Inferences About Human Demography Based on Multilocus Analyses of Noncoding Sequences

Anna Pluzhnikov,* Anna Di Rienzo* and Richard R. Hudson†,[1]

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

Manuscript received December 27, 2001
Accepted for publication March 19, 2002

ABSTRACT

Data from 10 unlinked autosomal noncoding regions, resequenced in 15 individuals from each of three populations, were used in a multilocus analysis to test models of human demography. Each of the 10 regions consisted of ~2500 bp. The multilocus analysis, based on summary statistics (average and variance of Tajima's *D* and Fu and Li's *D**), was used to test a family of models with recent population expansion. The African sample (Hausa of Cameroon) is compatible with a constant population size model and a range of models with recent expansion. For this population sample, we estimated confidence sets that showed the limited range of parameter values compatible with growth. For an exponential growth rate as low as 1×10^{-3}/generation, population growth is unlikely to have started prior to 50,000 years ago. For higher growth rates, the onset of growth must be more recent. On the basis of the average value of Tajima's *D*, our sample from an Italian population was found to be incompatible with a constant population size model or any simple expansion model. In the Chinese sample, the variance of Tajima's *D* was too large to be compatible with the constant population size model or any simple expansion model.

ELUCIDATING the history of the human population size is an important part of reconstructing human evolution and understanding patterns of human variation. Changes in population size are thought to mark important events in the history of a species, *e.g.*, geographic range expansions, development of technological innovations, and climatic changes. In addition, the estimation of the time since the most recent common ancestor (TMRCA), which has important implications for human evolution, relies critically on assumptions about human demography (Brookfield 1997; Stephens and Donnelly 2000; Thomson *et al.* 2000). Finally, inferences about the frequency of disease-causing alleles in human populations also rely on assumptions about the demographic history (Reich and Lander 2001). In the last 10 years, many data sets have been collected with the purpose of gaining insights into the history of population growth in humans. MtDNA sequences showed evidence in favor of rapid population growth from a small initial size such as that leading to a star-shaped genealogy. This conclusion was supported by the excess of rare variants and the analysis of mismatch distributions (Di Rienzo and Wilson 1991; Slatkin and Hudson 1991; Rogers and Harpending 1992). Microsatellite studies also support the idea of an ancient population growth. However, different studies considered different models of growth, which include growth from an initial small and constant size, growth following a bottleneck, and growth from an initial constant and nontrivial size (Di Rienzo *et al.* 1998; Kimmel *et al.* 1998; Reich and Goldstein 1998; Gonser *et al.* 2000; Zhivotovsky *et al.* 2000). In addition, different models of the mutation process have been assumed in different studies. As a consequence of this heterogeneity, and the uncertainty about the parameters of the mutation process, the results of microsatellite studies did not lead to consistent results on the evidence for growth and the populations that experienced it.

Similarly, studies of nuclear sequence variation lead to somewhat contrasting conclusions. In general, nuclear loci show more ancient coalescence times compared to mtDNA (as might be expected on the basis of the different effective population sizes), but no evidence for a star-shaped genealogy. The latter observation suggests that rapid growth from a small initial size is not compatible with the data and that, if ancient population growth occurred, it started from a population of nontrivial size. Unfortunately, studies of nuclear sequence variation also vary greatly with regard to the scheme for sampling populations (from population-based to grid sampling), the type of genomic regions studied (from coding to noncoding), and the method of variation detection (Przeworski *et al.* 2000). Wall and Przeworski (2000) considered a variety of demographic models, but none of them could account for the pattern of variation seen at all loci. Since many of the loci examined spanned coding regions, it could not be excluded that natural selection affected a portion of them. They suggested that models of bottleneck or geographic

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

structure underlie the patterns observed at most loci while a minority of loci are affected by directional selection.

More recently, 10 noncoding regions were surveyed in three population samples to characterize the decay of linkage disequilibrium and obtain population-based estimates of the crossing-over and gene conversion rates (Frisse *et al.* 2001). This survey also showed that the two non-African samples depart significantly from the expectations of an equilibrium model for several aspects of the data, including average and variance of Tajima's *D* across loci, higher-than-expected levels of linkage disequilibrium, and a high proportion of loci with significant Fay and Wu's *H*-tests (Hamblin *et al.* 2002). The African sample, however, appeared to be consistent with an equilibrium model with a constant long-term population size. Several scenarios of demography emerge as relevant models for human populations. One is a model of constant but nontrivial size followed by recent population growth. This model, which appears to be most relevant to African populations, is the main topic of this article. Additional models envision a population of constant and nontrivial size that experiences a bottleneck or temporary population subdivision, in both cases followed by population growth. These more complex models may be most relevant to European populations and perhaps non-African populations in general.

Here, we reanalyze the noncoding sequence data in Frisse *et al.* (2001) to test explicitly models of growth from an ancestral population at equilibrium. The analyses carried out here differ from those in Frisse *et al.* (2001) in that they properly take into account the effect of recombination occurring within loci and test a simple growth model for a broad range of parameter values. These tests are based on Tajima's *D* statistic (Tajima 1989) and Fu and Li's *D** statistics (Fu and Li 1993). The critical values of these statistics depend on the mutation rate, recombination rate, and population size history. We estimated these critical values for different population histories using coalescent-based Monte Carlo simulations. We incorporate uncertainty about certain genetic parameters (recombination and mutation rates) and their variability between loci by assuming these rates are random variables. Thus, in the simulations, the values of these rates are drawn from specified probability distributions. Furthermore, we restricted our attention only to combinations of parameter values expected to match the polymorphism levels observed in human variation studies.

In agreement with Frisse *et al.* (2001), we find that the African sample is compatible with the constant population size model, as well as a range of models with recent expansion. Since the growth models considered here imply an interdependence among the parameters, we estimated confidence sets for a combination of parameters rather than confidence intervals for individual parameters. Such confidence sets show that, for an exponential growth rate as low as 1×10^{-3}/generation, the growth phase is unlikely to have started earlier than ~50,000 years ago. For higher—possibly more realistic—growth rates, the onset of growth must be more recent. Both non-African samples are incompatible with the constant population size model or any version of our growth model.

MATERIALS AND METHODS

Data collection: A new scheme for data collection was developed to survey simultaneously and efficiently sequence variation and linkage disequilibrium (LD). This consisted of resequencing two segments of ~1 kb separated by ~8 kb in all individuals from three population samples. Each of these two-segment units is referred to as a "locus pair." The data set analyzed here consists of 10 such locus pairs that are unlinked to each other. The genomic regions were chosen according to a fixed set of criteria. These criteria were determined by the need to pool data from different locus pairs in the analysis and, thus, to select locus pairs with similar recombination and mutation rates. In addition, because the main goal of this analysis is to reconstruct demographic histories, it was necessary to reduce the probability that the surveyed genomic regions were affected by natural selection. This was achieved by choosing regions that do not contain or flank known or strongly predicted coding regions (the minimum distance between the regions surveyed and the closest known or strongly predicted gene was >25 kb). The details of the procedure for selecting genomic regions that fulfilled these criteria are described in Frisse *et al.* (2001). The surveyed regions have an average crossing-over rate of 1.29 cM/Mb and 35–45% G + C content. The 10 locus pairs were resequenced in all 15 individuals (30 chromosomes) of samples drawn from each of three large populations from the major ethnic groups: Hausa of Cameroon (Sub-Saharan Africa), Italians (Europe), and Han Chinese (Asia). Descriptive statistics of sequence variation are shown in Table 1.

Demographic history models in the coalescent framework: The basic model of population demography is the Wright-Fisher model, which assumes a panmictic population with nonoverlapping generations. We assume the diploid population size in the distant past was constant at size N_A. At a time, t_{onset} generations in the past, the population began exponentially growing until the present. Thus, measuring time in units of generations before present we assume the population size, $N(t)$, is

$$N(t) = \begin{cases} N_A, & t \geq t_{onset} \\ N_A e^{\alpha(t_{onset}-t)}, & 0 \leq t \leq t_{onset} \end{cases} \quad (1)$$

(see Weiss and von Haeseler 1998; Pritchard *et al.* 1999). The current population size, $N(0) = N_A e^{\alpha t_{onset}}$, is denoted N_0. The primary goal of this article is to determine the values of α and t_{onset} for which the model is compatible with the data. We assume a generation time of 20 years. Only positive values of α are considered.

Coalescent simulations with recombination were used to generate samples under this model. These simulations used standard methodology (Hudson 1983, 1990) except that the values of some parameters were drawn from distributions as described below. Recombination was assumed homogeneous at all sites within each locus pair, but the recombination rate for each locus pair was drawn from a gamma distribution as described below. Likewise, an infinite sites model of mutation was assumed with mutation parameters also drawn from a

TABLE 1

Summary statistics of sequence variation

		Hausa				Italians				Chinese			
Region	L^a	S^b	π^c (%)	TD^d	FLD*^{e}	S^b	π^c (%)	TD^d	FLD*^{e}	S^b	π^c (%)	TD^d	FLD*^{e}
1	2423	12	0.08	−1.27	−1.81	6	0.08	−0.16	0.44	3	0.03	−0.37	−0.26
2	2552	15	0.18	0.73	0.36	11	0.06	−1.47	−0.57	9	0.04	−1.69	0.20
3	2792	17	0.15	−0.03	0.87	13	0.16	1.31	0.17	9	0.11	1.00	0.79
4	2560	10	0.12	0.74	0.87	7	0.11	1.72	1.27	8	0.08	0.03	−0.58
5	3050	9	0.08	0.29	0.79	10	0.11	0.88	−0.21	9	0.05	−0.99	−2.71
6	2920	16	0.10	−0.93	−1.43	8	0.06	−0.53	0.69	9	0.04	−1.45	−2.12
7	2811	11	0.07	−0.96	0.94	7	0.12	2.70	1.27	10	0.10	0.37	0.87
8	2034	9	0.09	−0.69	−0.38	5	0.07	0.46	1.14	8	0.09	−0.27	0.69
9	1791	6	0.08	−0.22	−1.11	4	0.08	1.10	0.05	3	0.09	2.43	0.95
10	2110	15	0.15	−0.52	−0.43	9	0.13	0.63	0.20	8	0.14	1.53	1.32

[a] Length in base pairs of sequenced segment.
[b] Number of polymorphic sites.
[c] Nucleotide diversity per base pair.
[d] Tajima's D (Tajima 1989).
[e] Fu and Li's D^* (Fu and Li 1993).

gamma distribution as described below. The middle "8 kb" of the generated polymorphisms was ignored to produce samples analogous to our locus pair data.

Parameter values: Since the goal of this study is making inferences about various demographic scenarios, the parameters not directly associated with demography, such as the mutation rate, μ, and the recombination rate, c, can be thought of as *nuisance* parameters. Elimination of these nuisance parameters is easily achieved by adopting the Bayesian approach, namely by viewing them as random quantities and subsequently integrating them out (Severini 1999). Specifically, we model the uncertainty in the values of μ by a Gamma(2, β_μ) distribution. Similarly, we assume that the recombination rate, c, is an independent Gamma(2, β_c) random variable. The shape parameters β_μ and β_c are chosen so that the means $E\mu = 2\beta_\mu$ and $Ec = 2\beta_c$ of these distributions correspond to genome-wide estimates for these parameters, namely an average mutation rate of 2×10^{-8}/site/generation and a recombination rate between adjacent base pairs of 1×10^{-8}/generation. Hence, we set $\beta_\mu = 1 \times 10^{-8}$ and $\beta_c = 0.5 \times 10^{-8}$. The central 90% intervals for these distributions are (0.36×10^{-8}, 4.74×10^{-8}) and (0.18×10^{-8}, 2.37×10^{-8}) for μ and c, respectively. Recent findings point to 10- to 1000-fold variability in recombination rate over 1–2 kb in the MHC region (Jeffreys *et al.* 2001). If this heterogeneity of recombination rate is indeed typical of the human genome as a whole, our modeling of the variability in c may only partially describe the true recombinational landscape.

Estimates of the neutral mutation rate are based on observed levels of sequence divergence from a great ape outgroup from a number of surveys (Nachman and Crowell 2000; Przeworski *et al.* 2000; Chen and Li 2001); these estimates are in good agreement with those obtained by Frisse *et al.* (2001) for the data set analyzed here. Estimates of the recombination rate are based on the comparison of genome-wide genetic and sequence maps (Yu *et al.* 2001).

We also treat N_A, the *ancestral* effective population size (which is identical to N_0 in the constant population size model), as a random variable independent of all other parameters. In particular, we let N_A be randomly distributed as Gamma(4, β_A). We note that, with the other parameters fixed (α, t_{onset}, and β_μ), the value of N_A determines the mean number of polymorphic sites in samples. To incorporate prior information about observed levels of polymorphism in earlier studies, we chose the value of β_A ($= EN_A/4$) so that the expected number of segregating sites per kilobase in a sample of 30 chromosomes is 4. This choice for β_A is based on the following observations. In a large number of studies, Watterson's estimate of θ ($= 4N_e\mu$) is on average ~0.001 (somewhat larger in African populations and somewhat smaller in non-African populations). This is also the average estimated value of θ in the 10 locus pairs analyzed here. In a sample of 30 chromosomes, this value of θ leads to an expected number of polymorphic sites of 4/kb under the neutral constant population size model. Thus, for the constant population size model we chose β_A so that $4EN_AE\mu = 0.001$. For models with population growth, we also set the value of β_A so that the expected number of polymorphic sites is 4/kb. In this case a simple formula is not available but the appropriate value of β_A or EN_A can be obtained numerically for any specified value of t_{onset} and α, as shown in the appendix.

To complete the model specification under a growth scenario, the remaining two parameters, t_{onset} and α, are allowed to vary over a grid of fixed values $t_{onset} = 1K, 2K, \ldots, 8K$ generations, $\alpha = 0.5 \times 10^{-3}, 1 \times 10^{-3}, \ldots, 10 \times 10^{-3}$. Note that some combinations of t_{onset}, α, and EN_A would be omitted from consideration since to attain the specified mean number of polymorphic sites the corresponding values of EN_0 would have greatly exceeded the current size of the entire world population. Smaller values of α were not considered because they would result in models virtually indistinguishable from the equilibrium model.

Simulation procedure: The polymorphism data are simulated using methods of Hudson (1990) for constant and Slatkin and Hudson (1991) for variable population size. Specifically, for a single locus pair and contiguous sequences of fixed length L, for each combination of t_{onset} and α (including $t_{onset} = 0.0$ and $\alpha = 0.0$ corresponding to the constant population size), and a fixed sample size of n chromosomes, we generate independent random realizations following these steps:

Step 1. Simulate the parameter N_A as described above, and compute the current effective population size N_0.
Step 2. Simulate the parameters μ and c, and compute the

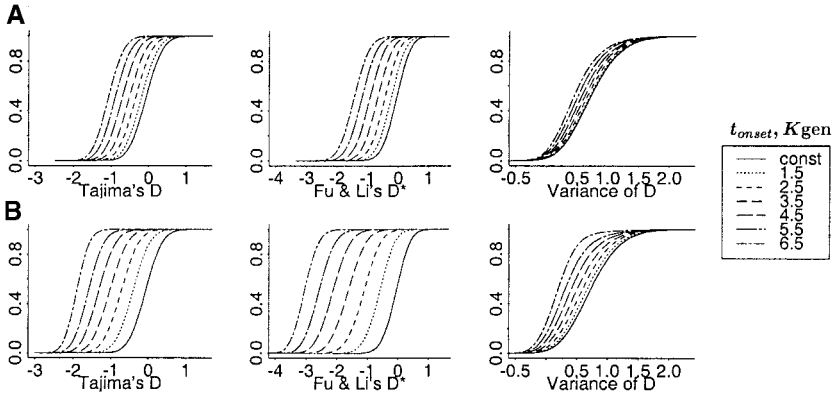

FIGURE 1.—Empirical cumulative distribution functions of Tajima's D and Fu and Li's D^* statistics for a single-locus pair and the sample variance of Tajima's D (over 10 locus pairs) for (A) $\alpha = 0.5 \times 10^{-3}$ and (B) 2.0×10^{-3}.

scaled mutation rate $\theta = 4N_0\mu L$ and recombination rate $\rho = 4N_0cL$, where L denotes the length of the sequence.

Step 3. Simulate the genealogical history with recombination events as described by HUDSON (1983).

Step 4. Simulate mutations on the genealogy assuming an infinite sites model and rates obtained in step 2.

For our preliminary investigations of the distribution of the statistics, we considered simulated samples of sequences with L = 10,000 bp, but only the mutations that fell in the two 1-kb flanking segments were considered. The polymorphic sites in the middle 8000 bp were ignored to match the structure of the locus pair data. This is referred to as simulations of the simplified data. For testing the models, a similar scheme was used except that the sequence length and distances between the sequenced segments were adjusted to match exactly the data. This is referred to as simulations of the real data. To generate samples of several genetically independent regions, for a single realization of N_A steps 2–4 were repeated for all unlinked loci in question, keeping the value of N_A the same while allowing other parameters to vary randomly from locus to locus. This effectively accounts for the mutation and recombination rate heterogeneity between loci. In addition, all realizations without polymorphic sites were discarded.

Summary statistics and hypotheses testing: Methods for using full data likelihoods are not available or feasible for the models tested here with recombination. Hence, we investigated the power of each of the following summary statistics at single-locus pairs to detect recent population growth: the mean pairwise nucleotide differences, π, the sample standard deviation of the pairwise nucleotide difference,

$$\nu = \sqrt{\left(\sum_{i \neq j} \frac{d_{ij}^2}{n(n-1)}\right) - \pi^2}$$

(HUDSON 1987), the number of distinct haplotypes in the sample, M, the haplotype diversity (sample heterozygosity), H, the number of singletons, η, Tajima's D statistic (TAJIMA 1989), and Fu and Li's D^* statistic (FU and LI 1993; see also SIMONSEN *et al.* 1995). The sampling properties of these statistics have been studied extensively by a number of authors (SIMONSEN *et al.* 1995; PRITCHARD *et al.* 1999) and were shown to be sensitive to population size changes.

In a preliminary investigation, 100,000 independent samples of locus pair sequences were generated using the parameters and procedures described above (simulations of the simplified data). The behavior of the summary statistics under different demographic scenarios was compared to choose the most informative one for detecting recent population expansion.

First, we obtained empirical distributions for the statistics under the null and alternative hypotheses and observed that, for a fixed α, they form a stochastically monotone family of distributions decreasing as t_{onset} increased, *i.e.*, went farther away into the past, until $\sim 20K$ generations, after which the direction of the change reversed (data not shown). We are interested in testing the hypothesis of a relatively recent population expansion ($t_{onset} < 5K$ generations); thus we limited our investigation to the time interval where monotonicity applies. Figure 1 illustrates this observation for two such families of empirical distributions—those of Tajima's D and Fu and Li's D^* statistics. Note that the monotonicity of these families of distributions implies monotonicity of power functions (CASELLA and BERGER 1990).

For all test statistics, we obtained empirical cutoff points for the average value of the statistic over 10 locus pairs corresponding to the 5% significance level. Likewise, the critical values of the variance of Tajima's D over 10 locus pairs were estimated. The power of each test was assessed as a function of the parameter t_{onset} for a range of fixed values of growth rate α. The empirical density functions can also be used to obtain P values for the experimental data.

The power of a test based on a summary statistic was estimated by simulating 100,000 realizations from an alternative model in question and counting the number of times the null hypothesis was rejected. The corresponding empirical power functions are shown in Figure 2. The plots clearly indicate that tests based on Tajima's D and Fu and Li's D^* statistics are by far more powerful than all other tests considered. These two statistics were used for testing the growth models. In addition, we carried out a test of the equilibrium model on the basis of the sample variance of Tajima's D; the critical values for this statistic were estimated on the basis of the same set of simulations described above. The use of this test was motivated by previous results of a similar test, which suggested a significantly large variance of Tajima's D in the Chinese sample. However, the test carried out here properly takes into

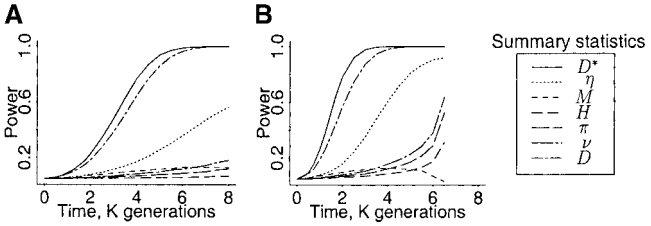

FIGURE 2.—Empirical power functions of various summary statistics averaged over 10 independent locus pairs on the basis of a one-tailed test. The null hypothesis is the equilibrium model.

TABLE 2

Results of the two-tailed multilocus tests for the 10 regions studied

Population	Average Tajima's D	P value	Variance of Tajima's D	P value	Average Fu and Li's D^*	P value
Hausa	−0.279	0.368	0.504	0.401	−0.134	0.690
Italians	0.663	0.012	1.399	0.138	0.447	0.088
Chinese	0.046	0.758	1.720	0.038	−0.085	0.818

account the structure of the data and incorporates the effect of recombination.

For each value of α and t_{onset}, a multilocus P value for the data was estimated as twice (*i.e.*, two-tailed test) the proportion of computer-generated samples with a more extreme average value of the test statistics than observed. The set of values of α and t_{onset} for which the P value was greater than a specified value constituted our estimated confidence set. Only values of α and t_{onset} such that N_0 is $<4 \times 10^9$ are considered. All multilocus P values were estimated on the basis of simulations of the real data. Note, however, that the power functions were calculated for a one-tailed test.

RESULTS

The results of testing the constant population size model are shown in Table 2. Based on the average value of Tajima's D and Fu and Li's D^*, the Hausa sample is compatible with the constant population size model. In addition, it is compatible with a set of models with recent population growth. The confidence regions for the parameters (α and t_{onset}) defining the growth model are shown in Figure 3. These are the set of parameter values for which the estimated P value is greater than the specified values 0.1, 0.05, 0.02, and 0.01. Such confidence sets show that, for an exponential growth rate as low as 1×10^{-3}/generation, the growth phase is unlikely to have started earlier than ~50,000 years ago.

As shown in Figure 3, the parameters of the growth model are interdependent: high growth rates are compatible with the data only for small t_{onset}, and models with large t_{onset} are accepted only for small growth rates. It should be noted that the expectation of N_A is varied across the confidence region plot with varying values of α and t_{onset} in such a way that the expected number of polymorphic sites is 4/kb (see MATERIALS AND METHODS). This variation of EN_A across combinations of α and t_{onset} values is exemplified in Table 3 for six points (labeled A–F) on the boundary of the 95% confidence set. These values range from 7800 to 10,300, depending on the growth rate and the test statistic applied. For any α value, EN_A decreases with decreasing t_{onset}. Hence, for all points in the 95% confidence set with $\alpha > 1 \times 10^{-3}$, EN_A must be >7800. As evident in the plot, the boundaries of the confidence sets sharply increase as α approaches zero. This implies that although a more ancient onset of growth is compatible with the data, the growth rate must be so small as to be essentially indistinguishable from the constant population size model. The dashed line in Figure 3 corresponds to points with $EN_A = 10{,}000$, a value that is often reported in human variation studies. This widely reported effective population size estimate is based on the implicit assump-

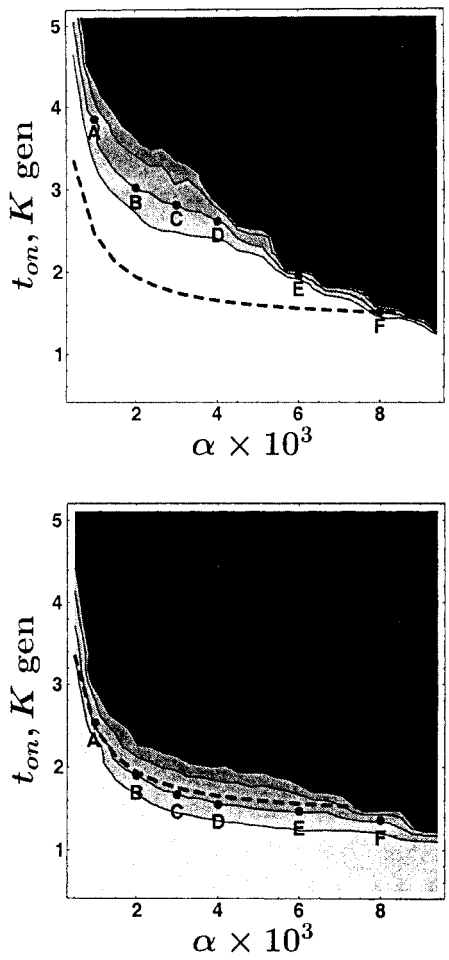

FIGURE 3.—Confidence sets for pairs of parameters (t_{onset}, α) based on Tajima's D (top) and Fu and Li's D^* (bottom) statistics, for the Hausa sample. Shown are the 0.9, 0.95, 0.98, and 0.99 levels. The points A–F on the boundary of the 95% confidence set are characterized in Table 3. The dashed line corresponds to the values of t_{onset} and α such that $EN_A = 10{,}000$. The rejection region is shaded in black. Confidence sets for the Italian and Chinese samples are not shown because these samples are incompatible with the constant population size model and the growth model for all parameter values.

TABLE 3

Combinations of parameter values on the boundary of the 95% confidence set for the Hausa sample

		Average Tajima's D			Average Fu and Li's D^*		
	$\alpha \times 10^3$	t_{onset}, K gen	EN_0, K	EN_A, K	t_{onset}, K gen	EN_0, K	EN_A, K
A	1	3.85	364	7.8	2.53	124	9.9
B	2	3.02	3,371	8.0	1.90	450	10.1
C	3	2.81	36,747	8.0	1.67	1,522	10.2
D	4	2.61	280,514	8.2	1.55	5,025	10.2
E	5	1.95	1,117,097	9.3	1.47	68,847	10.2
F	6	1.52	1,908,035	10.0	1.36	546,595	10.3

K gen, thousand generations.

tion of an equilibrium population and, as such, is not equivalent to an estimate of the ancestral population size in a model that incorporates growth.

In contrast, the Italian data show large positive average values of Tajima's D and Fu and Li's D^* across loci. The simulations showed that the observed average value of Tajima's D is too large to be compatible with the constant population size model. As expected, recent population growth shifts the distribution of these statistics toward smaller values (Figure 1). Thus, it follows that the Italian sample is also incompatible with the family of growth models tested here (for any positive growth rate). Likewise, the variance of Tajima's D across loci in the Chinese is significantly too large compared to the expectations for the constant population size model. Since we showed that the distribution of the variance of Tajima's D decreases monotonically with increasing time of onset of growth (Figure 1), these results allow us to rule out both the equilibrium model and the family of growth models tested here. These findings are consistent with the results in Frisse *et al.* (2001), which pointed to several significant departures from an equilibrium model in the non-African population samples.

DISCUSSION

Our multilocus analysis of noncoding sequences showed that the Hausa sample is compatible with a constant population size model as well as a model with recent population growth if the growth parameter and the time of initiation of growth are in a constrained range as shown in Figure 3. This figure shows a 95% confidence region based on Tajima's D and Fu and Li's D^*. For an exponential growth rate of 10^{-3}/generation, the earliest onset compatible with the observed Fu and Li's D^* in the Hausa sample is ~50,000 years ago. It should be noted that this growth rate is rather small, resulting in population size increasing only by a factor of 12 over 50,000 years. If the growth parameter is larger, the onset of growth must be more recent than 50,000 years ago.

Conversely, the equilibrium model or models with population growth from a population of nontrivial size are not compatible with the observed average value and variance of Tajima's D in the Italian and Chinese samples, respectively. More complex demographic models with population bottlenecks and/or with some degree of geographic substructure in the past may account for the non-African data.

These conclusions are consistent with the results of Frisse *et al.* (2001). However, the analyses carried out here properly take into account the structure of the data and allow for recombination within each region. As a consequence, in this analysis, the multilocus P values are somewhat lower and reach nominal levels of significance for both the Italian and Chinese samples (Table 2). It should be noted that no correction for multiple tests was carried out. However, other aspects of the data corroborate the conclusion of a departure from either the equilibrium or the simple growth models. For example, Fay and Wu's H statistic (Fay and Wu 2000) is significant at 4 of the 10 locus pairs (Frisse *et al.* 2001; Hamblin *et al.* 2002). It has been shown that models of geographic structure may account for this observation (Przeworski 2002). Also, higher-than-expected levels of linkage disequilibrium are observed in both non-African samples. These results taken together suggest that a more complicated demographic model is necessary to account for all aspects of the polymorphism data in the non-African samples. In contrast, we have yet to find an aspect of the Hausa data that is incompatible with the equilibrium model or with the simple growth model considered here.

Despite many attempts to infer the history of population size in humans, a coherent picture has yet to emerge. A synthesis of the available evidence is complicated by the heterogeneity of data used, including unlinked autosomal loci as well as nonrecombining uniparentally inherited loci such as those in the mtDNA genome and in the nonrecombining portion of the Y chromosome. A further level of heterogeneity results from the analysis of loci experiencing different mutation processes, namely nucleotide substitution and insertion/deletion (*i.e.*, microsatellites). Finally, the methods of analysis, the specific models tested, and the populations sampled vary greatly across studies.

Most human population samples show patterns of

mtDNA variation consistent with rapid population growth. These data were used to estimate the time of onset of growth to an interval that largely overlaps with our estimates for the Hausa sample (Rogers and Harpending 1992; Sherry *et al.* 1994; Ingman *et al.* 2000). However, our overall results differ from mtDNA findings in several important regards. First, mtDNA data show the most consistent signal of rapid population expansion in non-African populations (Di Rienzo and Wilson 1991; Sherry *et al.* 1994; Weiss and von Haeseler 1998; Ingman *et al.* 2000) while substantial heterogeneity exists across sub-Saharan African populations (Sherry *et al.* 1994; Watson *et al.* 1996). Interestingly, however, a different sample of Hausa showed a unimodal distribution of pairwise sequence differences, consistent with rapid population growth in the relatively recent past (21,000 years ago; Watson *et al.* 1996). Second, the patterns observed in the mtDNA data are broadly consistent with a "star-shaped" genealogy (Di Rienzo and Wilson 1991; Slatkin and Hudson 1991). Marjoram and Donnelly (1994) showed that such a pattern is expected only under a specific expansion model in which a population of very small size (*e.g.*, 500 females for mtDNA or 125 individuals for an autosomal locus) grows rapidly. This prediction is consistent with estimates of the ancestral population size (*i.e.*, before growth) obtained on the basis of a model of instantaneous growth applied to mtDNA data that range between zero and several hundred individuals (Rogers and Harpending 1992). Thus, although our estimates of the time of onset of growth are largely consistent with those obtained on the basis of mtDNA, our estimates of the ancestral population size appear to be exceedingly large compared to those based on mtDNA (even when the fourfold difference expected for uniparentally *vs.* biparentally inherited loci is taken into account).

Some interesting parallels exist between the mtDNA and the Y chromosome findings. Like mtDNA, Y chromosome loci show patterns consistent with rapid growth in most human populations (Pritchard *et al.* 1999; Thomson *et al.* 2000). A variety of methods and models have been applied to Y chromosome data to infer the time of onset of growth. Pritchard *et al.* (1999) analyzed microsatellite data under the assumption of a generalized stepwise mutation model and the same family of expansion models tested here. This analysis led to estimates of the time of onset of 18,000 years and exponential growth rate of 0.008, both consistent with our Hausa results. As with mtDNA data, however, the ancestral population size is estimated to be much smaller than our estimate for the Hausa sample, namely 900 males [95% confidence interval (C.I.) 50–3200]. To assess the compatibility of the conclusions of Pritchard *et al.* (1999) with our data, we calculated the number of polymorphic sites expected in a sample of 30 chromosomes, assuming the above parameter values. For a DNA segment of 10 kb, the parameter estimates of Pritchard *et al.* (1999) led to an expected number of 10.2 polymorphic sites (3.8–24.9), which were thus markedly less than observed at the locus pairs in the Hausa sample (*i.e.*, 47.9 based on Table 2 in Frisse *et al.* 2001). A study of Y chromosome sequences led to analogous conclusions (Thomson *et al.* 2000). A significant excess of rare variants was observed, consistent with rapid population growth. The TMRCA for a worldwide sample was estimated to be 59,000 years ago (95% central probability intervals 40,000–140,000) on the basis of a model of exponential growth throughout the history of the population. The TMRCA was estimated to be similar (70,000 years ago) on the basis of the average number of differences between each sequence and the root of the genealogy.

Overall, the data from uniparentally inherited nonrecombining loci differ markedly from our results in two main respects: the smaller-than-expected ancestral population size and the signal of growth in non-African samples. While the latter discrepancy might be reconciled by more complex demographic models, including a population size reduction before expansion (Fay and Wu 1999), the former may require nondemographic explanations such as natural selection acting on mtDNA and the Y chromosome (Di Rienzo and Wilson 1991; Thomson *et al.* 2000).

Although a number of autosomal microsatellite data sets agree in showing evidence for *some* population growth, many aspects of the results are incongruous, thus hindering any comparison to the locus pair data. Under the assumption of a more general stepwise mutation model, Kimmel *et al.* (1998) found evidence for a population size reduction followed by expansion in the non-African samples while the African sample fits the expectations of a constant population size model. These results are in qualitative agreement with the locus pair data; since no attempt was made at estimating the parameters of the model, it is impossible to compare our conclusions in greater detail. In a different study, a generalized stepwise mutation model was used to test the null model of constant population size (Gonser *et al.* 2000). A significant departure in the direction expected under rapid population growth was observed in all non-African samples as well as in one of two African samples. Under the assumption of a simple stepwise mutation model, Reich and Goldstein (1998) found some evidence for rapid population growth in 3 out of 8 population samples from sub-Saharan Africa and in none of the 12 samples from outside Africa. They estimated the maximum size of the ancestral population as 6600 individuals, possibly consistent with the Hausa results for low growth rates. The time of onset of growth was estimated to range within a 90% confidence interval of 49,000–640,000 years ago.

Probably due to the similar type of data and demographic models tested, our results are most consistent with those of Wall and Przeworski (2000) who analyzed sequence polymorphism data from eight nuclear loci. They considered one locus at a time and found

interval estimates of the time of onset of exponential growth on the basis of Tajima's *D* and Fu and Li's *D**. Only the case of an ancestral population of size 10,000 growing 10-fold and 100-fold is considered (hence, the exponential growth rate was varied with varying time of onset of growth). For the African samples, all eight loci are compatible with either no expansion or recent expansion, consistent with our results. However, their confidence intervals are quite large. In the non-African samples, four loci are incompatible with either the constant size or the growth models with respect to Tajima's *D* while two loci are consistent with the growth models only. These results suggest that the variance of Tajima's *D* in non-Africans is larger than expected under an equilibrium model as we observe in the Chinese sample.

A recent survey of sequence variation in 313 human genes showed a marked skew toward negative Tajima's *D* values (Stephens *et al.* 2001), which was interpreted as evidence in favor of population expansion. However, this analysis involved only a pooled sample of four different ethnic groups from the United States. In our data, pooling samples from different ethnic groups results in a more negative average Tajima's *D* than observed in the individual samples (data not shown). This raises the possibility that population-specific patterns of frequency spectrum are obscured in pooled samples and makes the interpretation of the global pattern questionable.

Our data and methods of analyses have several advantages over those of earlier studies. The use of single-nucleotide substitution rather than microsatellite data implies better estimates of the mutation rate at each locus and hence more reliable estimates of population parameters. Furthermore, avoiding coding regions reduces the probability that patterns of variation were shaped by natural selection rather than demography. The availability of sequence data from several independent loci in exactly the same population samples also eliminates the possibility that the observed interlocus variability is due to the different histories of the populations surveyed at different loci. Since evolutionary processes are highly stochastic, demographic inferences must of necessity rely on the analysis of many independent loci. Unless natural selection is thought to act on a specific subset of the loci, any demographic model should account for the data at *all* loci. Thus, a simultaneous analysis of multiple independent loci will lead to better estimates and more powerful tests. Accordingly, our multilocus analysis led to narrower confidence intervals and more easily interpretable results compared to those in Wall and Przeworski (2000) that were based on a set of single-locus *P* values. Also, our demographic model is more general than that of Wall and Przeworski (2000), in that a full range of ancestral population size and growth rate was considered, and is more general than a model of exponential growth throughout the history of the population. Our model is virtually identical to that of Pritchard *et al.* (1999). However, rather than providing confidence intervals for individual parameters, we obtain two-dimensional confidence sets that show the interdependence of the parameters. Unlike Wall and Przeworski (2000) and similarly to Pritchard *et al.* (1999), we incorporate the uncertainty about the mutation and recombination parameters in our estimation method.

It should be noted that, although we have rejected simple growth models for the non-African samples, the data may be consistent with other scenarios that include a growth phase as part of a more complex model. In this regard, it is interesting to note that a recent analysis of ascertained single nucleotide polymorphisms in humans supported a model that included growth in effective population size in the context of a subdivided population. However, when population subdivision was removed from the model, a simple equilibrium model could not be rejected (Wakeley *et al.* 2001). It follows that, when the equilibrium model cannot be rejected, more complex models may require some form of population growth to be compatible with the data.

We thank M. Przeworski, J. Pritchard, P. Donnelly, and S. Zoellner for comments on the manuscript. This work was supported by a National Institutes of Health grant (HG02098) to A.D.

LITERATURE CITED

Abramovitz, M., and I. A. Stegun, 1964 *Handbook of Mathematical Functions.* Dover, New York.

Brookfield, J. F., 1997 Importance of ancestral DNA ages. Nature **388:** 134.

Casella, G., and R. L. Berger, 1990 *Statistical Inference.* Warsworth & Brooks/Cole, Pacific Grove, CA.

Chen, F. C., and W. H. Li, 2001 Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. **68:** 444–456.

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

Di Rienzo, A., P. Donnelly, C. Toomajian, B. Sisk, A. Hill *et al.*, 1998 Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. Genetics **148:** 1269–1284.

Fay, J. C., and C.-I Wu, 1999 A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Mol. Biol. Evol. **16:** 1003–1005.

Fay, J. C., and C.-I Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Forsythe, G. E., M. A. Malkolm and C. B. Moler, 1977 *Computer Methods for Mathematical Computations.* Prentice-Hall, Englewood Cliffs, NJ.

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.

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Gonser, R., P. Donnelly, G. Nicholson and A. Di Rienzo, 2000 Microsatellite mutations and inferences about human demography. Genetics **154:** 1793–1807.

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

Hamblin, M. T., E. E. Thompson and A. Di Rienzo, 2002 Complex signatures of natural selection at the Duffy blood group locus. Am. J. Hum. Genet. **70:** 369–373.
Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.
Hudson, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:** 245–250.
Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.
Ingman, M., H. Kaessmann, S. Paabo and U. Gyllensten, 2000 Mitochondrial genome variation and the origin of modern humans. Nature **408:** 708–713.
Jeffreys, A. J., L. Kauppi and R. Neumann, 2001 Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat. Genet. **29:** 217–222.
Kimmel, M., R. Chakraborty, J. P. King, M. Bamshad, W. S. Watkins *et al.*, 1998 Signatures of population expansion in microsatellite repeat data. Genetics **148:** 1921–1930.
Marjoram, P., and P. Donnelly, 1994 Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics **136:** 673–683.
Nachman, M. W., and S. L. Crowell, 2000 Estimate of the mutation rate per nucleotide in humans. Genetics **156:** 297–304.
Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, 1996 *Numerical Recipes in Fortran 90: The Art of Parallel Scientific Computing*. Cambridge University Press, Cambridge, UK.
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:** 1791–1798.
Przeworski, M., 2002 The signature of natural selection at randomly chosen loci. Genetics **160:** 1179–1189.
Przeworski, M., R. R. Hudson and A. Di Rienzo, 2000 Adjusting the focus on human variation. Trends Genet. **16:** 296–302.
Reich, D., and D. Goldstein, 1998 Genetic evidence for a Paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. USA **95:** 8119–8123.
Reich, D. E., and E. S. Lander, 2001 On the allelic spectrum of human disease. Trends Genet. **17:** 502–510.
Rogers, A. R., and H. Harpending, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552–569.
Severini, T. A., 1999 On the relationship between Bayesian and non-Bayesian elimination of nuisance parameters. Stat. Sinica **9:** 713–724.
Sherry, S. T., A. R. Rogers, H. Harpending, H. Soodyall, T. Jenkins *et al.*, 1994 Mismatch distributions of mtDNA reveal recent human population expansions. Hum. Biol. **66:** 761–775.
Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.
Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.
Stephens, J. C., J. A. Schneider, D. A. Tanguay, J. Choi, T. Acharya *et al.*, 2001 Haplotype variation and linkage disequilibrium in 313 human genes. Science **293:** 489–493.
Stephens, M., and P. Donnelly, 2000 Inference in molecular population genetics. J. R. Stat. Soc. B **62:** 605–635.
Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.
Thomson, R., J. K. Pritchard, P. Shen, P. J. Oefner and M. W. Feldman, 2000 Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc. Natl. Acad. Sci. USA **97:** 7360–7365.
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.
Watson, E., K. Bauer, R. Aman, G. Weiss, A. von Haeseler *et al.*, 1996 mtDNA sequence diversity in Africa. Am. J. Hum. Genet. **59:** 437–444.
Weiss, G., and A. von Haeseler, 1998 Inference of population history using a likelihood approach. Genetics **149:** 1539–1546.
Yu, A., C. Zhao, Y. Fan, W. Jang, A. J. Mungall *et al.*, 2001 Comparison of human genetic and sequence-based physical maps. Nature **409:** 951–953.
Zhivotovsky, L. A., L. Bennett, A. M. Bowcock and M. W. Feldman, 2000 Human population expansion and microsatellite variation. Mol. Biol. Evol. **17:** 757–767.

Communicating editor: N. Takahata

APPENDIX

Here, we compute the expected value of N_A conditional on the expected number of polymorphisms S_n in the sample. We begin by deriving the expression for the expected number ES_n of polymorphic sites in a sample of haploid size n as a function of N_A and the parameters of the evolution model. First, assume that N_A and the mutation rate μ are constants rather than random variables. This assumption is not required for the recombination rate since the only property it affects is the variance of S_n.

In the infinite sites model of mutation under standard coalescent theory, S_n is identical to the number of mutations on a coalescent tree since the most recent common ancestor of the sample and is given by

$$ES_n = \frac{\theta}{2} E\mathcal{L}_n, \tag{A1}$$

where $\theta = 4l\mu N_0 = 4l\mu g(t_{\text{onset}})N_A$ is the scaled mutation rate per sequence of length l, and $E\mathcal{L}_n$ is the expected length of the ancestral tree (*i.e.*, the total length of all branches). This relationship holds for all models of demographic history of the population, in particular, for the exponential growth model in which $g(t_{\text{onset}}) \equiv G = e^{\alpha t_{\text{onset}}}$.

To find the expression for $E\mathcal{L}_n$ in terms of N_A and other parameters, we note that $\mathcal{L}_n$ can be partitioned as

$$\mathcal{L}_n = \mathcal{L}_n^c + \mathcal{L}_n^g, \tag{A2}$$

where $\mathcal{L}_n^c$ and $\mathcal{L}_n^g$ are the parts of the ancestral tree corresponding to the constant and growth phase, respectively. We observe that

$$E\mathcal{L}_n = \int_0^\infty [EA_n(t) - P(A_n(t) = 1)]\, dt, \tag{A3}$$

where $A_n(t)$ is the number of ancestors of the sample at time t in the past. For a sample from an exponentially growing panmictic population,

$$EA_n(t) = 1 + \sum_{k=2}^{n} e^{-(k(k-1)/2)\Lambda(t)} \frac{(2k-1)\, n_{[k]}}{n_{(k)}}, \tag{A4}$$

and

$$P(A_n(t) = 1) = 1 + \sum_{k=2}^{n} (-1)^{k-1}\, e^{-(k(k-1)/2)\,\Lambda(t)} \frac{(2k-1)\, n_{[k]}}{n_{(k)}} \tag{A5}$$

(Griffiths and Tavaré 1998), where $\Lambda(t) = (e^{\alpha t} - 1)/2\alpha G N_A$, $n_{[k]} = n(n-1)\cdots(n-k+1)$, and $n_{(k)} = n(n+1)\cdots(n+k-1)$. Hence, similarly to Slatkin and Hudson (1991),

$$\begin{aligned} E\mathcal{L}_n^g &= \int_0^{t_{\text{onset}}} [EA_n(t) - P(A_n(t) = 1)]\,dt \\ &= \sum_{k=2,4,\ldots,n} \frac{2(2k-1)\,n_{[k]}}{n_{(k)}} \int_0^{t_{\text{onset}}} e^{-(k(k-1)/2)\Lambda(t)}\,dt \\ &= \frac{1}{\alpha G N_A} \sum_{k=1}^{\lfloor n/2 \rfloor} \frac{(4k-1)\,n_{[2k]}}{n_{(2k)}}\, e^{k(2k-1)/(2\alpha G N_A)} \\ &\quad \times \left[E_1\!\left(\frac{k(2k-1)}{2\alpha G N_A}\right) - E_1\!\left(\frac{k(2k-1)}{2\alpha N_A}\right)\right], \end{aligned} \tag{A6}$$

where $E_1(\cdot)$ is the exponential integral (Abramovitz and Stegun 1964).

By the Markov property of the ancestral process $A_n(t)$,

$$\begin{aligned} E\mathcal{L}_n^c &= \int_{t_{\text{onset}}}^{\infty} [EA_n(t) - P(A_n(t) = 1)]\,dt \\ &= \int_0^{\infty} \sum_{k=2}^{n} E(A_n^c(t) \mid A_n^c(0) = k) P(A_n^g(t_{\text{onset}}) = k)\,dt \\ &= \frac{2}{G} \sum_{k=2}^{n} P(A_n^g(t_{\text{onset}}) = k) \sum_{j=1}^{k-1} \frac{1}{j}, \end{aligned} \tag{A7}$$

where $A_n^g(t)$ and $A_n^c(t)$ refer to the ancestral process during the growth and constant periods, respectively, and the probability distribution of $A_n^g(t_{\text{onset}})$ (Griffiths and Tavaré 1998) is given by

$$P(A_n^g(t_{\text{onset}}) = k) = \sum_{l=k}^{n} e^{-(l(l-1)/2)\Lambda(t_{\text{onset}})} \frac{(2l-1)(-1)^{l-k} k_{(l-1)} n_{[l]}}{k!(l-k)!\, n_{(l)}}. \tag{A8}$$

The factor of $1/G$ in expression (A7) is due to different timescales for $A_n^g(t)$ and $A_n^c(t)$.

For relatively small sample sizes ($n < 25$), expressions (A6) and (A7) can be easily evaluated by means of, for instance, the Numerical Recipes software package (Press *et al.* 1996) for a range of values of N_A and fixed values of μ, α, and t_{onset}. For larger n, numerical evaluation of (A7) becomes increasingly unstable due to a so-called "catastrophic cancellation" (Forsythe *et al.* 1977) in the alternating summation (A8). One way of getting around this problem is to utilize the 64-bit (quadruple-precision) arithmetic available on Sun SPARC and Power PC workstations. Another is to avoid the "exact," but unstable, calculations in (A8) altogether and instead solve numerically the system of differential equations,

$$y_i = y_i + \lambda_i y_{i-1}, \quad \lambda_i = \binom{i}{2}, \quad i = 1, \ldots, n, \tag{A9}$$

which gives rise to (A8). A Fortran 95 program that implements both of these approaches is available from the authors.

Finally, solving Equation A1 numerically with respect to N_A yields the desired result. Note that for a random, rather than fixed, N_A this gives only an approximation for EN_A; however, the approximation is sufficiently accurate due to the smoothness of ES_n as a function of N_A.