DOI: 10.1534/genetics.104.030171

Maximum-Likelihood Estimation of Demographic Parameters Using the Frequency Spectrum of Unlinked Single-Nucleotide Polymorphisms

Alison M. Adams*,[1] and Richard R. Hudson†

*Committee on Genetics and †Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637

Manuscript received April 17, 2004
Accepted for publication July 13, 2004

ABSTRACT

A maximum-likelihood method for demographic inference is applied to data sets consisting of the frequency spectrum of unlinked single-nucleotide polymorphisms (SNPs). We use simulation analyses to explore the effect of sample size and number of polymorphic sites on both the power to reject the null hypothesis of constant population size and the properties of two- and three-dimensional maximum-likelihood estimators (MLEs). Large amounts of data are required to produce accurate demographic inferences, particularly for scenarios of recent growth. Properties of the MLEs are highly dependent upon the demographic scenario, as estimates improve with a more ancient time of growth onset and smaller degree of growth. Severe episodes of growth lead to an upward bias in the estimates of the current population size, and that bias increases with the magnitude of growth. One data set of African origin supports a model of mild, ancient growth, and another is compatible with both constant population size and a variety of growth scenarios, rejecting greater than fivefold growth beginning >36,000 years ago. Analysis of a data set of European origin indicates a bottlenecked population history, with an 85% population reduction occurring ~30,000 years ago.

PATTERNS of genetic variation in contemporary populations can be used to make inferences about past population size changes. Ideally, likelihood methods using the full data would be applied to make such inferences. For the case of DNA sequence polymorphism and where no recombination occurs between the variable sites, methods are available for carrying out such inferences (Beerli and Felsenstein 2001; Kuhner *et al.* 1998; Nielsen 1999). With incomplete linkage between sites, such approaches are frequently computationally infeasible. An exception is the case in which only two chromosomes are sampled at each locus, where Marth *et al.* (2003) have shown that maximum-likelihood methods are feasible. These computational difficulties have led to the use of summary statistics such as Tajima's *D* (Tajima 1989) for making inferences about past demography. For example, Wall and Przeworski (2000) and Pluzhnikov *et al.* (2002) tested compatibility between observed values of Tajima's *D* and values observed in simulations under constant-size and alternative demographic scenarios. Weiss and Von Haeseler (1998) also focused on summaries of the data by implementing a likelihood approach based on mean pairwise differences and segregating sites for a model of complete linkage.

With free recombination between sites, the problem is greatly simplified. In this case, sites can be considered to be statistically independent of each other and the data are completely characterized by the number of polymorphic sites and the frequency spectrum. That is, we can represent the full data by $\mathbf{m} = (m_0, m_1, m_2, m_{n-1})$, where m_0 is the number of sites monomorphic in the sample, and, for $i > 0$, m_i is the number of polymorphic sites in which the derived allele is present i times in the sample of n chromosomes. We assume all polymorphic sites are biallelic. Then, $\Sigma_{i=0}^{n-1} m_i$ is L, the number of sites surveyed, and $\Sigma_{i=1}^{n-1} m_i$ is the total number of segregating sites in the sample, S. Also note that $m_0 = L - S$. In this case of free recombination between sites, full-likelihood approaches are computationally undemanding. This case has been examined by Wooding and Rogers (2002), Polanski and Kimmel (2003), and Marth *et al.* (2004) and is also the focus of our study. We examine the statistical properties of demographic inferences based on **m**, using maximum likelihood and assuming sites are independent. By utilizing the entire frequency spectrum, **m**, rather than a summary statistic such as Tajima's *D*, this approach captures all available information in data sets consisting of unlinked polymorphic sites.

With linkage between sites, there is a statistical nonindependence between polymorphic sites, and thus **m** for a set of linked sites would contain less information than that for a set of unlinked sites. It follows that our results for unlinked sites give an idea of the best one can do using **m** or summary statistics, such as Tajima's *D*, which can be calculated from it. It is important to note that for linked sites **m** does not completely characterize the data, and that full likelihood, which incorporates in-

[1]*Corresponding author:* Department of Ecology and Evolution, 1101 E. 57th St., Room Z302, University of Chicago, Chicago, IL 60637. E-mail: alison1@uchicago.edu

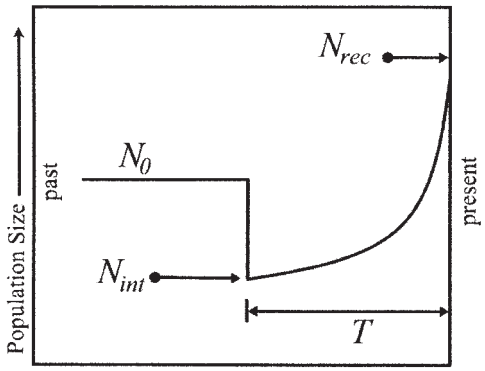


FIGURE 1.—Demographic model. f_{int} (= N_{int}/N_0), f_{rec} (= N_{rec}/N_0), and T are the estimated parameters.

formation about linkage disequilibrium, for instance, might result in better inferences.

The models examined here consist of either exponential growth or an instantaneous decrease followed by exponential growth, which requires simultaneous estimation of either two or three parameters, respectively. We illustrate that, particularly for recent growth scenarios, data sets consisting of large numbers of segregating sites are required to produce good estimates based solely on frequency spectrum data. Our results provide a theoretical perspective on the feasibility of frequency spectrum-based parameter estimation with a modest amount of data, and we present methods to determine the approximate variance and covariance associated with such estimators under any demographic scenario of interest. The maximum-likelihood method is also applied to three human data sets. The first is an African data set consisting of the original data set of FRISSE *et al.* (2001) as well as 40 additional locus pairs (A. DI RIENZO, unpublished data). The Seattle single-nucleotide polymorphisms (SNPs) data, consisting of both African-American and European data sets (http://pga.gs.washington.edu), is also examined. Each of these data sets consists of linked segregating sites within effectively unlinked loci, and a procedure is outlined by which one can extend this method to such data and construct the appropriate confidence regions associated with the estimators.

MODEL AND METHODS

Model: The demographic model considered is that of a population of constant effective size N_0 until time T when there was an instantaneous decrease to an intermediate population size (N_{int}) followed by exponential growth to the current population size (N_{rec}). As illustrated in Figure 1, this model involves four demographic parameters: N_0, N_{int}, N_{rec}, and T, where T is the time at which the instantaneous size change occurred. T is measured in units of $4N_0$ generations before the present. We assume the mutation rate per site, u, is small, so that the occurrence of more than one mutation occurring in the history of the sample at a single site can be ignored. We find it convenient to introduce the parameters $f_{int} = N_{int}/N_0$ and $f_{rec} = N_{rec}/N_0$ and specify the model by $4N_0u$, f_{int}, f_{rec}, and T. This demographic model is flexible and can be generalized to the case of exponential growth with no bottleneck by setting f_{int} equal to one or to the case of a population reduction with no recovery by setting f_{rec} equal to f_{int}. We also assume that the population is unstructured (panmictic) and that the polymorphic sites are unlinked.

Maximum-likelihood method: The maximum-likelihood approach followed here is that of WOODING and ROGERS (2002) and POLANSKI and KIMMEL (2003). Our analyses require a population survey of variation at a set of L unlinked sites. For L unlinked sites, $\mathbf{m}$ is multinomially distributed,

$$\text{Prob}(\mathbf{m}) = \binom{L}{m_0 \;\; m_1 \;\; \ldots \;\; m_{n-1}} \prod_{i=0}^{n-1} P_i^{m_i}, \qquad (1)$$

where P_0 is the probability that a site is monomorphic in the sample, and, for $i > 0$, P_i is the probability that a site is polymorphic with i copies of the derived allele. The P_i's are functions of the four parameters of the demographic model (θ_0, f_{int}, f_{rec}, and T) and the sample size n.

To obtain the maximum-likelihood estimates of the parameters one maximizes the right-hand side of (1). We note, however, that we can write the probability of the data as

$$\text{Prob}(\mathbf{m}) = \binom{L}{m_0 \;\; m_1 \;\; \ldots \;\; m_{n-1}} P_0^{(L-S)} (1 - P_0)^S \prod_{i=1}^{n-1} \frac{P_i^{m_i}}{(1 - P_0)^{m_i}} \qquad (2)$$

$$= \binom{L}{m_0 \;\; m_1 \;\; \ldots \;\; m_{n-1}} P_0^{(L-S)} (1 - P_0)^S \prod_{i=1}^{n-1} p_i^{m_i}, \qquad (3)$$

where p_i (= $P_i/(1 - P_0)$) is the probability that a site is polymorphic with i copies of the derived allele, conditional on the site being polymorphic in the sample. P_0 and the p_i's can be written in terms of θ_0 and mean properties of sample gene trees. For example, for small θ_0, P_0 can be expressed in terms of the mutation parameter and the mean total length of the gene genealogy,

$$P_0 \approx 1 - \theta_0 \tau(n), \qquad (4)$$

where $\tau(n)$ is the mean total length of the gene tree of a sample of n chromosomes measured in units of $4N_0$ generations (HUDSON 1990). We define an i-branch to be a branch of the gene tree such that a mutation that occurs on the branch results in i copies of the mutation in the sample. The mean total length of i-branches in units of $4N_0$ generations we denote by $\tau_i(n)$. Then, P_i is $\sim\theta_0\tau_i(n)$, and

$$p_i \approx \frac{\theta_0 \tau_i(n)}{\theta_0 \tau(n)} = \frac{\tau_i(n)}{\tau(n)}, \quad i > 0. \qquad (5)$$

When time is measured in units of $4N_0$ generations,

$\tau_i(n)$ and $\tau(n)$ are functions of f_{int}, f_{rec}, and T, but do not depend on N_0 or θ_0. Thus, to find the maximum-likelihood estimates of the four parameters, we can first find the maximum-likelihood estimates, $\hat{f}_{\text{int}}$, $\hat{f}_{\text{rec}}$, and $\hat{T}$, by maximizing $\prod_{i=1}^{n-1} p_i^{m_i}$. The maximum-likelihood estimate of θ_0, if desired, can then be obtained as $\hat{\theta}_0 = (S/L)/\hat{\tau}(n)$, where $\hat{\tau}(n)$ is the mean gene tree length with f_{int}, f_{rec}, and T set equal to the maximum-likelihood estimates. In this article, we focus on estimation of the parameters f_{int}, f_{rec}, and T, which requires maximizing $\prod_{i=1}^{n-1} p_i^{m_i}$ and does not require specifying L or m_0. Equivalently, we can consider estimation of the parameters based on the probability of $\mathbf{m}$ conditional on S, which is

$$\begin{aligned}\text{Prob}(\mathbf{m}|S) &= \frac{\text{Prob}(\mathbf{m})}{\text{Prob}(S)} \\ &= \frac{\text{Prob}(\mathbf{m})}{\binom{L}{m_0} P_0^{(L-S)}(1-P_0)^S} \\ &= \binom{S}{m_1 \;\; m_2 \;\; \ldots \;\; m_{n-1}} \prod_{i=1}^{n-1} p_i^{m_i}, \qquad (6)\end{aligned}$$

which does not depend on L or m_0.

To find the maximum-likelihood estimates of f_{int}, f_{rec}, and T, we estimate $\text{Prob}(\mathbf{m}|S)$ at a set of points on a rectangular grid of values in the three-dimensional space of f_{int}, f_{rec}, and T values. For each point in the grid, we estimate the $\tau_i(n)$'s by generating 100,000 replicate gene trees with simple one-site coalescent simulations. The $\tau_i(n)$'s can also be obtained as described elsewhere (Griffiths and Tavaré 1998; Wooding and Rogers 2002; Polanski and Kimmel 2003), and this method can be generalized to any demographic model for which the relevant $\tau_i(n)$'s can be calculated or estimated. From the estimated τ_i's, the p_i's are calculated, which in turn are used to calculate the product, $\prod_{i=1}^{n-1} p_i^{m_i}$. Since we have ignored the problem of estimating θ_0, we do not require L, and the results are all given conditional on specified numbers of polymorphic sites.

Required data: Our analyses require data in the form of unlinked polymorphic sites. Ascertainment bias is not considered in this article, so we assume that sites are randomly chosen with no prior knowledge of polymorphism and sequenced in each sampled chromosome. Frisse *et al.* (2001) sequenced ~25 kb and found 120 segregating sites in an African Hausa sample of 30 chromosomes. If we consider genome-wide polymorphism levels to be similar to that data, then ~104,000 sites would have to be sequenced in 30 chromosomes to assemble a data set consisting of 500 segregating sites. These 104,000 sites could be sequenced in small unlinked segments throughout the genome to obtain frequency spectrum data from unlinked polymorphic sites. We consider only biallelic polymorphic sites and assume that the ancestral/derived status of each allele is known. However, not having knowledge of the ancestral state makes only a minimal difference to our results (data not shown). We also assume that all sites are surveyed in a sample of n chromosomes (or $n/2$ diploid individuals), but more general sampling is easily accommodated. For example, one can separate a data set into a series of frequency spectra, each with a different n. A global likelihood may then be obtained for the entire data set by multiplying the result of Equation 6 for sets of sites with different sample sizes.

Sample size comparison: We compare the effect of sample size and number of unlinked polymorphic sites on both the power to reject the null hypothesis of constant population size and the quality of estimates of specific demographic parameters. The number of segregating sites, when indicated, is scaled on the basis of the average total branch length of a random gene genealogy (τ), which will vary according to sample size and demographic scenario. For example, suppose we wish to compare sample sizes of 50 and 100 chromosomes under a growth scenario where 40-fold expansion occurred beginning 10,000 years ago. In this case $\tau(50)$ (in units of $4N_0$ generations) for a sample size of 50 chromosomes is 4.93, $\tau(100)$ for a sample size of 100 chromosomes is 6.06, and $\tau(100)/\tau(50)$ is 1.23. In words, the average total branch length of a random gene genealogy is 1.23 times greater for a sample size of 100 than for a sample size of 50. This indicates that, for every 500 segregating sites discovered in a sample size of 50, ~615 segregating sites would be found in a sample size of 100 if the same number of sites were sequenced. Thus, when we compare sample size 50 to sample size 100, we compare $n = 50$, $S = 500$ to $n = 100$, $S = 615$. Normalizing the number of sites in this way serves to facilitate comparisons between analyses of different sample size because it takes into account the expected number of polymorphic sites in the samples of each size.

Asymptotic properties: To investigate whether asymptotic approximations of confidence regions and variances of estimators are applicable to data sets of modest size, we first determined whether 95% confidence regions obtained from the log-likelihood ratio have the expected coverage properties. Confidence regions include those points on the grid with a log-likelihood ratio <3 or 3.9 for simultaneous estimation of two or three parameters, respectively. We also compared the observed variances and covariances of the estimators with the approximate variances and covariances calculated by estimating the inverse of the information matrix,

$$I_{ij} = -E\left(\frac{\partial^2}{\partial\xi_i \partial\xi_j} \log L\right). \qquad (7)$$

We estimate the expected log-likelihood with

$$E(\log L) = \sum_{i=1}^{n-1} P_i(f_{\text{rec}_0}, f_{\text{int}_0}, T_0) \log(P_i(f_{\text{rec}}, f_{\text{int}}, T)) \qquad (8)$$

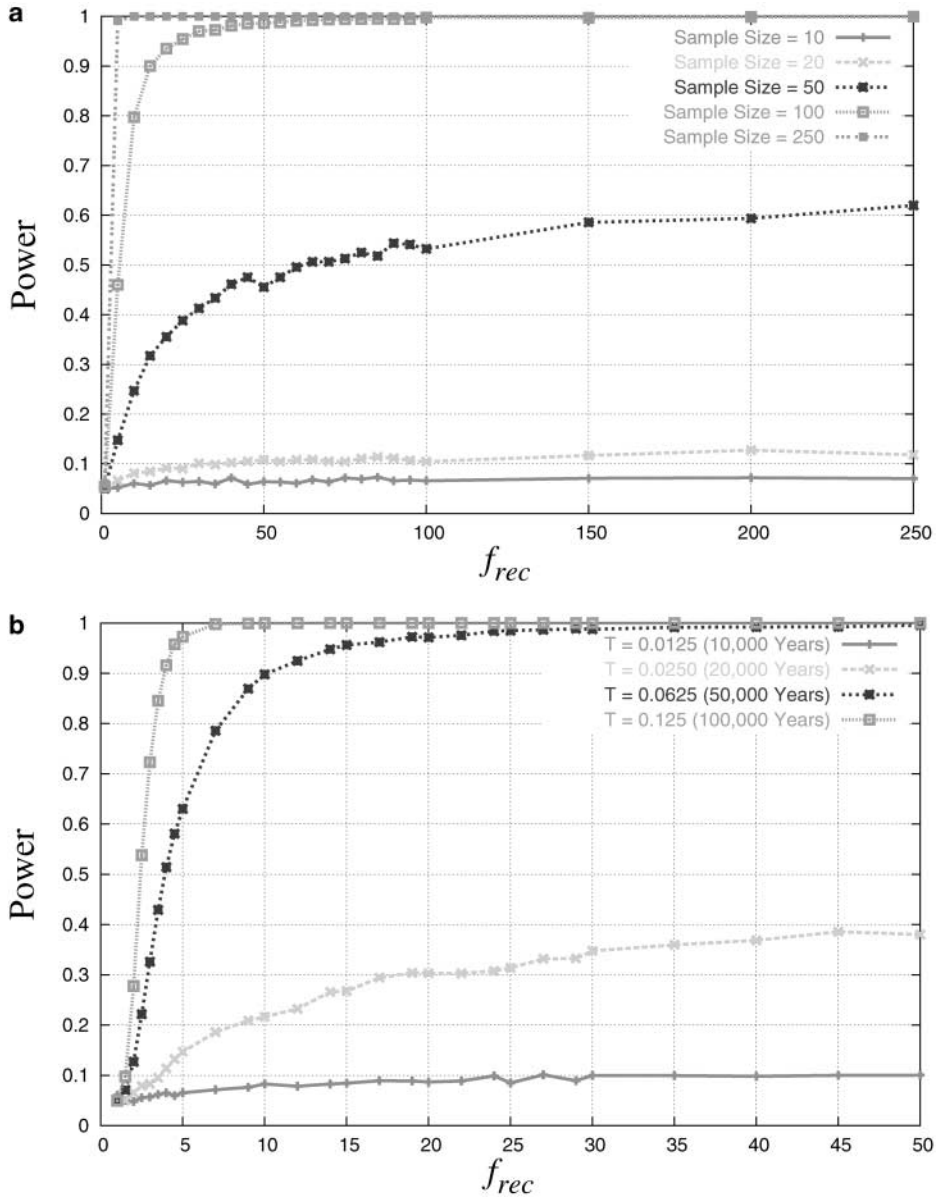


FIGURE 2.—Power to detect growth with ~500 unlinked sites. The number of sites used for each point in a curve is scaled on the basis of 500 sites in a sample size of 20. (a) Effect of sample size on power to detect recent growth beginning 10,000 years ago ($T = 0.0125$). (b) Effect of the onset time of growth on power to detect growth with a sample size of 20.

with the $P_i(f_{rec_0}, f_{int_0}, T_0)$ estimated by coalescent simulation for a set of points on a narrow grid of f_{rec}, f_{int}, and T values around the true parameter values of f_{rec_0}, f_{int_0}, and T_0. The second partial derivatives relevant to the information matrix are then approximated from best-fit second-degree polynomial curves. The variance and covariance observed from simulation can then be compared to the appropriate terms of the inverse of I_{ij} to determine whether the asymptotic approximations apply to data sets of modest size.

RESULTS

Accuracy and precision of estimated τ_i's: As described in MODEL AND METHODS, we estimate the relevant $\tau_i(n)$'s from 100,000 replicate gene trees generated by one-site coalescent simulations. We find that our $\tau_i(n)$'s, estimated from simulation, are in very close agreement to $\tau_i(n)$'s calculated numerically by the method of POLANSKI and KIMMEL (2003). For the case of $f_{rec} = 2.0$, $f_{int} = 0.15$, and $T = 0.0375$ for a sample size of 46, the maximum-likelihood parameters for the Seattle SNPs European data set, we find that our simulated $\tau_i(46)$'s differ, at most, by 0.05% from the calculated $\tau_i(46)$'s. Additionally, we calculate the log-likelihood of the Seattle SNPs European data set using 10 independent $\tau_i(46)$ estimates, each resulting from 100,000 replicate gene trees, and find that the likelihoods calculated from our simulated $\tau_i(46)$'s differ little between trials, ranging from −11987.278 to −11987.302. Since the log-likelihood ratio critical values relevant for our construction of confidence regions range from 3.86 to 9.1, such a negligible fluctuation would not affect our inferred acceptance regions.

TABLE 1

Evaluation of asymptotic confidence region

	Sample size = 50	Sample size = 100
1,000 sites	0.02332	0.01911
5,000 sites	0.03774	0.06186
10,000 sites	0.03750	0.05155
20,000 sites	0.06122	0.05096

Proportion of simulations where the log-likelihood ratio lies outside the two-dimensional asymptotic confidence region (log-likelihood ratio >3) is shown. Each value is based on 5000 repetitions with parameter values of $f_{rec} = 5$, $f_{int} = 0.5$, and $T = 1$.

TABLE 2

Asymptotic and simulated variance of two-dimensional estimators

	Variance			
	$\hat{f}_{int}$	$\hat{T}$	Covariance	Correlation
Asymptotic				
1,000 sites	0.03111	0.08806	−0.01804	−0.34
5,000 sites	0.006223	0.017612	−0.003608	−0.34
10,000 sites	0.003112	0.008806	−0.001804	−0.34
20,000 sites	0.001556	0.004403	−0.000902	−0.34
Simulated				
1,000 sites	0.016781	0.056648	−0.00745	−0.24
5,000 sites	0.005966	0.014237	−0.002720	−0.3
10,000 sites	0.002367	0.007119	−0.001810	−0.44
20,000 sites	0.001580	0.004182	−0.001040	−0.4

Asymptotic variance is obtained from the estimated information matrix as described in the text (Equations 7 and 8). Results are based on a demographic scenario of $f_{rec} = 5.0$, $f_{int} = 0.5$, and $T = 1.0$ and a sample size of 50 chromosomes. We assume f_{rec} is known and sites are unlinked.

Power curves: Power analyses were conducted using a chi-square test with $n - 2$ d.f. for a sample size of n chromosomes to determine the degree of growth that would be required to reject the null hypothesis of constant population size using only frequency spectrum information. For smaller numbers of segregating sites, the degrees of freedom may vary slightly, as frequency categories with expected site counts of less than five are collapsed. The expected frequency spectrum under the null hypothesis is calculated from

$$p_i = \frac{1/i}{\sum_{j=1}^{n-1}(1/j)}, \quad 1 \le i \le n-1 \tag{9}$$

(Ewens 1979) by multiplying each p_i by the number of segregating sites. The observed frequency spectrum is obtained by estimating the p_i's from 100,000 replicates for each combination of f_{int}, f_{rec}, and T values and then multinomially sampling from these simulated p_i's. For each sample size, the number of segregating sites at each f_{rec} value is scaled on the basis of 500 polymorphic sites in a sample size of 20, as described in MODEL AND METHODS.

Recent growth beginning 10,000 years ago: Figure 2a shows power curves for the scenario of recent growth beginning at $T = 0.0125$ (which, for humans, would correspond to 10,000 years ago on the basis of a generation time of 20 years and N_0 of 10,000, roughly corresponding to the advent of agricultural society). We consider sample sizes of 10, 20, 50, 100, and 250 chromosomes with 500 polymorphic sizes (scaled as described above), which, for a sample size range of 10–250, corresponds to a range of 398–859 sites at constant population size ($f_{rec} = 1$) to 390–1137 sites at the most extreme growth scenario considered ($f_{rec} = 250$). With sample sizes of ≤20, the power to reject the null hypothesis of constant population size never exceeds 0.15, even with 500-fold growth. With a sample size of 50, power reaches ~0.5 with 50-fold growth, but barely rises above 0.6 at the largest magnitude of growth considered. As sample size reaches 100, one can reliably detect 20-fold growth, and a sample size of 250 allows for a power near 1 to reject the null hypothesis with only 5-fold growth (Figure 2a). It is clear that recent rapid growth can be reliably detected with frequency spectrum data only with fairly large samples (>100 chromosomes), and the most modest growth scenarios may be detected only with samples consisting of at least 250 chromosomes when data sets consist of only 500 (scaled) polymorphic sites.

More ancient growth onset: Power to reject the constant size hypothesis is also dependent upon the time that exponential growth begins, as illustrated in Figure 2b. While power is minimal for small sample sizes with growth beginning 10,000 years ago, power increases dramatically with more ancient growth. For example, while a sample size of 20 with 500 polymorphic sites yields virtually no power to detect any magnitude of growth beginning 10,000 years ago, if growth instead began 50,000 years ago, a sample size of 20 with the same number of sites would be sufficient to reliably detect 10-fold growth.

Asymptotic properties: We evaluate the distribution of our maximum-likelihood estimates to determine whether asymptotic theory provides an adequate approximation of the 95% confidence regions and variance associated with our parameter estimates. Table 1 illustrates the proportion of maximum-likelihood estimates for which the true value of the parameters lies outside the asymptotic 95% confidence region. Our simulations indicate that for large amounts of data, asymptotic theory does provide a good approximation of the 95% confidence region for the demographic scenario examined. For smaller amounts of data, the asymptotic approximation appears to be conservative, with the true parameter values lying within the 95% confidence region in ~97–98% of the runs. We also examine a specific case corresponding to

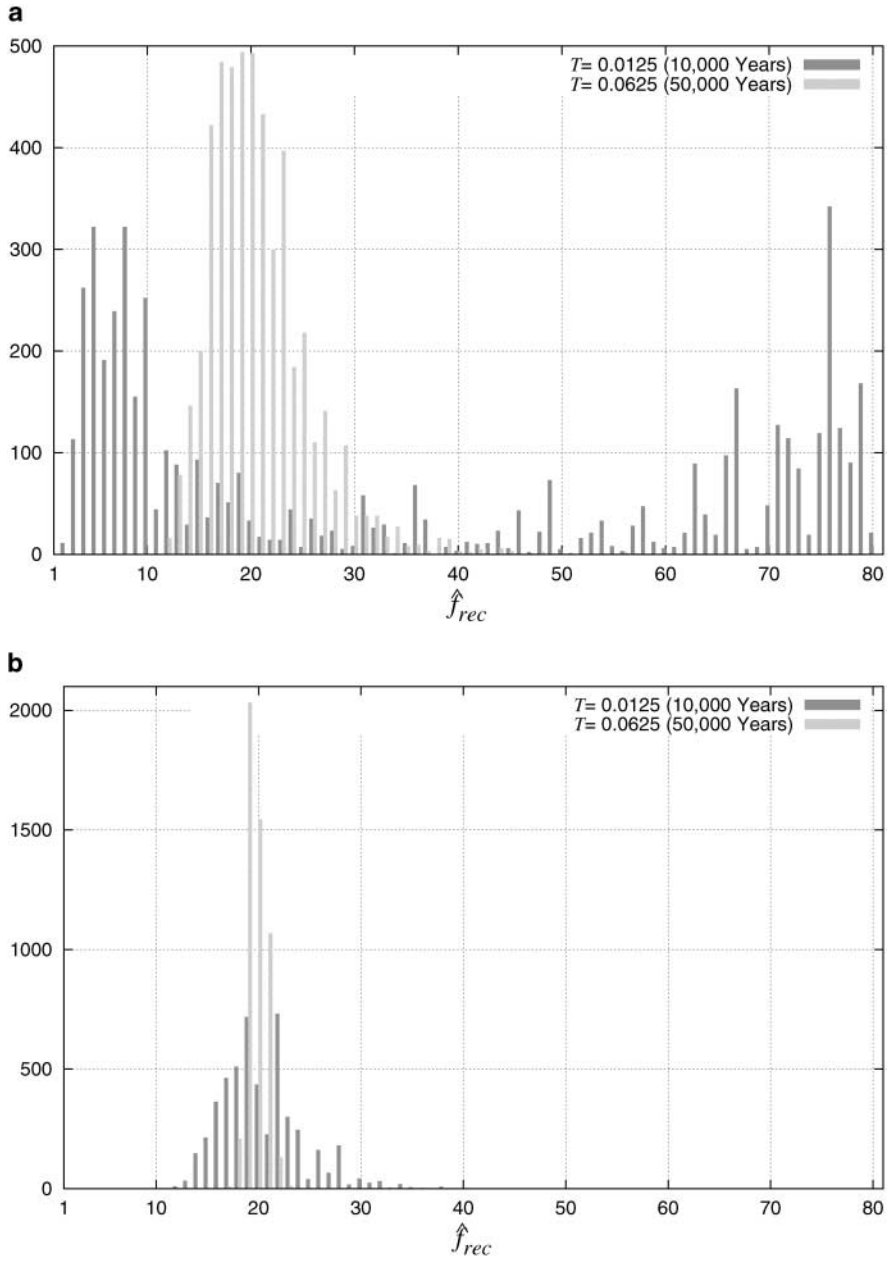


FIGURE 3.—Distribution of $\hat{f}_{rec}$ estimates. Histograms are based on 5000 simulated data sets where f_{rec} = 20 and f_{int} is fixed at 1. (a) 50 chromosomes, 10,000 sites; (b) 250 chromosomes, 16,244 sites (T = 0.0125) and 20,322 sites (T = 0.0625).

the Hausa data set, which consists of 597 sites in a sample size of 30. For a data set of this size [simulated from the Hausa maximum-likelihood estimate (MLE)], we find that asymptotic approximation is especially conservative, rejecting the true parameter values in only 2.04% of the 5000 simulated data sets.

We also determine the variance and covariance of our maximum-likelihood estimators in two dimensions by both asymptotic theory and simulation (Table 2). In this analysis, we assume that N_{rec} is known and is fivefold $>N_0$ (f_{rec} = 5), while f_{int} and T are jointly estimated. For this demographic scenario, asymptotic theory provides a good approximation for the simulated variance and covariance only when the data set consists of a large number of segregating sites.

Quality of estimators: We evaluate the quality of our maximum-likelihood estimators by examining the distribution of the estimates under both two-dimensional and three-dimensional models.

Two-dimensional estimators: A recent growth scenario was examined in which the population size was constant until exponential growth occurred beginning 10,000 years ago. Data sets were simulated with f_{rec} ranging from 10- to 320-fold growth, f_{int} fixed at 1, and T = 0.0125 (10,000 years ago based on a generation time of 20 years and N_0 of 10,000). For this scenario of recent growth, a sample size of at least 250 chromosomes with ~16,000 segregating sites is required for 90% of the distribution of $\hat{f}_{rec}$ to lie within a factor of four of the true f_{rec} value for all magnitudes of growth examined. This is illustrated in Figure 3, which compares the $\hat{f}_{rec}$ distribution under this recent growth scenario for sample sizes of 50 and 250 for

TABLE 3

Distribution of $\hat{T}$

f_{rec}	Mean	SD	0.025	0.5	0.975
10	0.015254	0.007262	0.0077	0.0125	0.0365
20	0.014790	0.005698	0.0089	0.0125	0.0269
40	0.014647	0.004850	0.0089	0.0125	0.0281
80	0.014118	0.003349	0.0101	0.0125	0.0221
160	0.013706	0.002479	0.0101	0.0137	0.0197
320	0.013331	0.001947	0.0101	0.0125	0.0173

Time of expansion is ~10,000 years ($T = 0.0125$), and f_{int} is fixed at 1. Simulated data sets consist of 5000 unlinked sites in a sample size of 50. The $\hat{T}$ grid includes 40 grid points from $\hat{T} = 0.1(T)$ to $\hat{T} = 4(T)$.

20-fold growth. As the magnitude of growth increases, $\hat{f}_{rec}$ becomes biased more severely upward. This result is similar to that obtained for growth beginning at $T = 0.0625$ (Table 4).

The estimates of T, however, are not subject to the upward bias seen in $\hat{f}_{rec}$. Instead, estimates of T are improved as the magnitude of growth increases (Table 3). Sample size also has a dramatic effect on the distribution of $\hat{T}$, as illustrated in Figure 4, which compares the $\hat{T}$ distribution for sample sizes of 50 and 250. With 5000 sites in a sample size of 50 chromosomes, 95% of $\hat{T}$ estimates lie within a factor of 3 of the true T value for 10-fold growth, with 95% of the distribution lying within a factor of 1.5 for 320-fold growth, the most severe growth scenario examined.

We explored another growth scenario in which the onset of growth was more ancient, beginning 50,000 years ago. In this case, the quality of the estimators improved, and 90% of the $\hat{f}_{rec}$ distribution was within a factor of four of the true f_{rec} value for a sample size of 50 with 5000 sites (as opposed to a sample size of 250 and ~16,000 sites under the more recent growth scenario). Figure 3 reveals the improvement in the $\hat{f}_{rec}$ estimator with the more ancient time of growth onset. As in the recent growth scenario, increasing the degree of growth both increased the bias and widened the quantiles of the $\hat{f}_{rec}$ distribution (Table 4). The T estimates under the more ancient growth scenario were also improved over the analogous recent growth estimates, with 95% of the $\hat{T}$ distribution within a factor of two of the true T value for all magnitudes of growth examined with data sets as small as a sample size of 50 with 1000 sites.

Three-dimensional estimators: We consider a three-dimensional model of a constant-sized population that experienced an instantaneous decrease to 0.05 times its initial size 100,000 years in the past, followed by exponential growth until the present to a final size of 5 times the initial population size ($f_{rec} = 5$; $f_{int} = 0.05$; $T = 0.125$). All three parameters were estimated for 5000 simulated data sets. Under this model, 90% of the distribution of each of the three estimators falls within a factor of 4 of the respective true values with data sets as small as 500 sites in a sample size of 30 (Table 5). If a large data set consisting of 10,000 polymorphic sites in 50 chromosomes were available, 95% of the estimates of all three parameters would lie within a factor of 1.5 of the true parameter values. As would be expected, estimates of any of the three parameters are improved by fixing one of the parameters at its true value (data not shown), indicating that incorporation of prior knowledge of one of the parameters would be beneficial.

Applications: We apply the maximum-likelihood method to data obtained from an African Hausa population (A. Di Rienzo, unpublished data) as well as to the African-American and European (CEPH) samples of the Seattle SNPs data set (http://pga.gs.washington.edu).

Hausa data: The Hausa data set consisted of the data of Frisse *et al.* (2001) in conjunction with additional unlinked locus pairs (A. Di Rienzo, unpublished data), which resulted in a data set consisting of 30 chromosomes and 597 polymorphic sites in an African sample, the Hausa of Cameroon. The sites in this data set include linked polymorphic sites within 50 effectively unlinked loci, but in the maximum-likelihood analysis we treat each site as though it provides independent information. As seen in Table 6, $\hat{f}_{rec} = 3.1$, $\hat{f}_{int} = 1$, and $\hat{T} = 6.1$ for this data set.

We perform a χ^2 goodness-of-fit test on the Hausa data set to determine whether the maximum-likelihood parameters can be accepted as an explanation of the Hausa data. However, this test assumes that each site is independent, which is not the case for this data set. Because the linkage between sites will affect the 95% critical value of the χ^2 test statistic, we determine the critical value of the test statistic distribution for this data set by coalescent simulation with recombination (Hudson 1983, 2002). We simulate 5000 data sets, each consisting of 30 chromosomes and 50 unlinked loci. The input parameters for the simulation included Watterson's estimate of θ (estimated to be 0.0012/bp), the recombination rate (estimated to be 5.99×10^{-4}/bp), the average locus length (10,286 bp), and a gene-conversion to crossing-over ratio of 2. Polymorphic sites within the middle 8000 bp were ignored to mimic the locus pair data (Frisse *et al.* 2001). Because the ancestral/derived status of each allele was not considered, each simulated frequency spectrum was folded at frequency 0.5 prior to performing the χ^2 goodness-of-fit test. On the basis of these simulations, we found the 95% critical value of the χ^2 test statistic to be 39.39, as opposed to a critical value of 23.68 (14 d.f.) if all sites were independent. The χ^2 goodness-of-fit test statistic for the Hausa data set under its maximum-likelihood estimate of $\hat{f}_{rec} = 3.1$, $\hat{f}_{int} = 1$, and $\hat{T} = 6.1$ is 26.30 ($P = 0.304$), indicating that this scenario cannot be rejected at the 0.05 significance level. Note, however, that this demographic scenario would have been rejected without properly accounting for the linkage within the data set.

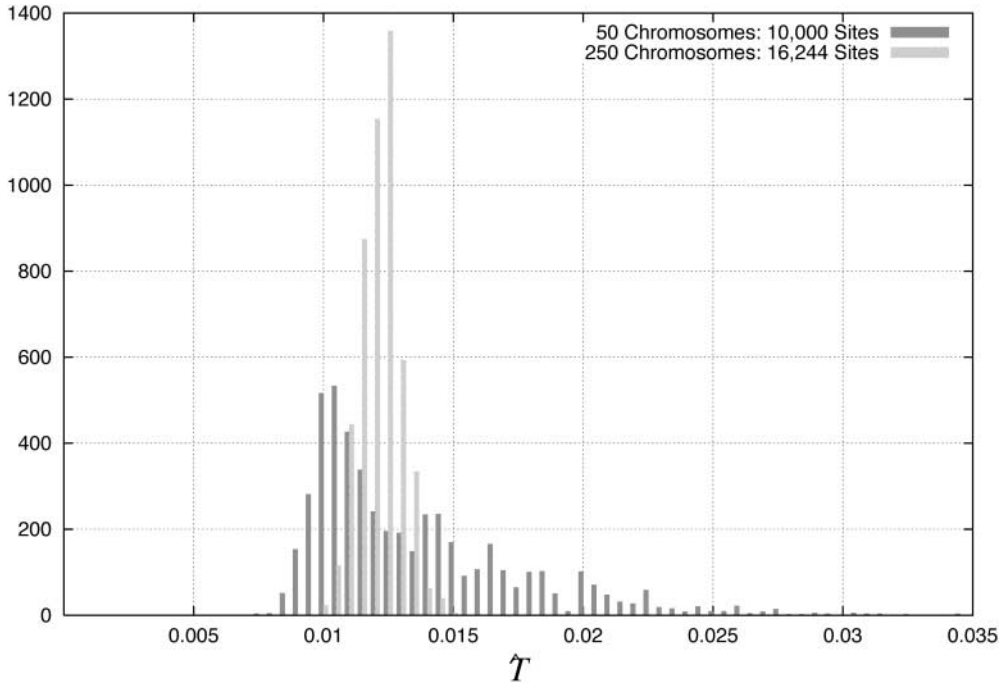


FIGURE 4.—Distribution of $\hat{T}$ estimates. Histograms are based on 5000 simulated data sets with parameters $f_{rec} = 20$, $f_{int} = 1$ (fixed), $T = 0.0125$, each consisting of 10,000 polymorphic sites in 50 chromosomes or 16,244 polymorphic sites in 250 chromosomes.

We also considered the equilibrium model of constant population size for this data set and obtained a χ^2 goodness-of-fit test statistic of 29.24 ($P = 0.204$). On the basis of an analysis of 10 locus pairs (a subset of the 50 locus pairs examined here), FRISSE *et al.* (2001) also concluded that the Hausa data are consistent with the equilibrium model.

Seattle SNPs: We also examined both the African-American and the European samples of the Seattle SNPs data, which are located at the University of Washington-Fred Hutchinson Cancer Research Center (UW-FHCRC) Variation Discovery Resource (http://pga.gs.washington.edu). These data sets consist of 12,587 and 7712 total SNPs across 138 loci in the African-American and European samples, respectively. We considered only those SNPs that were sequenced in the entire panel of 48 (African-American) or 46 (European) chromosomes to facilitate evaluation of confidence regions and goodness-of-fit by simulation. Additional analyses incorporating more of the SNPs are described in the DISCUSSION. The frequency spectrum of nonsynonymous SNPs has been shown to differ from that of synonymous SNPs (CARGILL *et al.* 1999; FAY *et al.* 2001; WOODING and ROGERS 2002), so we also removed all SNPs that result in an amino-acid coding change to minimize the inclusion of those SNPs subject to nonneutral evolutionary processes. This resulted in a final data set of 5892 SNPs for the African-American data set and 4211 SNPs for the European data set. We applied our maximum-likelihood method to these data sets, treating all SNPs as unlinked, and found that the three-dimensional maximum-likelihood estimates of f_{rec}, f_{int}, and T are $\hat{f}_{rec} = 1.9$, $\hat{f}_{int} = 1$, and $\hat{T} = 0.27$ for the African-American data set and $\hat{f}_{rec} = 2$, $\hat{f}_{int} = 0.15$, and $\hat{T} = 0.0375$ for the European data set (Table 6). These estimates suggest a scenario of very slow growth over a long period of time with no bottleneck for the African-Americans and a fairly recent population bottleneck with ~13-fold recovery for the Europeans.

TABLE 4

Distribution of $\hat{f}_{rec}/f_{rec}$

f_{rec}	Mean	SD	0.05	0.5	0.95
10	1.0674	0.2926	0.7	1.0	1.6
20	1.0925	0.3862	0.7	1.0	1.7
40	1.1629	0.5722	0.6	1.0	2.2
80	1.2980	0.8200	0.5	1.0	3.3
160	1.3906	1.0092	0.4	1.0	3.9
320	1.5072	1.1824	0.3	1.0	3.9

Time of expansion is ~50,000 years ($T = 0.0625$), and f_{int} is fixed at 1. Simulated data sets consist of 5000 unlinked sites in a sample size of 50. The $\hat{f}_{rec}$ grid includes 40 grid points from $\hat{f}_{rec} = 0.1(f_{rec})$ to $\hat{f}_{rec} = 4(f_{rec})$.

TABLE 5

Distribution of three-dimensional MLEs

	Mean	SD	0.05	0.5	0.95
$\hat{f}_{rec}$	6.017479	4.233278	1.5	5	15
$\hat{f}_{int}$	0.056695	0.042797	0.01	0.045	0.14
$\hat{T}$	0.112958	0.038210	0.055	0.115	0.175

Simulated data sets consist of 500 sites in a sample size of 30 chromosomes, where $f_{rec} = 5.0$, $f_{int} = 0.05$, and $T = 0.125$. The three-dimensional grid includes $\hat{f}_{rec}$ values from 0.5 to 14.5 (at 0.5 intervals), $\hat{f}_{int}$ values from 0.01 to 0.15 (at 0.005 intervals), and $\hat{T}$ values from 0.025 to 0.225 (at 0.01 intervals).

TABLE 6

Hausa and Seattle SNPs analysis

	$\hat{f}_{rec}$	$\hat{f}_{int}$	$\hat{T}$	Likelihood[a]	*P*-value[b]
Hausa data set (A. Di Rienzo, unpublished data)					
MLE	3.1	1	6.1	−1411.14	0.304
Constant population size	1	1	—	−1413.34	0.204
Seattle SNPs African-American MLE	1.9	1	0.27	−1411.69	0.113
African-American data set (http://pga.gs.washington.edu)					
MLE	1.9	1	0.27	−15448.17	2×10^{-4}
Constant population size	1	1	—	−15544.63	$\ll 1 \times 10^{-4}$
European data set (http://pga.gs.washington.edu)					
MLE	2.0	0.15	0.0375	−11987.28	0.2015
Constant population size	1	1	—	−12022.41	$< 1 \times 10^{-4}$

[a] Note that this is not true likelihood since the segregating sites are not entirely unlinked.

[b] *P*-values are calculated from a χ^2 goodness-of-fit test where the distribution of the χ^2 test statistic is simulated for each data set, accounting for linkage as described in the text.

To determine whether the demographic model we consider is compatible with the Seattle SNPs data, we simulate the distribution of the goodness-of-fit test statistic for this data set as described for the Hausa data set. For these simulations, each data set consisted of 48 chromosomes and 138 loci. The input parameters were that of the Hausa data set, except substituting the average length of a Seattle SNPs locus. In this case, each locus was simulated, fixing the number of segregating sites to be the average number of segregating sites per locus in the African-American or European Seattle SNPs data set, so each simulated data set contained the same total number of segregating sites as our observed Seattle SNPs African-American or European data set. Because the Seattle SNPs data sets do not specify the ancestral/derived status of each allele, each simulated frequency spectrum is again folded. The 95% critical values of the distribution were found to be 48.68 (African) and 137.36 (European) as opposed to 35.17 (23 d.f.) and 33.92 (22 d.f.) if all sites were unlinked.

Using these simulated critical values, a χ^2 goodness-of-fit test indicates that the maximum-likelihood parameters produce an expected frequency spectrum that is not significantly different from the observed Seattle SNPs European data ($\chi^2 = 122.98$; $P = 0.2015$). Therefore, we can accept our simple bottleneck model as a reasonable explanation for this data set. The same test indicates that a constant population size model is not compatible with the European data ($\chi^2 = 207.286$; $P < 1 \times 10^{-4}$).

However, the χ^2 goodness-of-fit test on the African-American data set reveals that the frequency spectrum predicted by the maximum-likelihood estimates of f_{rec}, f_{int}, and T is significantly different from the empirical Seattle SNPs African-American frequency spectrum ($\chi^2 = 86.64$; $P = 2 \times 10^{-4}$); therefore, our simple demographic model cannot be accepted as a complete explanation of the African-American data set, although the fit is better than that predicted by the constant population size model ($\chi^2 = 268.66$; $P \ll 1 \times 10^{-4}$). Figure 5 provides a visual comparison of the observed Seattle SNPs frequency spectrum to the frequency spectra predicted by both the maximum-likelihood parameters and constant population size parameters, indicating that the lack of fit of the maximum-likelihood parameters does not seem to be confined to any particular nonsingleton frequency class. However, our demographic model with the maximum-likelihood parameters appears to provide a better fit to the data than the equilibrium model, particularly in the singleton class. In addition, we note that a χ^2 goodness-of-fit test shows that the Hausa data are compatible with $f_{rec} = 1.9$, $f_{int} = 1$, and $T = 0.27$, the estimates obtained from the Seattle SNPs African-American data set ($\chi^2 = 33.46$; $P = 0.113$).

Because the sites in the Hausa and Seattle SNPs data sets are not entirely unlinked, asymptotic approximation of confidence intervals is not appropriate. We simulated 10,000 data sets as described above for both the Hausa and Seattle SNPs data sets, using their respective maximum-likelihood estimates for input parameters, and applied the maximum-likelihood method to the folded frequency spectrum of each simulated data set. For the Hausa and Seattle SNPs African-American data sets, we estimated both f_{rec} and T, fixing f_{int} at 1, which was the maximum-likelihood estimate for both data sets. All three parameters were estimated for the data sets simulated from the Seattle SNPs European parameters. The ratio of the log-likelihood at the maximum-likelihood parameters to the log-likelihood at the parameters from which the data set was simulated could then be calculated. From the log-likelihood ratio distribution, we determined the 95% critical value to be 3.86 for the Hausa data set and 4.85 for the Seattle SNPs African-American data set as compared to the asymptotic critical

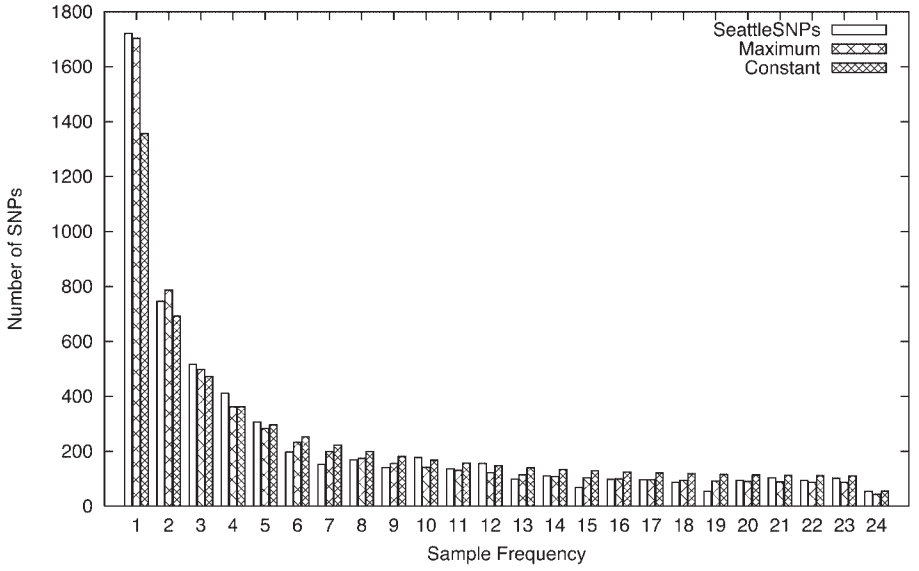


FIGURE 5.—African-American Seattle SNPs folded frequency spectra comparison. Empirical Seattle SNPs frequency spectrum and the expected frequency spectrum for demographic parameters corresponding to the Seattle SNPs maximum-likelihood estimate ($f_{rec} = 1.9$, $f_{int} = 1$, and $T = 0.27$) and constant population size are shown. The number of SNPs at a sample frequency of i is equal to the total number of SNPs (5892) times p_i (folded). For constant population size, p_i's are obtained from Equation 9, and, for the maximum-likelihood parameters, p_i's are obtained from simulation as described in the text.

value of 3.0 for two-dimensional maximum-likelihood estimates. The 95% critical value of 9.1 was found for the European data set as compared to the asymptotic critical value of 3.9 for three-dimension estimation. Using the critical values from simulation, we can easily reject the constant-size population model for the Seattle SNPs African-American and European data sets since the log likelihood ratios are 96 and 35, respectively (Table 6).

Figure 6a provides a visual representation of the 95, 99, and 99.9% confidence regions of the Hausa data set obtained by including all parameter values for which the log-likelihood ratio is ≤3.86, 6.38, and 10.08, respectively. Likewise, Figure 7 illustrates the analogous confidence regions for the Seattle SNPs African-American (Figure 7a) and European (Figure 7b) data sets.

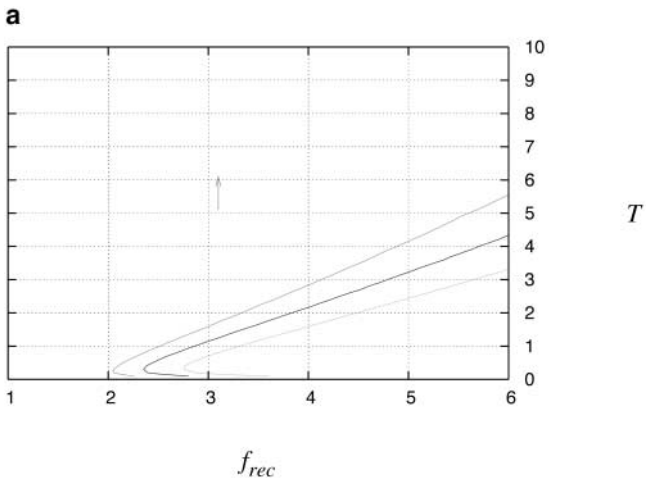
DISCUSSION

Our power analyses on models with exponential growth beginning 10,000 years ago illustrate that the frequency spectrum does not provide sufficient information to reject the null hypothesis of constant population size when either small sample sizes (<50 chromosomes) or small numbers of unlinked sites (<1000) are available. This result should serve as a cautionary note to researchers interested in demographic models involving expansion as recent as 10,000 years ago. Prior knowledge of the model of interest should also be considered when determining whether the frequency spectrum retains the requisite information for demographic inference, as the power to detect departures from constant size increases with both the extent of growth (Figure 2, a and b) and the time since the onset of growth (Figure 2b).

Application of the maximum-likelihood method on recent growth scenarios reveals that data sets consisting of at least 250 chromosomes with at least 10,000 scaled segregating sites (15,838–17,140 segregating sites depending upon the true f_{rec} value) are required for the $\hat{f}_{rec}$ distribution to have 95% critical values that fall within a factor of four of the true f_{rec} value if growth began as recently as 10,000 years ago ($T = 0.0125$). Unless large sample sizes and many unlinked sites are surveyed, the frequency spectrum alone provides little information about the magnitude of growth that has occurred relatively recently. As f_{rec} increases, the frequency spectrum becomes more distinct from what would be expected under a constant size scenario. However, with increasingly extreme recent growth, the frequency spectrum becomes less distinguishable from that of other severe growth scenarios, and it becomes more difficult to estimate the f_{rec} parameter with frequency spectrum information alone.

While it is difficult to accurately estimate f_{rec} for scenarios of recent growth, T can be estimated with more modest amounts of data. The distribution of $\hat{T}$ has 95% critical values that fall within a factor of four of the true T value for sample sizes as small as 50 chromosomes and 5000 sites, as compared to a sample size of 250 and 15,838–17,140 sites required to estimate f_{rec} to the same accuracy. Additionally, $\hat{T}$ is not subject to the upward bias seen in $\hat{f}_{rec}$ and estimates of T actually improve with increasing f_{rec}. Estimates of both f_{rec} and T improve as the onset of growth becomes more ancient. This observation is consistent with our observation that power to reject the null hypothesis of constant population size with frequency spectrum data increases with scenarios of more ancient growth (Figure 2b).

Simultaneous estimation of all three parameters results in estimator distributions where 90% of the estimates lie within a factor of four of the true parameter values with data sets as small as 500 segregating sites in a sample size of 30 for a model where the population decrease and subsequent expansion began 100,000 years



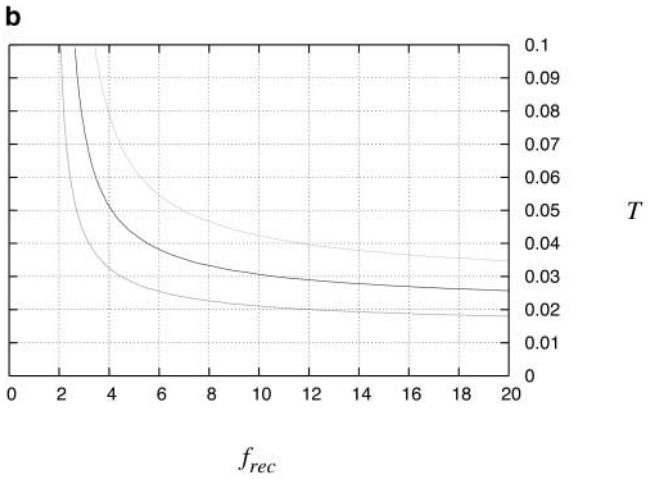
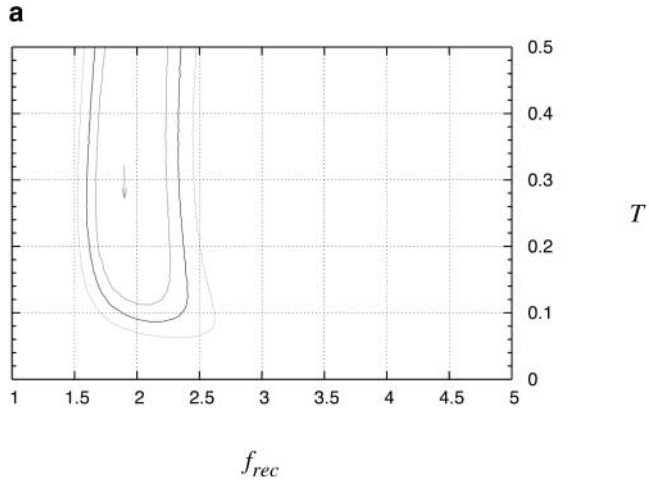


FIGURE 6.—Hausa confidence region. The third dimension, f_{int}, is fixed at 1. (a) Maximum-likelihood estimate (MLE) is indicated by the arrow ($\hat{f}_{rec}$ = 3.1, $\hat{T}$ = 6.1). (b) Focus on recent growth times with expanded f_{rec} range. The leftmost, middle, and rightmost contours represent the 95, 99, and 99.9% confidence intervals (3.86, 6.38, and 10.08 log-likelihood units, respectively).



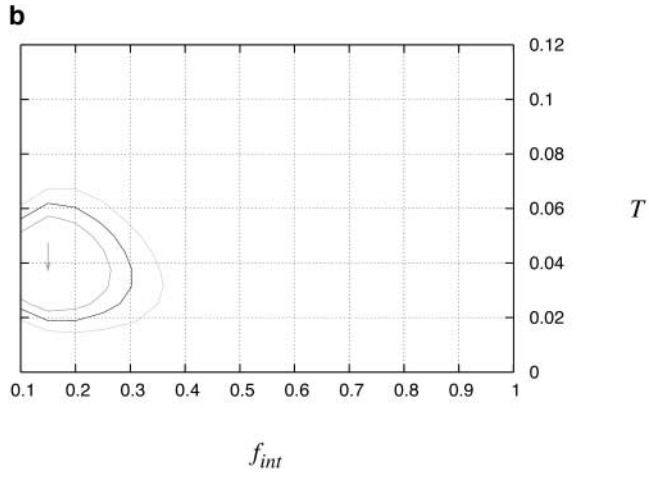


FIGURE 7.—Seattle SNPs confidence regions. (a) African-American data set, with MLE indicated by the arrow ($\hat{f}_{rec}$ = 1.9, $\hat{T}$ = 0.27). The third dimension, f_{int}, is fixed at the MLE of $\hat{f}_{int}$ = 1. (b) European data set, with MLE indicated by the arrow ($\hat{f}_{int}$ = 0.15, $\hat{T}$ = 0.0375). The third dimension (f_{rec}) is fixed at the MLE of $\hat{f}_{rec}$ = 2.0. The innermost, middle, and outermost contours surrounding the MLE represent the 95, 99, and 99.9% confidence regions, respectively.

ago and the present population is only five times the initial population size. These estimates benefit from both a more ancient time of growth onset and a modest magnitude of growth that is not subject to the upward bias seen in more severe growth scenarios. The ability of the frequency spectrum alone to elucidate the time and magnitude of population size change events is, therefore, greatly dependent upon the underlying demographic model. While ancient demographic events may be inferred relatively accurately from contemporary frequency spectrum patterns, more recent and severe episodes of growth are problematic for this method and require exceedingly large amounts of unlinked data. For these recent growth scenarios, it is possible that more informative estimates could be obtained by using a method that uses linked polymorphic sites and considers additional aspects of the data such as levels of linkage disequilibrium.

Evaluation of the asymptotic properties of our maximum-likelihood estimators indicates that asymptotic theory provides a reasonable approximation of the confidence intervals associated with the estimators. As we illustrate with the Hausa and Seattle SNPs data sets, it is also possible to construct these confidence intervals around a maximum-likelihood estimate through simulation. By simulating data sets that closely match the properties of the observed data set, one can estimate the critical value of this log-likelihood ratio distribution and construct corresponding confidence regions. This procedure is particularly relevant when asymptotic approximation is not appropriate, such as when the segregating sites in a data set are not unlinked.

We apply the maximum-likelihood method to both the African Hausa data set and the African-American and European samples of the Seattle SNPs data set.

In both the Hausa and the Seattle SNPs data sets, the segregating sites are not entirely unlinked, but the maximum-likelihood analysis treats them as though each site provides independent information. However, we illustrate how one may use simulation to construct confidence regions and use goodness-of-fit tests that take into account the linkage between sites.

In the Seattle SNPs African-American data set, the simulated 95% confidence interval clearly allows for rejection of the constant population size model, since the log-likelihood of observing the data is almost 100 units less with the constant size parameters than with the estimated parameters. The maximum-likelihood estimates of $\hat{f}_{rec} = 1.9$, $\hat{f}_{int} = 1$, and $\hat{T} = 0.27$ based on the Seattle SNPs African-American data correspond to a slow, ancient growth scenario where growth began >200,000 years ago to a present size of approximately two times the initial population size.

The simulated 95% confidence region around the Seattle SNPs African-American maximum-likelihood estimates, shown in Figure 7a, includes only a very narrow range of f_{rec} values within 1.6–2.5. However, the confidence region includes a wide range of T values ranging from as recent as 80,000 years ago ($T = 0.1$) to the most ancient time examined, 800,000 years ago ($T = 1$), assuming a generation time of 20 years and an N_0 of 10,000. Even with the most recent compatible T value, it is not surprising that this data set allows for rejection of the constant size hypothesis with an estimate of only twofold growth. Our power analyses show that a data set consisting of 50 chromosomes has a power of 0.9 to reject the constant size hypothesis with only 1000 unlinked sites for twofold growth beginning 100,000 years ago. While the Seattle SNPs data set does not consist of entirely unlinked sites, our analysis included >5000 polymorphic sites across 138 loci, which should allow for comparable power.

Despite the compact confidence region (Figure 7a), visually reasonable fit of the frequency spectrum under the maximum-likelihood parameters to the observed data (Figure 5), and compatibility with the Hausa data set, a χ^2 goodness-of-fit test indicates that our simple three-dimensional demographic model with the maximum-likelihood estimates obtained from the Seattle SNPs African-American data set is incompatible with the Seattle SNPs data ($P = 2 \times 10^{-4}$), even when linkage is taken into account. The visual comparison between the observed and maximum-likelihood frequency spectra in Figure 5 seems to indicate that the number of singletons expected under the maximum-likelihood parameters is very close to the observed value, and therefore the incompatibility must be due to some combination of the other frequency categories. The loci in the Seattle SNPs data set were chosen because of their role in inflammatory pathways and may reflect the action of evolutionary forces other than population size changes. Although we removed those SNPs that result in amino-acid coding changes, which are more apt to be subject to natural selection, it is still probable that this data set includes SNPs that are mildly deleterious and may influence the frequency spectrum toward greater numbers of low-frequency variants and mimic evidence of growth.

The African-American population sampled for the Seattle SNPs data set may also be subject to population structure and admixture, which could affect the frequency spectrum and confound our inference about demographic history (Ptak and Przeworski 2002). To determine the effect of European admixture on maximum-likelihood estimates obtained from an African data set, we randomly combined 6 Italian chromosomes (A. Di Rienzo, unpublished data) with the 30 Hausa chromosomes at each of the 50 locus pairs of the Hausa data set, which resulted in a total data set of 657 polymorphic sites in 36 chromosomes. This represents ~17% European admixture, which is consistent with admixture estimates obtained from African-American populations (Parra *et al.* 1998). Admixture of this proportion had virtually no effect on the maximum-likelihood estimates or confidence intervals based on the original Hausa data set (data not shown). Regardless, that does not eliminate the possibility that either the true population structure could involve admixture in different proportions or admixture in a larger data set such as the Seattle SNPs would produce a more prominent effect. The frequency spectrum of the Seattle SNPs data set is certainly not consistent with an equilibrium model of constant population size, although the degree of growth predicted is less than that of some previous reports based on African populations (Aris-Brosou and Excoffier 1996; Pritchard *et al.* 1999). However, our estimate of twofold growth beginning as recently as 80,000 years ago is consistent with a recent study based on the frequency spectrum in an African-American population (Marth *et al.* 2004).

The maximum-likelihood parameters estimated from this data set are consistent with the Hausa data set, which contains noncoding loci that are less likely to be subject to confounding factors such as selection. However, this analysis does not preclude population structure within Africa as a potential influence on the maximum-likelihood estimates of the Hausa data set. The maximum-likelihood estimates from the Hausa data set ($\hat{f}_{rec} = 3.1$, $\hat{f}_{int} = 1$, and $\hat{T} = 6.1$) correspond to a scenario of slow, ancient threefold growth, beginning several million years ago. However, the confidence region associated with this data set (Figure 6a) is consistent with a wide range of growth scenarios, including both the demographic history estimated from the Seattle SNPs data set ($\hat{f}_{rec} = 1.9$, $\hat{f}_{int} = 1$, and $\hat{T} = 0.27$) and $f_{rec} = 1$, which corresponds to constant population size. Additionally, Figure 6b provides a close-up view of the acceptance region of Figure 6a, considering only more recent values of T where the onset of growth occurs no more than 80,000 years ago. If we focus on these T values, it is

clear that our confidence regions on this data set do not exclude scenarios of 20-fold or more growth, provided that the time of onset is correspondingly more recent. For example, if we believe that the Hausa population has undergone growth >5-fold, then our analysis indicates that the growth must have begun no earlier than $T = 0.045$ (36,000 years ago if N_0 is 10,000 assuming a generation time of 20 years). Growth of that magnitude or larger is rejected at the 1% level (Figure 6b) for values of $T > 0.045$ and <3 ($\sim$36,000–2.4 million years ago).

Our analysis of the Seattle SNPs European data set reveals an estimated demographic history of $\hat{f}_{\text{rec}} = 2.0$, $\hat{f}_{\text{int}} = 0.15$, and $\hat{T} = 0.0375$, which corresponds to an 85% reduction in population size at $T = 0.0375$ (30,000 years ago assuming $N_0 = 10{,}000$ and a 20-year generation time) and then $\sim$13-fold exponential growth to a current population size of twice the ancestral size. In constructing our data set, we exclude all SNPs that are not successfully typed in every chromosome to facilitate construction of appropriate confidence regions and estimation of χ^2 critical values through simulation. However, we note that if all SNPs that were typed in at least half of the sampled chromosomes (7410 SNPs) were included in this analysis, we get only a slightly different estimate ($\hat{f}_{\text{rec}} = 1.25$, $\hat{f}_{\text{int}} = 0.2$, and $\hat{T} = 0.05$) that differs by <2 log-likelihood units from our maximum-likelihood estimate based on the filtered data ($-20{,}668.96$ *vs.* $-20{,}670.93$ when all SNPs are included). Since the 95% confidence region includes all parameter values within 9.1 log-likelihood units from the maximum, it is not likely that this filtering of the data would result in a significant shift in our acceptance region.

A χ^2 goodness-of-fit test indicates that frequency spectrum produced by the estimated parameters ($\hat{f}_{\text{rec}} = 2.0$, $\hat{f}_{\text{int}} = 0.15$, and $\hat{T} = 0.0375$) is a reasonable match to the observed Seattle SNPs European frequency spectrum with a *P*-value of 0.2015. The constant size model is both rejected by the goodness-of-fit test and excluded by the simulated likelihood-ratio confidence region for this data set (Table 6, Figure 7b). These results implicate a bottlenecked history for this European data set, which is consistent with previous studies (Marth *et al.* 2003, 2004) and the "Out of Africa" model for human population history (Harpending *et al.* 1998). Since the Seattle SNPs European data set is composed of the same coding loci as the Seattle SNPs African-American data set, it seems reasonable that the lack of agreement between the frequency spectrum predicted by the maximum-likelihood parameters and the observed African-American frequency spectrum is more likely to be due to population structure than to the presence of slightly deleterious variants in the data set. Of course, the good fit of the European maximum-likelihood parameters does not preclude the possibility of population structure or selection within the European data set as well.

As we have indicated earlier, there are a number of confounding factors to consider when attempting to infer demographic history based on frequency spectrum information, including population structure (past or present) and selection. An additional complication that is not considered by these analyses is ascertainment bias and genotyping error. It has been shown that ascertainment bias can lead to large errors in maximum-likelihood-based demographic inference (Kuhner *et al.* 2000; Wakeley *et al.* 2001). Polanski and Kimmel (2003) have also shown that exclusion or misclassification of low-frequency SNPs can result in estimated growth rates that are significantly lower than the true value. Note, however, that the sites represented in the Di Rienzo and Seattle SNPs data sets were chosen without prior indication of polymorphism status. Therefore, the analyses on these data sets would not be influenced by ascertainment bias due to using a discovery sample for SNP identification. However, we cannot exclude the possibility that genotyping errors have biased our inferences.

Conclusions: Analysis of this maximum-likelihood method indicates that demographic inferences can be drawn from frequency spectrum data when sufficient amounts of data are available. Asymptotic theory or simulation can be used to determine the variance and covariance associated with these estimators to determine whether the maximum-likelihood estimates would be meaningful for a particular demographic model and amount of data that may be available. However, our results show that very large amounts of data may be required to obtain practical confidence regions, particularly in models involving recent growth. For growth beginning as recently as 10,000 years ago, the power to reject the hypothesis of constant population size is very low with sample sizes of <20 chromosomes. To make accurate inferences under this type of recent-growth model using the frequency spectrum alone, both large samples (>100 chromosomes) and a large number of unlinked sites (>5000 sites) are required, although estimators improve as the time of onset of growth becomes more ancient. In scenarios of extreme growth, there is also a severe bias in $\hat{f}_{\text{rec}}$, even with large amounts of data. However, T can be estimated with more modest amounts of data, and $\hat{T}$ is not subject to the bias seen in $\hat{f}_{\text{rec}}$, indicating that one may obtain reasonable estimates of the time of population size-change events, even if the magnitude is biased. This maximum-likelihood method incorporates all available information contained in unlinked polymorphic sites, and parameter estimation methods based on summaries of the frequency spectrum require even larger amounts of data to be equally as informative. Therefore, for scenarios where the entire frequency spectrum of modest data sets does not provide an adequate amount of information, it may be necessary to incorporate additional aspects of linked data to improve estimates of demographic parameters.

Application of the maximum-likelihood method to

three human data sets implicates differing demographic histories for African *vs.* European data sets. The African Hausa data set is compatible with a wide range of growth scenarios, ranging from slow, ancient growth to some scenarios of very recent, rapid growth. However, we can reject episodes of greater than 5-fold growth beginning >36,000 and <2.4 million years ago on the basis of this data set. The Seattle SNPs African-American data set also supports a model of growth, although a goodness-of-fit test indicates that the best-fit model of ancient, slow growth is not sufficient to explain the observed frequency spectrum. Maximum-likelihood analysis of the Seattle SNPs European data set reveals that the best-fit model is one of a population bottleneck occurring ~30,000 years ago, reducing the population to 15% of the ancestral size, followed by 13-fold growth to a current population size that is twice the ancestral size.

We thank A. Di Rienzo for access to the Hausa data set prior to publication. Construction of the Hausa data set was supported by National Institutes of Health (NIH) grant HG-02098. A. Adams was supported by NIH/National Institute of General Medical Sciences grant 5 T32 GM07197.

LITERATURE CITED

Aris-Brosou, S., and L. Excoffier, 1996 The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. Mol. Biol. Evol. **13** (3): 494–504.

Beerli, P., and J. Felsenstein, 2001 Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc. Natl. Acad. Sci. USA **98:** 4563–4568.

Cargill, M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie *et al.*, 1999 Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. **22:** 231–238.

Ewens, W. J., 1979 *Mathematical Population Genetics.* Springer-Verlag, New York.

Fay, J. C., J. Wyckoff and C.-I Wu, 2001 Positive and negative selection on the human genome. Genetics **158:** 1227–1234.

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. **69:** 831–843.

Griffiths, R. C., and S. Tavaré, 1998 The age of a mutation in the general coalescent tree. Stoch. Models **14:** 273–295.

Harpending, H. C., M. A. Batzer, M. Gurven, L. B. Jorde, A. R. Rogers *et al.*, 1998 Genetic traces of ancient demography. Proc. Natl. Acad. Sci. USA **95:** 1961–1967.

Hudson, R. R., 1983 Properties of the neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

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

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

Kuhner, M. K., P. Beerli, J. Yamato and J. Felsenstein, 2000 Usefulness of single nucleotide polymorphism data for estimating population parameters. Genetics **156:** 439–447.

Marth, G., G. Schuler, R. Yeh, R. Davenport, R. Agarwala *et al.*, 2003 Sequence variations in the public human genome data reflect a bottlenecked population history. Proc. Natl. Acad. Sci. USA **100** (1): 376–381.

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

Nielsen, R., 1999 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154:** 931–942.

Parra, E. J., A. Marcini, J. Akey, J. Martinson, M. A. Batzer *et al.*, 1998 Estimating African American admixture proportions by use of population-specific alleles. Am. J. Hum. Genet. **63:** 1839–1851.

Pluzhnikov, A., A. Di Rienzo and R. R. Hudson, 2002 Inferences about human demography based on multilocus analyses of noncoding sequences. Genetics **161:** 1209–1218.

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

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun and M. W. Feldman, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. **16** (12): 1791–1798.

Ptak, S. E., and M. Przeworski, 2002 Evidence for population growth in humans is confounded by fine-scale population structure. Trends Genet. **18:** 559–563.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Wakeley, J., R. Nielsen, S. N. Liu-Cordero and K. Ardlie, 2001 The discovery of single-nucleotide polymorphisms—and inferences about human demographic history. Am. J. Hum. Genet. **69:** 1332–1347.

Wall, J. D., and M. Przeworski, 2000 When did the human population size start increasing? Genetics **155:** 1865–1874.

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

Wooding, S., and A. Rogers, 2002 The matrix coalescent and an application to human single-nucleotide polymorphisms. Genetics **161:** 1641–1650.

Communicating editor: M. Nordborg