 2.05 \times 10^4$, and $\gamma = 0.00303 \pm 0.00052$, where the range of error is defined by the approximate 95% confidence intervals assuming a bivariate normal distribution of estimated N_0 and γ. Our estimates thus appear to be “unbiased” and the uncertainty in the estimate of the growth rate is quite small, approximately 28% of the value of γ.

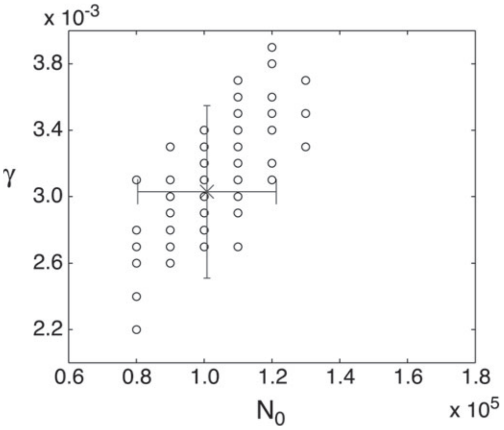


FIG. 8. The parameters estimated by minimizing the squared deviations for 300 different simulated gene genealogies, each from a different simulation with $N_0 = 10^5, \gamma = 0.003$. The average of these realizations and the 95% confidence intervals are $N_0 = 1.008 \times 10^5 \pm 2.05 \times 10^4, \gamma = 0.00303 \pm 0.00052$. The sample size used is $n_0 = 500$, the sequence length is $L = 400$ base pairs and the mutation rate is $\mu = 0.01$ per generation.

Here, we presented one example of our fitting process. Before presenting a wide range of results, we should first describe the limitations and scope of the technique. Generally, the time horizon imposed by the use of polymorphism data is T_{MRCA}. Our method, however, assumes an almost deterministic behavior of NLFT; in the cases that we have studied, this deterministic approximation works as long as the number of lineages is larger than five or so, and thus one has to replace T_{MRCA} with $T_{det} \sim T_{n=5}$. Within this time horizon, three conditions should be fulfilled.

- $C1$: The change in the population size should be large enough to be detectable, and thus $1/\gamma$ must be much smaller than T_{det}.
- $C2$: T_{det} defined above should allow for enough mutations to occur, and thus the mutation rate must satisfy $1/\mu \ll T_{det}$.
- $C3$: When too many mutations occur, the number of repeated mutations is large and the information about the far history is lost. This implies that the mutation rate should not exceed $\mu \cdot T_{det} \ll L$.

With these conditions in mind, let us take a look at figure 9. For each pair of γ and N_0, we have simulated 1,000

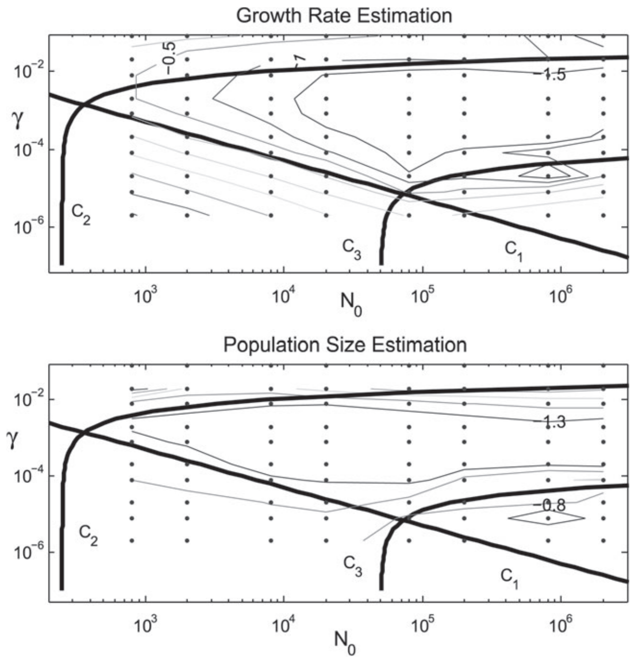


FIG. 9. A contour plot of the $\log_{10}$ of ratio between the average deviation from the real values and the real values, with panel a for the growth rate γ and panel b for the effective population size N_0. The dots indicate the pairs that were checked. The thick lines are the boundaries of the validity region: below the C_1 line ($\exp(\gamma \cdot T_{\mathrm{MRCA}} = 2)$) the change in the population size is too small. Above C_2 ($T_{\mathrm{MRCA}} \cdot \mu = 10$) the number of mutations is too small, and below C_3 ($T_{\mathrm{MRCA}} \cdot \mu = L = 1{,}000$)) there are too many repeated mutations. The T_{MRCA} was calculated using a parallel expression to equation (9), which is given in supplementary equation 1, Supplementary Material online.

genealogies and inferred the parameters using our technique. An objective measure for the quality of the inference is the ratio between the real parameter and its recovered value. Contours of this ratio are plotted in the (N_0, γ) plane together with thick black lines for the conditions C1–C3. One can see that inside the validity region, the error of the estimates is less than 5% of the real values, indicating that our method is not biased. Even outside the validity region, the estimates are not bad. In supplementary table S1, Supplementary Material online, we present the average estimate and 95% confidence interval for each of the points in figure 9.

Figure 10 shows how the inference quality depends on the mutation rate. The average estimates for N_0 and γ are shown, together with their standard deviation, for different values of μ. Clearly seen is a wide "validity region," where the estimates are unbiased and quite tight. Again, even beyond the validity region, the bias and the errors are not terribly large, and the true values are within one standard deviation of the average estimated value.

We have focussed on the case of an exponentially growing population, but our method of estimation can easily be applied to populations of constant size. As above, we define the fitting function $\tilde{n}(t)$ to be the average over many simulation replicates, $\tilde{n}(t) \equiv \langle n_e(t) \rangle$. Supplementary figure S6, Supplementary Material online, shows the average best fit N and its standard deviation (over 200 replicates) for population sizes ranging from $N = 25{,}000$ to $N = 190{,}000$ in increments of 1,000. The average best fit N was in all cases very close to the true population size, and the standard deviation was about 20%. This whole process took only about 10 cpu hours on a regular desktop. The sample size was $n_0 = 50$, and even so the estimates are reasonably good.

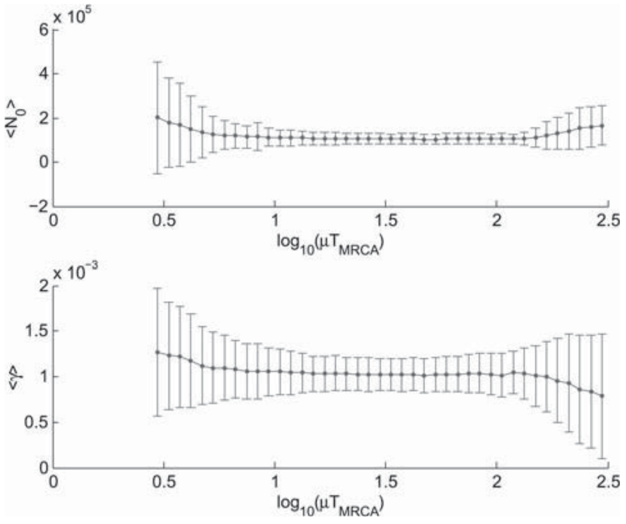


FIG. 10. The average estimated parameters, with error bars representing one standard deviation, as a function of the mutation rate. The parameters were inferred from simulated data, where all histories share the same current population size ($N_0 = 10^5$), growth rate ($\gamma = 0.001$), and sequence length ($L = 500$). One thousand histories were produced for each mutation rate, and the parameters of each were inferred from the polymorphism data. Inside the "validity zone" both the effective population size (upper panel) and the growth rate (lower panel) are estimated without a significant bias. Outside the validity region the estimates are reasonable but slightly biased. The sample size used was $n_0 = 100$; for larger samples, the validity zone is even wider.

A third demographic scenario to which we will briefly relate is the case of a stepwise growth model. This model assumes that the population size was fixed in the past at N_2, and then at time T_c, the population suddenly grew to a new size N_1. We use this model as an example of a case for which there is no elegant analytic solution for the NLFT, yet the simulation-obtained function can be used. (Also, we choose to analyze this model because in the coming section, we will consider how to differentiate between an exponential growth model and a stepwise growth model as done by Polanski et al. (1998).

In this case also we define the function $\tilde{n}(t)$ to be the average over many simulation replicates, $\tilde{n}(t) \equiv \langle n_e(t) \rangle$ and scan the (three dimensional) parameters space to find the global minima of χ^2. From this procedure, all three parameters, N_1, T_c, and N_2, are inferred. In table 1, we present of the average estimates and 95% confidence intervals (over 1,000 replicates) for a few sets of parameters. These estimates appear to be good and with a small error range. A more thorough study of the stepwise growth model is outside the scope of this paper.

Table 1. Stepwise Growth: The Average Estimates of the Parameters for the Stepwise Growth Model. For Each Set of Parameters, We Used 1, 000 Replicates. The Rest of the Parameters are $n_0 = 200$, $\mu = 0.01$, $L = 1000$, $\kappa = 20$, and $\alpha = 0.25$.

Real N_1	Real T_c	Real N_2	$E(N_1)$	$E(T_c)$	$E(N_2)$
5×10^5	8×10^3	8×10^3	5.2×10^5 [2.5×10^5–1×10^6]	8.3×10^3 [4×10^3–1.2×10^4]	9.2×10^3 [316 - 2.5×10^4]
1.5×10^6	1×10^3	$1.5. \times 10^4$	2.4×10^6 [3.1×10^5–7.9×10^6]	1.2×10^3 [4×10^2–2×10^3]	1.9×10^4 [5×10^3 – 3.9×10^4]

Model Differentiation

Throughout this paper, we have assumed a model for the population demography (for example, an exponential growth at a fixed rate) and have tried to infer the real rates from the observed data. One question needed to be asked is about the ability of our technique to distinguish between demographic scenarios. Obviously, it is easy to differentiate between very different scenarios like an exponentially growing population and a fixed population. On the other hand, one cannot distinguish between an exponential growth and a growth occurring through many small steps. Here, we compare an exponential growth model with the stepwise model we presented above. It has been shown by Di Rienzo and Wilson (1991) and Slatkin and Hudson (1991) that it is indeed difficult to differentiate between these two scenarios. Methods like Polanski et al. (1998) that do not assume a specific growth model are able to discriminate between the two scenarios, but we will show that despite the fact that our technique assumes a model, it can still differentiate between the two.

One of the problems in assuming a model is that even when the model is wrong, it supplies some results. For example, if a data set from a population that underwent a stepwise growth is taken and fitted to an exponential growth model, the fitting procedure will find the values of N_0 and γ that minimize the value of the χ^2. These wrong values will be considered the description of the history of the population. However, there is a way to detect this false procedure and to identify that the model is incorrect.

When a wrong model is assumed to describe the history, the function of squared deviations between the model and the data, χ^2_{D}, will be relatively large. By saying large, we mean relative to the χ^2_{M} between the model and a data set that was obtained from this model. Ideally in order to determine this, the distribution of the χ^2_{M} (i.e., the distribution between all the possible data sets obtained from the model with specific parameters and the model) should be used. If the χ^2_{D} of the data is in the (very)upper end of this distribution, this means that the model does not describe the data set well and thus should be rejected. However, if it is in range, the model should not be rejected but rather be kept as a plausible description of the history.

Practically, we implement this test in the following way which we will explain by way of example. Assume we obtain a data set from a stepwise growth model and are trying to fit it to an exponential growth model. The fitting process will produce some values of γ and N_0 rendered by our algorithm. We then simulate 50 replicates of exponentially growing populations with these parameters and fit them to the exponential growth model obtaining the minimal χ^2 of each replicate. Thus, instead of the whole distribution of χ^2_{M}, we have a set of χ^2_{M} between the model and the replicates. A model is rejected if χ^2_{D} is larger than the maximum χ^2_{M}, which is equivalent (on average) to saying that a model is rejected if less than 2% of the χ^2_{M} are larger than the χ^2_{D}. This procedure may be used, of course, for any possible model. Note that we cannot use the standard P-value test because this test requires knowledge about the error estimates of the data set.

Table 2 summarizes the results of our rejection test. We have generated 1, 000 replicates of the exponential growth model and 1, 000 of the stepwise growth model and fitted both models to both data sets in order to see the rejection percentages of the wrong model and of the true model. One clearly observes that the wrong model has been rejected almost certainly, whereas the chance of the right model being rejected is small.

One last point that we should emphasize about our rejection test is that it cannot positively confirm a model but can only reject or not reject it. If the model is not rejected, it is always possible that another model fits the data as well. Thus, when one compares a data set to two models, both models can be rejected, one can be rejected but not the other one or neither be rejected.

Human Growth Rates

As a first illustration of our method, we applied the method to sequences of hypervariable region 1 (HRV1) of human mtDNA for which a very large number of sequences is available at HvrBase++ (http://www.hvrbase.org) (Handt et al. 1998; Kohl et al. 2006). Specifically, we obtained 1, 212 sequences of HVR1 from China and aligned them using the tools available at HvrBase++. The aligned data set contains 377 sites, and is available from the authors upon request. We used an HVR1 mutation rate of $\mu = 0.0024$ per 377 bp sequence per generation (Sigurdardottir et al. 2000) as the mutation rate in our simulations to estimate N_0 and γ.

We estimated the growth rate and effective size for China using the methods described above. The one modification we made was to correct the pairwise sequence differences in HVR1 for multiple mutations. We used the F84 model with gamma-distributed rates across sites (Felsenstein 2004), as it is implemented in Dnadist (http://evolution.genetics.washington.edu/phylip/doc/dnadist.html) with equal base frequencies, $\kappa = 20$, and $\alpha = 0.25$ (Rosset et al. 2008). We used this same mutation model in the simulations for estimating N_0 and γ and obtaining confidence intervals.

Table 2. The Percentage of Rejections (out of 1,000 realizations) of a Model When Applied to Simulated Data. The Parameters for the Exponential Model Are $n_0 = 200$, $N_0 = 8 \times 10^5$,$\gamma = 2 \times 10^{-4}$,$\mu = 0.01$, $L = 1,000$, $\kappa = 20$, and $\alpha = 0.25$. For the stepwise growth, the parameters are $N_1 = 8 \times 10^5$, $T_c = 8,000$, and $N_2 = 8,000$. The Rest of the Parameters Are the Same.

Fitted Model / True Model	Exponential Growth	Stepwise Growth
Exponential growth	0.033	0.999
Stepwise growth	0.980	0.165

Figure 11*a* shows the surface of squared deviations for these data. The estimates from the global minimum correspond to $N_0 = 210,000 \pm 46,000$ and $\gamma = 0.0021 \pm 0.0005$, where the ranges represent the approximate 95% confidence intervals for each parameter assuming a normal distribution of errors estimated using our parametric bootstrap approach. We also show the global minima of 100 replicates from our bootstrap simulations (i.e., estimating N_0 and γ from data sets simulated with $N_0 = 210,000$ and $\gamma = 0.0021$) for a visual sense of the error of our parameter estimates. Figures 11*b* and *c* demonstrate the approximate normality of the distributions of our estimates obtained using our parameter bootstrap method.

We applied the rejection test presented above to the application of the exponential growth model to the Chinese data. We built the χ^2_M distribution from 10,000 replicates. The χ^2_{China} of the data was compared to this distribution and 2% of the values were larger than χ^2_{China}. Thus, we conclude that the exponential growth model is plausible. We also fitted the data set to the stepwise growth model. We found that the χ^2_{China} between the NLFT and the stepwise growth model was larger than all the χ^2_M of the stepwise growth replicates. From this, we conclude that the stepwise growth model is not a plausible explanation for the data.

Our idealized model of the exponential growth of a single population does not likely fit the truth for the ancestry of these samples from China. Thus, we may call our estimates of N_0 and γ the current effective population size and growth rate. Note that here we mean a very short-term effective population size, that is, inbreeding or variance effective size (Crow and Denniston 1988) and not a long-term effective size such as might be estimated from genetic polymorphism data. Figure 12 compares our predictions for the effective population size in the ancestry of these 1, 212 samples from China to the results (Atkinson et al. 2007) obtained using 28 mtDNA sequences from “North and Central Asia” using a Bayesian skyline plot. The agreement between the median estimates is very good indicative of a strong signal of population growth in these human data. Our estimates of the approximate 95% confidence intervals for past effective population sizes are quite narrow, reflecting the size of the data set and the nearly deterministic behavior of the NLFT for large samples. We obtained percentiles for estimates of past effective sizes by sorting the population sizes predicted from our model of exponential growth for each of 10,000 bootstrap replicates at each generation in the past. We should mention that the confidence interval estimates that we obtained assume a model without any stochasticity. However, the history of the real population is probably more

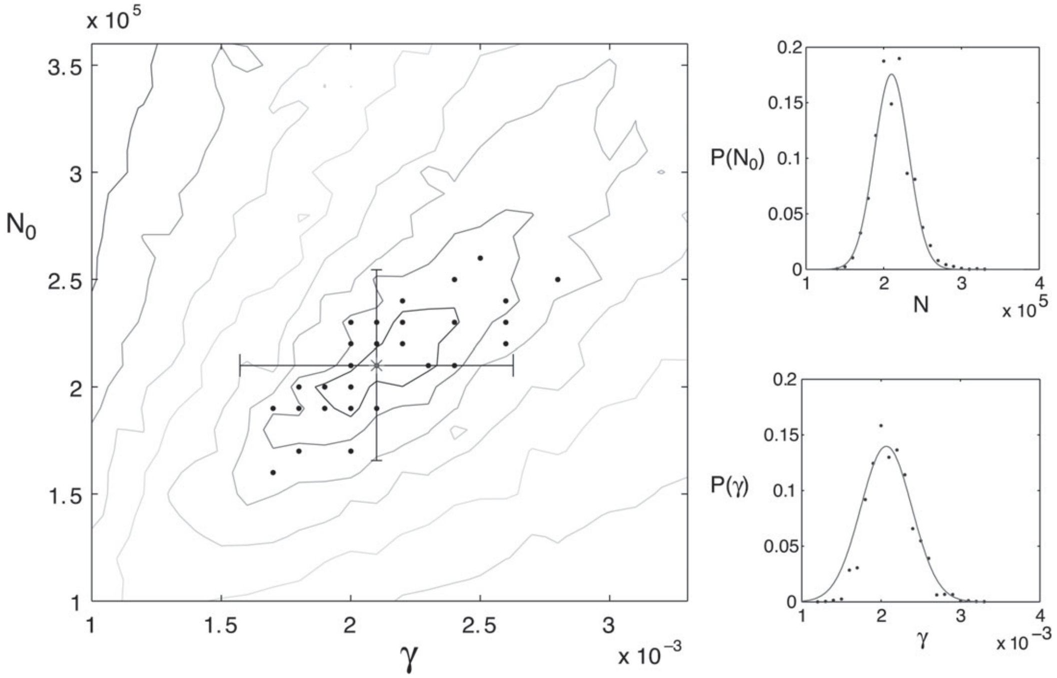


Fig. 11. Panel a: The surface of squared deviations (on a log scale) for the mtDNA from China. The asterisk is the global minima at $N = 210,000$ and $\gamma = 0.0021$. The dots are parameters estimated for 40 Monte Carlo realizations with the same parameters. Panels b and c: Comparison of the distribution of estimates of N_0 and γ, from 10,000 realizations, with the best fit normal distributions. The 95% approximate confidence interval obtained from these realizations yields $N_0 = 210,000 \pm 46,000$ and $\gamma = 0.0021 \pm 0.0005$.

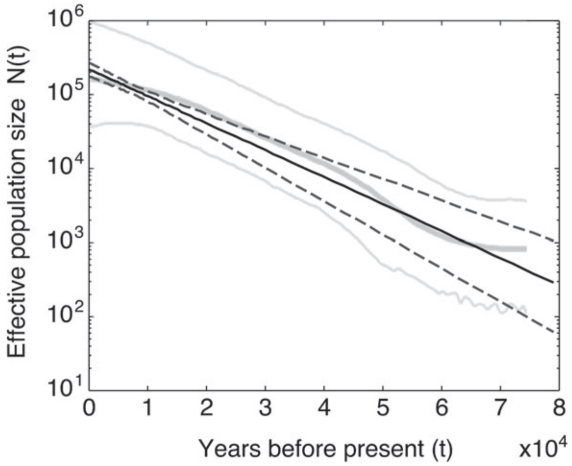


FIG. 12. Comparison of predicted effective population sizes backward in time for our estimates from HVR1 sampled from China with the prehistorical effective population sizes inferred by Atkinson et al. (2007) for North and Central Asia (includes China). The solid curve shows our median estimate of the effective population size, whereas the lower and upper dashed curves show the 2.5% and 97.5% cutoffs at each time. We assumed an average generation time of 25 years. The curves in the background are the median and 95% highest posterior density from a Bayesian skyline plot analysis and are redrawn from figure 1*c* of Atkinson et al. (2007).

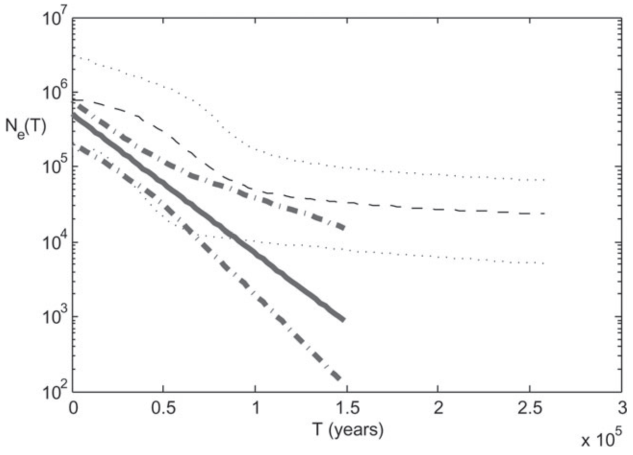


FIG. 13. Comparison of predicted effective population sizes backward in time for our estimates from HVR1 sampled from BWs with the estimates of the Skyline plot in BEAST using the same data set. The solid curve shows our mean estimate of the effective population size, whereas the lower and upper dashed-dotted curves show the 2.5 and 97.5 cutoffs at each time. We assumed an average generation time of 15 years. The dashed line is the average estimate obtained by using BEAST, and the dotted lines are the 95% highest posterior density from a Bayesian skyline plot analysis.

complicated, and thus the error range of the estimated parameters for real data are larger than for simulated data.

Note that our model of growth, which has been used by others as well (Slatkin and Hudson 1991; Kuhner et al. 1998), is not realistic in another sense, which is that the population size will approach zero if followed sufficiently long into the past. This is exemplified by the fact that our predictions continue to decline linearly (on a log scale) with time at the far right of figure 12. Thus, we consider this a model only for the recent growth of a population, and in this sense, our estimates are in good agreement with those of Atkinson et al. (2007). It is also for the most recent times that our estimates of past effective sizes show the least variability.

Note also that there is considerable uncertainty in the literature concerning the mutation rate in human mtDNA (e.g., see Howell et al. 2003). In general, using a smaller (respectively, larger) value of μ in our method would result in larger (respectively, smaller) predicted values for the effective population size in the past. In terms of the parameters, smaller past values of effective size are achieved by smaller values of N_0 and larger values of γ. Our predictions in figure 12 also depend on the number of years per human generation, which we assume to be 25 years. Adopting a shorter generation time would predict a steeper decline of past effective population size.

Blue Whale Population

As a further illustration of our technique, we have analyzed the blue whale (BW) data. We used 63 sequences of the control region of the mtDNA with $L = 299$ obtained from GenBank (Benson et al. 2005). The sequences were aligned using ClustalW Larkin et al. (2007), with its default parameters. The mutation rate per sequence per generation is $\mu = 0.00035$ (Jackson et al. 2009). We again used the F84 model with Γ-distributed rates across sites as implemented in Dnadist with equal base frequencies, $\kappa = 20$, and $\alpha = 0.25$.

An exponential growth fit, when applied to the BW data, yields N_0 =650,000 [135,000–1,575,000] and $\gamma = 0.00067$ [0.00024–0.0012]. The confidence intervals have been produced using 10, 000 replicates. Note that the BW parameters are slightly outside the validity region, so the average estimate is weakly biased. In figure 13, we compare our results with the estimates of population size through time obtained from this data set using the Bayesian skyline model in the BEAST package (Drummond et al. 2005). One can see that both methods give more or less similar results. Our rejection test has been applied to the exponential growth model for the BW growth using 10, 000 replicates. We found that 1.3$\%$ of χ^2_M were larger than χ^2_{BW}, and thus the model is plausible. It is important to explain that the timescale that we are dealing with is much larger than the timescale of the massive whaling of the previous century. Thus, we cannot detect this event. In addition, taking samples from a population a short time after a catastrophe is like sampling from a representative sample of the whole population (if of course the whaling was not biased) and thus it takes time until the polymorphism of the survivor population will be effected by the change in the population size.

We also fitted the stepwise growth model to the BW data and obtained the following results: $N_1 = 1.4 \times 10^6$ $\left[1.2 \times 10^{55}–6 \times 10^6\right]$, $T_c = 3,548$ $\left[2.2 \times 10^3–6.3 \times 10^3\right]$, and $N_2 = 5 \times 10^3$ $\left[1 \times 10^3–1.2 \times 10^4\right]$. The error range was obtained from the bootstrap technique using 10, 000 replicates. We applied the rejection test to this model and the χ^2_D of the data was smaller than 15% of the χ^2_M of the model. Thus, this model was not rejected either. The reason neither of the models were rejected is that in a small sample the stochasticity of the system is large, resulting in a large χ^2 even when a true model is applied. In addition, most probably, the history was somewhere between the two models (as suggested by the skyline method result). We can, however, negate the fixed population assumption suggested by some authors (Jackson et al. 2009). In particular, our results call for a reexamination of previous statements regarding the genetics and demographics of the BW, like the value of the mtDNA mutation rate (Jackson et al. 2009) obtained under the assumption of a fixed population size.

Discussion

We have presented a new method of estimating the effective size and growth rate of a population using a large sample of sequences from a single genetic locus. As a key part of this, we obtained new analytic expressions for the number of lineages ancestral to a sample, as a function of time back into the past. Our simulations and analyses show that the behavior of the NLFT is largely deterministic for a large sample from a large population, suggesting that estimates based on the NLFT will be relatively efficient. Importantly, estimates of the population growth rate using our method appear, from simulations, to be unbiased. Applying our technique to a large sample of human mtDNA from China gives an estimate of the trajectory of the ancestral population size that agrees with other recent estimates. Likewise, our estimate of the BW population agrees with that we obtained by applying another method to the BW data set.

Clearly, our simple population model lacks many important features it would be desirable to include, for example, in the context of our application to human mtDNA from China. As noted in the Blue Whale Population section, its utility diminishes for long times in the past. In addition, for reproduction, we have used the idealized Wright–Fisher model of a single well-mixed population in which all genetic variation is selectively neutral. We have further ignored the fact that China is connected to other regions via complicated migration patterns that have probably also changed over time. We use the phrase effective population size to cover some of the deviations from our model (Möhle 1998; Nordborg and Krone 2002; Sjödin et al. 2005), but future work will be needed to fully deal with population structure and migration, natural selection, and other complications.

Our general aim has been the development of methods of inference tailored to large data sets, such as those that are rapidly accumulating for humans and several other species. Within population genetics, it has been known for some time that taking larger and larger samples, in terms of the number of sequences, is not necessarily the best strategy to improve estimates of population parameters (Pluzhnikov and Donnelly 1996; Felsenstein 2006). However, as we have shown here, some properties of large samples are very desirable and can lead to new approaches to data analysis. Our overall strategy of focusing on certain aspects of genetic variation in large samples, and avoiding those parts of the data that are strongly stochastic, appears to be a fruitful way to extract reliable information from sequence data. Given the phenomenal accumulation of sequence data, methods geared toward large samples may be key to the further development of studies of genetic variation.

Supplementary Material

Supplementary figures S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank David Kessler, Peter Smouse, and Yoram Louzoun for many helpful discussions. The work of N.S. and Y.M. was supported by the CO3 STREP of the Complexity Pathfinder of NEST (EC FP6).

References

Atkinson QD, Gray RD, Drummond A. 2007. mtDNA variation predicts population size in humans and reveals a major southern Asian chapter in human prehistory. *Mol Biol Evol.* 25:468–474.

Baldwin BG, Sanderson MJ. 1998. Age and rate of diversification of the Hawaiian silversword alliance. *Proc Natl Acad Sci U S A.* 95:9402–9406.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.

Becquet C, Przeworski M. 2007. A new approach to estimate parameters of speciation models with application to apes. *Genome Res.* 17:1505–1519.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler D. 2005. Genbank. *Nucleic Acids Research* 33(Special Issue):D34–D38.

Bininda-Emonds OR, Cardillo M, Jones KE, MacPhee RD, Beck RM, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A. 2007. The delayed rise of present-day mammals. *Nature* 446:507–512.

Crow JF, Denniston C. 1988. Inbreeding and variance effective population numbers. *Evolution* 42:482–495.

David FN, Barton DE. 1962. Combinatorial chance. London: Charles Griffin.

Di Rienzo A, Wilson AC. 1991. Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proc Natl Acad Sci U S A.* 88:1597–1601.

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

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates, Inc.

Felsenstein J. 2006. Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Mol Biol Evol.* 23:691–700.

Felsenstein J. 2007. Trees of genes in populations. In: Gascuel C, Steel M, editors. Reconstructing evolution: new mathematical and computational advances. Oxford: Oxford University Press. p. 3–29.

Fisher RA. 1930. The genetical theory of natural selection. Oxford: Clarendon.

Fu Y-X, Li W-H. 1997. Estimating the age of the common ancestor of a sample of DNA sequences. *Mol Biol Evol.* 14:195–199.

Griffiths RC, Tavaré S. 1994a. Simulating probability distributions in the coalescent. *Theor Popul Biol.* 46:131–159.

Griffiths RC, Tavaré S. 1994b. Ancestral inference in population genetics. *Stat Sci* 9:307–319.

Handt O, Meyer S, von Haessler A. 1998. Complilation of human mtDNA control region sequences. *Nucleic Acids Res.* 26: 126–129.

Hein J, Schierup MH, Wiuf C. 2005. Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford: Oxford University Press.

Howell N, Bogolin Smejkal C, Mackey DA, Chinnery PF, Turnbull DM, Herrnstadt C. 2003. The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. *Am J Hum Genet.* 72:672–670.

Hudson RR. 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37:203–217.

Jackson JA, Baker CS, Vant M, Steel DJ, Medrano-González L, Palumbi SR. 2009. Big and slow: phylogenetic estimates of molecular evolution in baleen whales (suborder mysticeti). *Mol Biol Evol.* 26: 2427–2440.

Johnson NL, Kotz S. 1977. Urn models and their application. New York: Wiley.

Jukes TH, Cantor RC. 1969. Evolution of protein molecules. In: Mammalian protein metabolism. New York: Academic. p. 21–132.

Kingman JFC. 1982a. The coalescent. *Stoch Process Appl.* 13:235–248.

Kingman JFC. 1982b. On the genealogy of large populations. *J Appl Probab.* 19A:27–43.

Kohl J, Paulsen I, Laubach T, Radtke A, von Haessler A. 2006. Hvr-Base++: a phylogenetic database for primate species. *Nucleic Acids Res.* 34:D700–D704.

Kuhner MK, Smith LP. 2007. Comparing likelihood and Bayesian coalescent estimation of population parameters. *Genetics* 175: 155–165.

Kuhner MK, Yamato J, Felsenstein J. 1995. Estimating effective population size and mutation rate from sequence data using Metropolois-Hastings sampling. *Genetics* 140:1421–1430.

Kuhner MK, Yamato J, Felsenstein J. 1998. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* 149:429–434.

Larkin MA, Blackshields G, Brown NP, et al. (13 co-author). 2007. Clustal w and clustal x version 2.0. *Bioinformatics* 23:21, 2947–2948.

Leman SC, Chen Y, Stajich JE, Noor MA, Uyenoyama MK. 2005. Likelihoods from summary statistics: recent divergence between species. *Genetics* 171:1419–1436.

Minin VN, Bloomquist EW, Suchard MA. 2008. Smooth Skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol.* 25:1459–1471.

Möhle M. 1998. Robustness results for the coalescent. *J Appl Probab.* 35:438–447.

Moreau CS, Bell CD, Vila R, Archibald SB, Pierce NE. 2006. Phylogeny of the ants: diversification in the age of Angiosperms. *Science* 312:101–104.

Nee S, Holmes EC, Rambaut A, Harvey PH. 1995. Inferring population history from molecular phylogenies. *Philos Trans R Soc Lond Ser B.* 349:25–31.

Nordborg M, Krone SM. 2002. Separation of time scales and convergence to the coalescent in structured populations. In: Slatkin M, Veuille M, editors. Modern developments in theoretical population genetics: the legacy of Gustave Malécot. Oxford: Oxford University Press. p. 194–232.

Pluzhnikov A, Donnelly P. 1996. Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* 144:1247–1262.

Polanski A, Kimmel M, Chakraborty R. 1998. Application of a time-dependent coalescence process for inferring the history of population size changes from dna sequence data. *Proc Natl Acad Sci U S A.* 95:10, 5456–5461.

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol.* 116:1791–1798.

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

Rauch EM, Bar-Yam Y. 2005. Estimating the total genetic diversity of a spatial field population from a sample and implications of its dependence on habitat area. *Proc Natl Acad Sci U S A.* 102: 9826–9829.

Rosset S, Wells S, Soria-Hernanz DF, Tyler-Smith C, Royyuru AK, Behar DM. 2008. Maximum likelihood estimation of site-specific mutation rates in human mitochondrial DNA from partial phylogenetic classification. *Genetics* 180:1511.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4:406–425.

Sigurdardottir S, Helgason A, Gulcher JR, Stefansson K, Donnelly P. 2000. The mutation rate in the human mtDNA control region. *Am J Hum Genet.* 66:1599–1609.

Sjödin P, Kaj I, Krone S, Lascoux M, Nordborg M. 2005. On the meaning and existence of an effective population size. *Genetics* 169:1061–1070.

Slatkin M, Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129:555–562.

Sneath PHA, Sokal RR. 1973. Numerical taxonomy, the principles and practice of numerical classification. San Francisco (CA): W. H. Freeman.

Sokal RR, Michener CD. 1958. A statistical method for evaluating systematic relationships. *Univ Kansas Sci Bull.* 38:1409–14384.

Stadler T. 2008. Lineages-through-time plots of neutral models for speciation, *Math. Biosci* 216:163–171.

Stephens M. 2001. Inferences under the coalescent. In: Balding DJ, Bishop MJ, Cannings C, editors. Handbook of statistical genetics. Chichester, UK: John Wiley & Sons. p. 213–238.

Strimmer K, Pybus OG. 2001. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Mol Biol Evol.* 18:2298–2305.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.

Tavaré S. 2004. Ancestral inference in population genetics. In: Cantoni O, Tavaré S, Zeitouni O, editors. École d'Été de Probabilités de Saint-Flour XXXI—2001. Lecture Notes in Mathematics, vol. 1837. Berlin: Springer-Verlag. p. 1–88.

Tavaré S, Balding DJ, Griffiths RC, Donnelly P. 1997. Inferring coalescence times from DNA sequence data. *Genetics* 145:505–518.

Wakeley J. 2008. Coalescent theory: an introduction. Greenwood Village (CO): Roberts & Company Publishers.

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

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

Weiss G, von Haeseler A. 1998. Inference of population history using a likelihood approach. *Genetics* 149:1539–1546.

Wright S. 1931. Evolution in Mendelian populations. *Genetics* 16:97–159.

Supplementary Material for Maruvka, Shnerb, Bar-Yam, and Wakeley

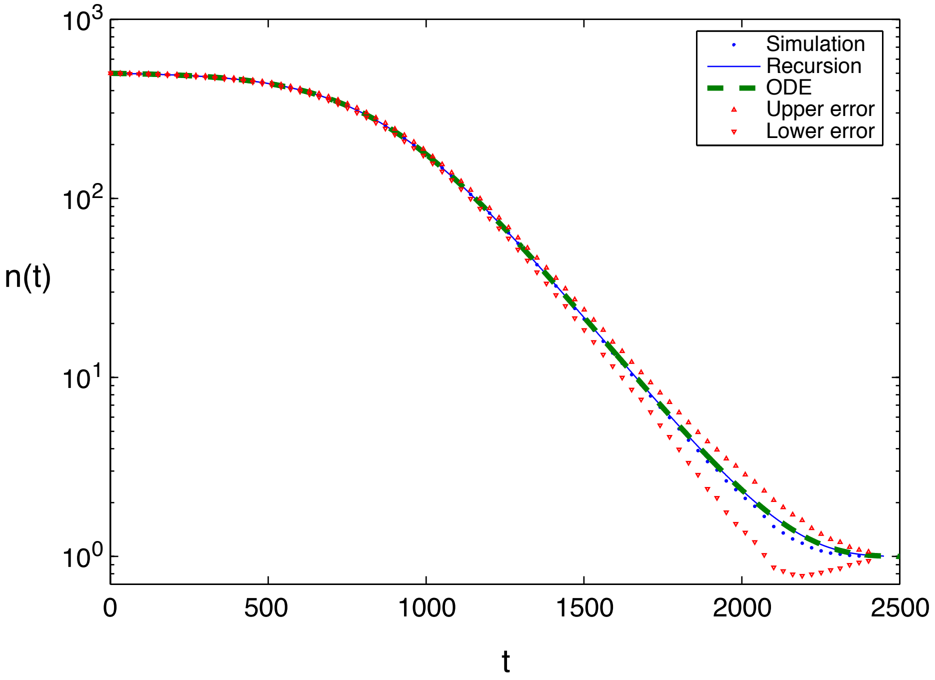
NLFT for an exponentially growing population

Eq. 11 of the main text governs the deterministic evolution of the NLFT in the case of an exponentially growing population. This ordinary differential equation takes into account terms of order $\mathcal{O}(\backslash(\sqcup)/\mathcal{N}(\sqcup))$, as in the case of a fixed population. When these small terms are neglected, the NLFT is:

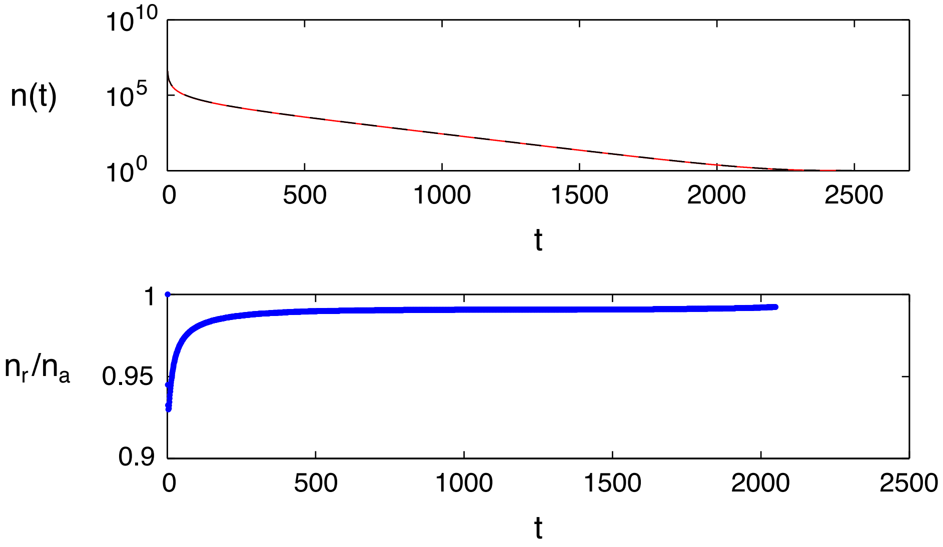
$$n(t) = \frac{2N_0\gamma n_0}{2N_o\gamma + n_0(\exp(\gamma t) - 1)}. \tag{1}$$

As long as $n(t)$ is large there is no difference between the expression 1 and the result of equation 11 in the main text. However, for long times when $n(t)$ becomes smaller the difference becomes apparent. Of course Eq. 11 yields a better description of the average number of lineages as a function of time. However, the expression (1 fits better the average time of the $i'th$ coalescent event (including the T_{MRCA}), and thus it has been used in figure 9 of the main text.

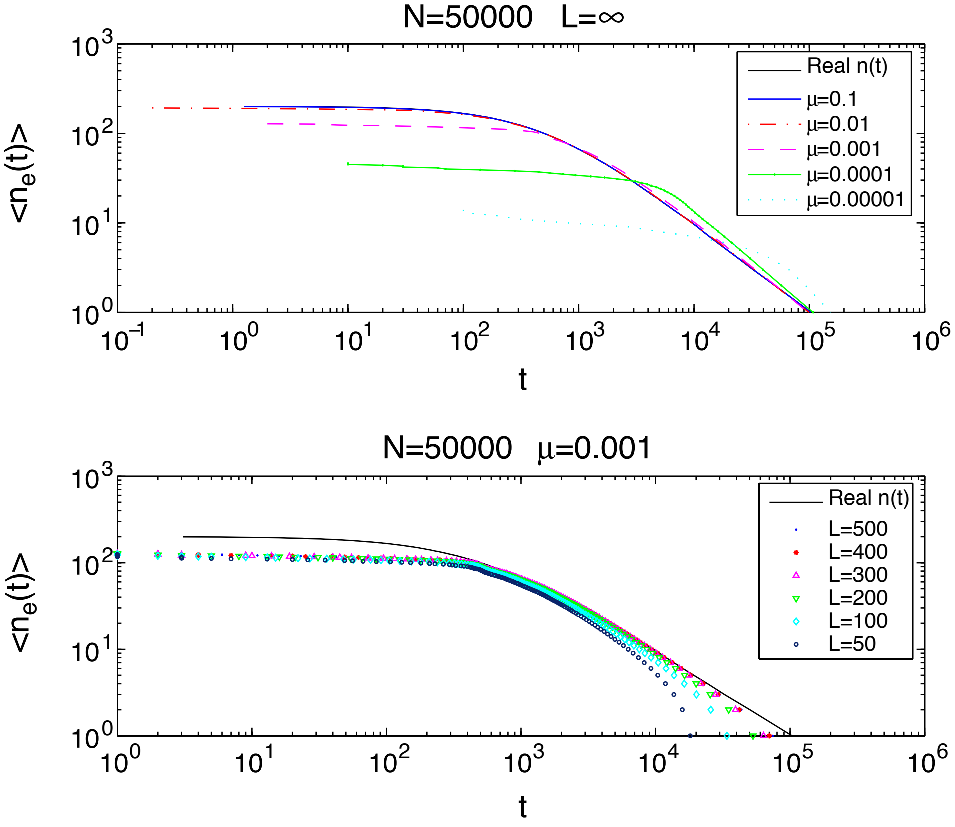
Supplementary figures



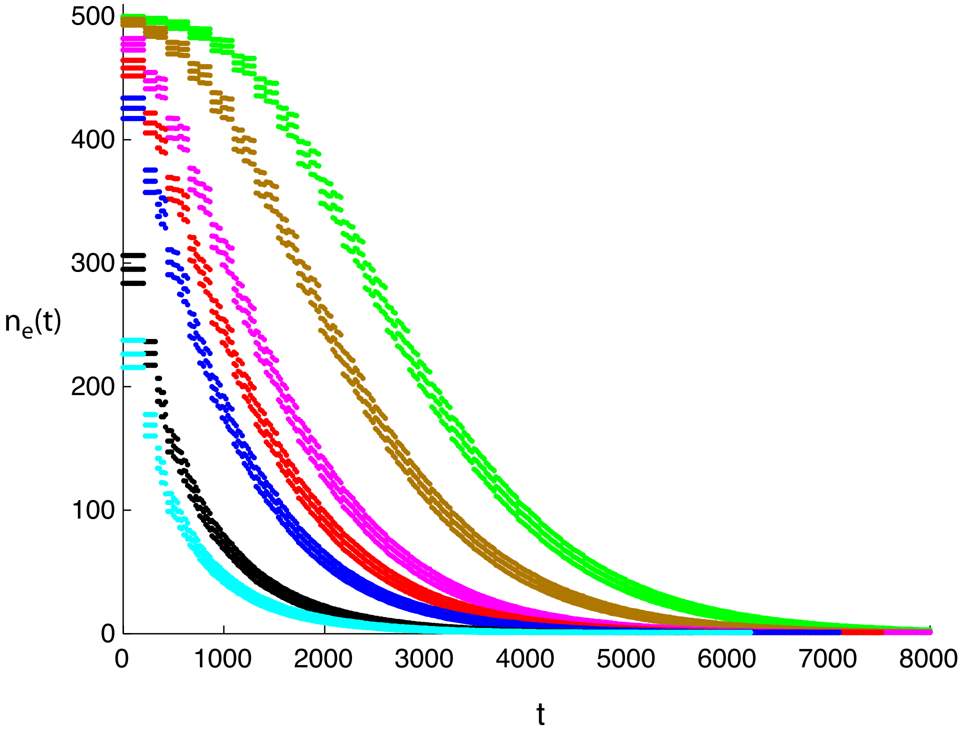
SUPPLEMENTARY FIG. S1—Similar to figure 1 in the main text, but here for a growing population, with parameters $N_0 = 5 \times 10^5$, $\gamma = 0.005$ and $n_0 = 500$.



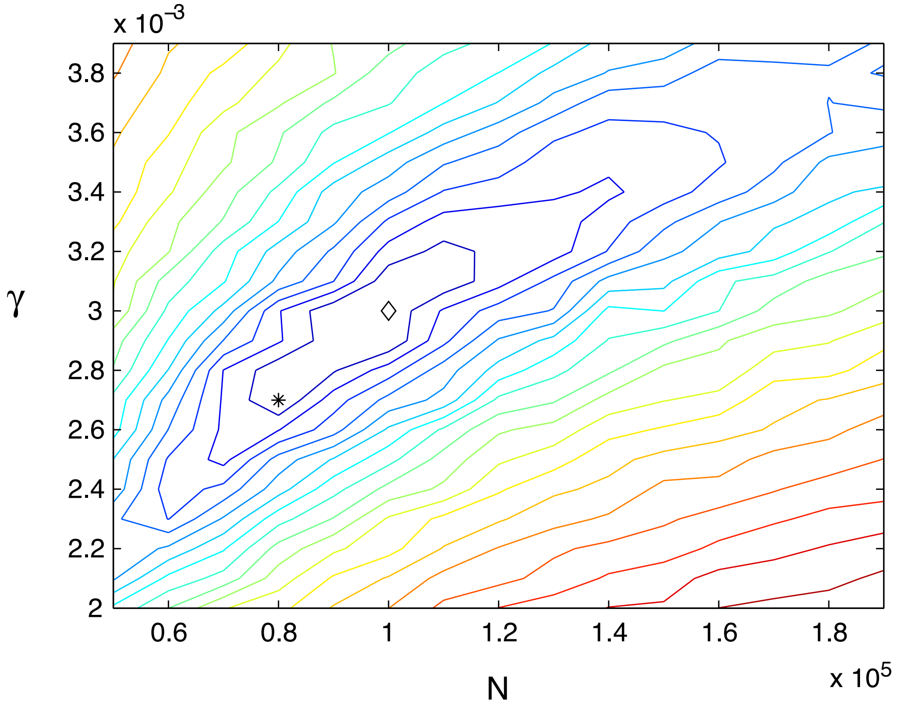
SUPPLEMENTARY FIG. S2—The upper panel shows the NLFT from the recursion equation (red line) and the continuum approximation (black dashed line), for the full sample $n_0 = N_0$. The lower panel, which shows the ratio of the two, illustrates that this difference is always less than about 5%. In both panels, $N_0 = 4000000$ and $\gamma = 0.005$.



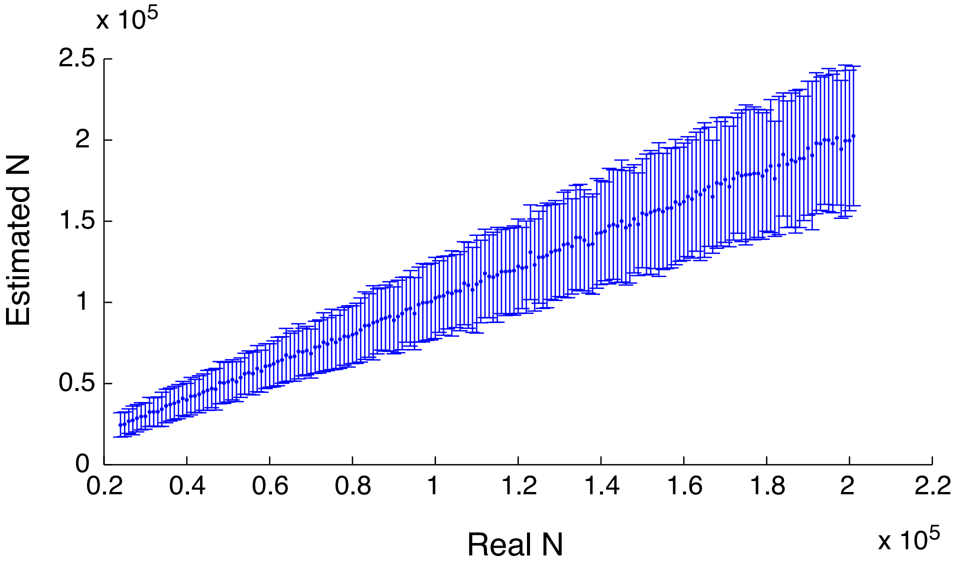
SUPPLEMENTARY FIG. S3—Dependence of the average estimated NFLT, $\langle n_e(t) \rangle$ on the mutation rate (upper panel) and the sequence length (lower panel) for a constant-sized population. Each line is the average of 500 simulation replicates.



SUPPLEMENTARY FIG. S4—The average ± one standard deviation of the NLFT, $n_e(t)$, estimated from simulated data, for a series of different current population sizes N_0. The populations sizes are 5×10^4 (turquoise), 10^5 (black), 5×10^5 (blue), 10^6 (red), 2×10^6 (magenta), 10^7 (brown) and 4×10^7 (green). Other parameters are $\gamma = 0.002$, $n_0 = 500$ and $\mu = 0.0023$.



SUPPLEMENTARY FIG. S5—A contour plot of surface of squared deviations over the demographic parameter space. The simulation runs with growth rate $\gamma = 0.003$. When the size of the population reaches $N = 10^5$, $n_0 = 50$ individuals were sampled and their genetic sequences were used to establish $n_e(t)$ using the WPGM algorithm. The color scale is from red (larger deviations) to blue (smaller deviations). The black star is the location of the global minimum, *i.e.* this gives us the estimated demographic parameters ($N = 8 \times 10^4,\ \gamma = 0.0027$). The black triangle marks the location of the real parameters ($N = 10^5,\ \gamma = 0.003$).



SUPPLEMENTARY FIG. S6—Estimated size of the population (middle line) and its standard deviation plotted against the real N.