The Allele Frequency Spectrum in Genome-Wide Human Variation Data Reveals Signals of Differential Demographic History in Three Large World Populations

Gabor T. Marth,[1] Eva Czabarka, Janos Murvai and Stephen T. Sherry

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894

Manuscript received April 15, 2003
Accepted for publication September 4, 2003

ABSTRACT

We have studied a genome-wide set of single-nucleotide polymorphism (SNP) allele frequency measures for African-American, East Asian, and European-American samples. For this analysis we derived a simple, closed mathematical formulation for the spectrum of expected allele frequencies when the sampled populations have experienced nonstationary demographic histories. The direct calculation generates the spectrum orders of magnitude faster than coalescent simulations do and allows us to generate spectra for a large number of alternative histories on a multidimensional parameter grid. Model-fitting experiments using this grid reveal significant population-specific differences among the demographic histories that best describe the observed allele frequency spectra. European and Asian spectra show a bottleneck-shaped history: a reduction of effective population size in the past followed by a recent phase of size recovery. In contrast, the African-American spectrum shows a history of moderate but uninterrupted population expansion. These differences are expected to have profound consequences for the design of medical association studies. The analytical methods developed for this study, *i.e.*, a closed mathematical formulation for the allele frequency spectrum, correcting the ascertainment bias introduced by shallow SNP sampling, and dealing with variable sample sizes provide a general framework for the analysis of public variation data.

THE analysis of statistical distributions of genetic variations has a rich history in classical population genetic studies (Crow and Kimura 1970), and recent genome-scale data collection projects have positioned the field to apply, challenge, and improve traditional theory by examining data from thousands of loci simultaneously. The two most frequently studied distributions of nucleotide sequence variation are the marker density (MD), or mismatch distribution (Li 1977; Rogers and Harpending 1992; *i.e.*, the distribution of the number of polymorphic sites observed when a collection of sequences of a given length are compared), and the allele frequency spectrum (AFS; Ewens 1972; *i.e.*, the distribution of diallelic polymorphic sites according to the number of chromosomes that carry a given allele within a sample). The latter distribution is immediately applicable to the genotype data produced by projects that are characterizing a large subset of currently available single-nucleotide polymorphisms (SNPs) with measures of individual allele counts (genotypes) for three ethnic populations (http://snp.cshl.org/allele_frequency_project/). In addition to data availability, the AFS has other, analytical advantages over MD data, most notably its independence from the effects of recombination or mutation rate heterogeneity as we show below.

Modeling the distribution of allele frequency: Prior study of the AFS has been restricted to properties of summary statistics such as Tajima's *D* (Tajima 1989), or the proportion of rare- to medium-frequency alleles (Fu and Li 1993). There has been very little analysis of the general shape of observed spectral distributions. The analytical shape of the AFS, under a stationary history of constant effective population size, was derived by Fu (1995) who showed that, within *n* samples, the expected number of mutations of size *i* is inversely proportional to *i*. Important properties of the coalescent process under deterministically changing population size have been derived in publications of Griffiths and Tavare (1994a,b) and Tavare *et al.* (1997). These results show that, for the purposes of genealogy, varying population size can be treated by appropriate scaling of the coalescent time. Applying these results to obtain a formula for the allele frequency spectrum is not trivial, however, because mutations occur in nonscaled time. More recently, Wooding and Rogers (2002) derived a method called the matrix coalescent that overcomes these difficulties and calculates the AFS under arbitrarily changing population size histories. Their approach solves the problem for the general case, but leads to an involved computational procedure requiring numerical matrix inversion. In this study, we have taken a different ap-

[1]*Corresponding author:* Department of Biology, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA 02467.
E-mail: marth@bc.edu

proach. By extending Fu's result from a stationary population history to a more general shape, a profile of demographic history characterized by an arbitrary number of epochs such that the effective population size is constant within each epoch, we have arrived at a very simple, easily computable formula for the AFS. The price we pay is the lack of generality of arbitrary shapes. In many practical situations, however, these shapes can be approximated by a piecewise constant effective size profile. The advantage is a formulation that permits very rapid generation of AFS under a large number of competing histories for accurate data fitting and hypothesis testing. This result is applicable when the sites under consideration are selected randomly and the number of successfully genotyped samples is identical at each site. For the data set we are considering both of these assumptions are violated. First, the sites in question were selected for the population allele frequency characterization of a large subset of SNPs from a genome-wide map (Sachidanandam *et al.* 2001) of SNPs discovered by computational means, in large mining efforts in the public (Altshuler *et al.* 2000; Mullikin *et al.* 2000; Lander *et al.* 2001; Marth *et al.* 2003) and private (Venter *et al.* 2001) domains, numbering millions of sites. Common in these efforts is that SNP discovery was carried out in samples of a small number of chromosomes (two or three). The samples used in the discovery phase were different from the samples used in the consequent genotype characterization experiments, and they represented an unknown mixture of ethnicities. Second, because of genotyping failures, the number of successful genotypes varies from site to site, raising the question of how to compare allele counts across these sites. In this work, we propose methods to deal with these practical problems. The resulting suite of tools enables us to analyze the shape of the AFS observed in the data directly and to evaluate competing scenarios of demographic history on the basis of how well they fit the observations.

Demographic history: The reconstruction of human demographic history is of direct biological and anthropological interest. Additionally, the history of effective population size has a profound effect on important quantities such as the extent of linkage disequilibrium and is therefore important for medical association studies. There have been many attempts for demographic inference from contemporary molecular data representing different molecular mutation systems such as mitochondrial DNA polymorphisms (Di Rienzo and Wilson 1991; Rogers and Harpending 1992; Sherry *et al.* 1994; Ingman *et al.* 2000), microsatellites (Di Rienzo *et al.* 1998; Kimmel *et al.* 1998; Reich and Goldstein 1998; Relethford and Jorde 1999; Gonser *et al.* 2000; Zhivotovsky *et al.* 2000), and, more recently, SNPs in nuclear DNA (Harding *et al.* 1997; Clark *et al.* 1998; Cargill *et al.* 1999; Zhao *et al.* 2000; Reich *et al.* 2001; Sachidanandam *et al.* 2001; Yu *et al.* 2001). For both global samples of human diversity, or specific subpopulations, practically all possible simple shapes of population history have been proposed: constant effective size (stationary history), growth relative to an ancestral effective size (population expansion), size reduction (collapse), and bottleneck (a phase of size reduction followed by a phase of growth or recovery); see Figure 1. These claims as well as the underlying data have been reviewed by various authors (Harpending and Rogers 2000; Wall and Przeworski 2000; Jorde *et al.* 2001; Rogers 2001; Ptak and Przeworski 2002; Tishkoff and Williams 2002). It is generally agreed that variation patterns in mitochondrial DNA show rapid expansion of effective size in all human populations. Results in microsatellite data are less unanimous about which populations experienced expansion or what the magnitude and starting time of such demographic events were. Recent studies of SNP data sets in nuclear DNA propose the possibility of a population collapse to explain reduced haplotype diversity (Clark *et al.* 1998; Reich *et al.* 2001, 2002; Gabriel *et al.* 2002), especially in samples of European ancestry, a hypothesis consistent with our observations in the current data set.

METHODS

Allele frequency spectrum under stepwise constant effective population size: We show that, for a population evolving under the Wright-Fisher model, and under selective neutrality, the expectation for the number of mutations Ψ_i of size i, within a sample of n chromosomes under a demographic history of multi-epoch, piecewise constant effective population size is

$$E(\Psi_i) = \frac{4\mu N_1}{i} + \sum_{m=1}^{M-1}\left\{4\mu\frac{N_{m+1}-N_m}{i}\binom{n-1}{i}^{-1} \times \sum_{k=2}^{n}\left[\binom{n-k}{i-1}\sum_{j=k}^{n}\left(e^{-\binom{j}{2}\tau_m^*}\prod_{\substack{k\le l\le n\\ l\neq j}}\frac{l(l-1)}{l(l-1)-j(j-1)}\right)\right]\right\}, \quad (1)$$

where μ is the (constant) per-locus mutation rate, N_m is the effective population size in epoch m, T_m is the corresponding epoch duration, and $\tau_m^* = \sum_{l=1}^{m} T_l/2N_l$, the normalized epoch boundary time. A detailed derivation of this result is given in the appendix. The normalized distribution of these expectations according to the frequency is the *allele frequency spectrum*:

$$P_n(i) = \Pr(\text{a given segregating site is size } i \text{ in } n \text{ samples}) = \frac{E(\Psi_i)}{\sum_{j=1}^{n-1}E(\Psi_j)}, \quad i = 1, \ldots, n-1. \quad (2)$$

It is sometimes useful to consider the "full" allele frequency spectrum, $P_n^{\text{full}}(i)$, considering sizes 0 and n, *i.e.*, when all samples carry the ancestral or the derived

allele, respectively. We have verified the accuracy of the complete allele frequency spectrum derived from this formulation by coalescent simulations (supplemental Figure S1 at http://www.genetics.org/supplemental/). Three important properties of the allele frequency spectrum are clear from Equation 1. First, the expectation for a given frequency is linear under simultaneous scaling of all effective population sizes and epoch durations (*i.e.*, as long as T_m and N_m are multiplied by the same constant for each m), hence the relative frequency spectrum remains unchanged. This fact can be exploited to reduce the number of parameters that characterizes a given demographic model under consideration. Second, the expected number of mutations of a given size for more than one nucleotide site is simply the sum of the individual expectations, without regard to any possible correlation among the site genealogy of proximal sites. Therefore, our results for the expected number of segregating sites as well as the allele frequency spectrum are also valid for polymorphisms at a single locus of arbitrary sequence length, without regard to possible recombination within the locus, or for polymorphisms collected from throughout the genome. This latter consideration allows us to apply the theoretical expectations derived here for the data set examined, without regard to the amount and structure of linkage between the sites represented within the set. Third, the allele frequency spectrum is independent of the actual value of the per-nucleotide, per-generation mutation rate, as long as this rate is uniform for every site considered.

Minor allele frequency spectrum (folded spectrum): In situations where allele frequency is determined experimentally by counting the two alternative alleles within a sample of n chromosomes, it is uncertain which of the two alleles is the mutant allele. In such situations, instead of the true frequency, we work with the frequency of the *less frequent* (or *minor*) allele (Fu 1995). The distribution of minor allele frequency is described by the *folded spectrum* defined as

$$\tilde{P}_n(i) = P_n(i) + P_n(n-i), \quad \because i \le \frac{n}{2}. \tag{3}$$

By this definition, if n is even, $\tilde{P}_n(n/2) = 2P_n(n/2)$, *i.e.*, twice the value we would expect to measure, leading to a "doubling effect." This fact needs to be taken into account during the interpretation of measured data. Because in many data sets available for analysis the ancestral allelic state is currently unknown, the folded spectrum is important in practice.

Numerical calculation of the allele frequency spectrum: Frequency spectrum calculations were implemented in the C programming language. Some care must be taken when calculating the expected spectrum, because computing Equation 1 requires the evaluation of alternating sums, a source of numeric instability when the individual terms are close in value. Instability can be avoided by accurate calculation of each term. The higher the sample size, the more accurately each term has to be evaluated. We do not have a systematic way to predict the accuracy requirement as a function of sample size, hence we determined the accuracy requirement for a given sample size by trial and error. In our implementation, we have used high-accuracy numeric libraries with settable numeric precision. Our experience has been that, up to a sample size $n = 100$, a numeric precision of 100 decimal places was sufficient for our calculations. Evaluation of the allele frequency spectrum for a sample size of 1000 required a numerical precision of ~500 decimal places.

Correcting ascertainment bias: To describe the situation where polymorphic sites discovered in a set of samples are genotyped in a second, independently drawn set of samples for frequency characterization we divide the two independent groups of samples into a "discovery" group consisting of k samples and a "genotyping" group consisting of n samples. The discovery process is modeled by considering only those sites within the $n + k$ samples that are polymorphic (*i.e.*, are of size between 1 and $k - 1$) within the discovery group of depth k and discarding those sites that are monomorphic in this group, as these sites would not be considered for subsequent genotyping. The conditional probability, $P_{n|k}(i)$, that a site is of size i within the n genotyping samples given that it is polymorphic in the k discovery samples is:

$$\begin{aligned}
P_{n|k}(i) &= \Pr(\text{size } i \text{ in } n \text{ samples} | \text{size between 1 and } k-1 \text{ in } k \text{ samples}) \\
&= \frac{\Pr(\text{size } i \text{ in } n \text{ samples AND size between 1 and } k-1 \text{ in } k \text{ samples})}{\Pr(\text{size between 1 and } k-1 \text{ in } k \text{ samples})} \\
&= \frac{\sum_{l=1}^{k-1} \Pr(\text{size } i+l \text{ in } n+k \text{ samples AND size } l \text{ in } k \text{ samples})}{\Pr(\text{size between 1 and } k-1 \text{ in } k \text{ samples})} \\
&= \frac{\sum_{l=1}^{k-1} \Pr(\text{size } l \text{ in } k \text{ samples} \mid \text{size } l+i \text{ in } n+k \text{ samples}) \cdot \Pr(\text{size } l+i \text{ in } n+k \text{ samples})}{\Pr(\text{size between 1 and } k-1 \text{ in } k \text{ samples})} \\
&= \frac{1}{\sum_{l=1}^{k-1} P_k^{\text{full}}(l)} \sum_{l=1}^{k-1} \binom{k}{l}\binom{n}{i} P_{n+k}^{\text{full}}(i+l) = \frac{\sum_{l=1}^{n+k-1} P_{n+k}^{\text{full}}(l)}{\sum_{l=1}^{k-1} P_k^{\text{full}}(l)} \sum_{l=1}^{k-1} \binom{k}{l}\binom{n}{i} P_{n+k}(i+l) \\
&= C \sum_{l=1}^{k-1} \binom{k}{l}\binom{n}{i} P_{n+k}(i+l).
\end{aligned} \tag{4}$$

It is possible that a site that appears polymorphic within the k discovery samples is monomorphic within the n genotyping samples. As a result, the conditional probabilities $P_{n|k}(0)$ and $P_{n|k}(n)$ are typically nonzero, and one has to renormalize after the transformation to get the AFS. It is easy to verify that Equation 4 is also valid for calculating the folded conditional spectrum $\tilde{P}_{n|k}(i)$, as defined in Equation 3, provided that both folded spectra $\tilde{P}_k(i)$ and $\tilde{P}_{n+k}(i)$ are available. This property makes it possible to account for the ascertainment bias when only the folded allele frequency distributions are available. For the sake of completeness, we include the conditional

spectrum for the important special case, $k = 2$, *i.e.*, ascertainment within a pair of chromosomes:

$$P_{n|2}(i) = \frac{2\sum_{k=1}^{n+1} P_{n+2}^{\text{full}}(k)}{P_2^{\text{full}}(1)} \cdot \frac{(i+1)(n+1-i)}{(n+1)(n+2)} P_{n+2}(i+1)$$
$$= C(i+1)(n+1-i)P_{n+2}(i+1). \quad (5)$$

It is easy to show that under a stationary history the spectrum is a linear function of i, and the folded spectrum is constant (Figure 2a).

We point out that our method of ascertainment bias correction improves on an earlier method based on using the measured discrete allele frequency as an estimator for the overall allele frequency within the population (Sherry *et al.* 1997; see supplemental Figure S2 at http://www.genetics.org/supplemental/).

Reduction of allele frequency counts to equivalent counts at a lower sample size: Often allele frequency data are the result of genotyping a target number, n_t, of individuals at a collection of polymorphic sites. Because of genotyping failures, however, the actual number of genotypes available at different locations is smaller and often varies from site to site. At sites where an identical number, n, of successfully determined chromosomal allelic states are available we denote the distribution of allele counts by $C_n(i)$ and the corresponding probability distribution obtained by normalizing these counts by $P_n(i)$. Sites with different numbers of successful genotypes are not directly comparable. To enable joint analysis of allele counts observed at all sites genotyped in the experiment, we have devised a procedure that, given an observed distribution of allele frequencies among samples, produces an equivalent distribution at a lower sample size, m. This is achieved by, first, considering all possible choices of m subsamples selected from the total n available samples, in such a way that each choice is equally likely and, second, requiring that the total number of observations remains the same. Under these assumptions, the "equivalent" allele counts, $\overline{C}_m(i)$, for m subsamples are

$$\overline{C}_m(i) = E(C_m(i)) = \sum_{j=i}^{n-m+i} \frac{\binom{m}{i}\binom{n-m}{j-i}}{\binom{n}{j}} C_n(j), \quad i = 0, \ldots, m, \quad (6)$$

$$\overline{P}_m(i) = \sum_{j=i}^{n-m+i} \frac{\binom{m}{i}\binom{n-m}{j-i}}{\binom{n}{j}} P_n^{\text{full}}(j), \quad i = 0, \ldots, m. \quad (7)$$

Note that this procedure does not allow one to generate a higher sample size distribution on the basis of a lower sample size distribution. Also note that, even if the higher sample size distribution was a relative allele frequency spectrum, the resulting lower sample size distribution will contain nonzero terms for size 0 and for size m. Clearly, the first case is the result of the possibility that the omission of $n - m$ chromosomes left us with 0 mutant alleles, and the second is that only mutant alleles remained. This results in a slight reduction of the total number of relative counts as compared to the original observations. To obtain the AFS, one omits sizes 0 and m in Equation 7 and renormalizes. It is easy to verify that the equivalence reduction also works for the folded allele frequency distribution.

We point out that our reduction procedure is not equivalent to frequency binning, a procedure sometimes employed to compare allele counts available at different samples sizes. Aggregating discrete allele frequency data on the basis of a nominal allele frequency c/n, the ratio of allele counts and the sample size, results in data distortion stemming from two sources. First, for a given sample, the inherent base frequency is $f_n = n^{-1}$. In general, only window sizes that are integer multiples of f_n will preserve the uniform appropriation of allele sizes into frequency bins. This may be impossible if multiple sample sizes are present in the data. Second, sites with identical nominal allele frequencies but different sample sizes are not equivalent; *e.g.*, a site with a minor allele count of 1 in 3 samples is clearly not equivalent to a site with a minor allele count of 10 in 30 samples. Distortions from both sources are most pronounced at lower sample sizes. Our equivalence reduction procedure is a technique of data aggregation that is free of such distortions. This point is further illustrated in supplemental Figure S3 at http://www.genetics.org/supplemental/, where we compared the AFS resulting from simple binning of all available data for the European samples to the AFS we obtain by the equivalence data reduction procedure presented here.

Coalescent simulations and tabulation of linkage disequilibrium: We used coalescent simulations to verify the accuracy of our allele frequency spectrum calculations (supplemental Figure S1), to tabulate measures of linkage disequilibrium, and to tabulate distributions of mutation age. To perform these simulations, we have implemented a widely used, direct coalescent algorithm (Hudson 1991). The simulation software was first implemented in Perl for rapid coding and error checking and then reimplemented in C++ for increased computational speed. To verify the direct formula, we have run coalescent simulations under a variety of population history scenarios, tabulated the allele frequency spectra, and compared them to the computed predictions. To verify the conditional spectrum calculations, we have simulated $n + k$ chromosomes within a common genealogy, designated k samples as the discovery group, and n samples as the genotyping, or frequency measurement, group. Of all the sites that were polymorphic within the $n + k$ samples, we discarded those sites that were monomorphic within the k discovery samples and kept the remaining sites. We then tabulated the allele frequency counts at these sites among the n genotyping samples.

Expectations for the extent of linkage disequilibrium were generated according to a previously published method (Kruglyak 1999). For each population, we

used the best-fitting three-epoch model for the coalescent simulations, with samples size $n = 100$. Marker allele frequencies were restricted to the range between $0.25n$ and $0.75n$. For each value of recombination fraction, we tabulated r^2, a commonly used measure of linkage disequilibrium defined as

$$r^2 = \frac{(p_{AB} - p_A \cdot p_B)^2}{p_A \cdot p_a \cdot p_B \cdot p_b}, \quad (8)$$

where A and a denote the mutant and the ancestral alleles at the first marker location, and B and b are the alternative alleles at the second marker location. The quantities p_A, p_a, p_B, and p_b are the corresponding allele frequency measurements, and p_{AB} is the measured frequency of the haplotype defined by the combination of allele A at the first marker position and B at the second marker position. Finally, marker age was tabulated by registering the time of occurrence for each of the mutations during the simulations.

Model fitting to observed allele frequency spectra: The primary objective of the fitting experiments is to determine the distribution of the posterior probability of the model parameters given the observed data: P(model|data). With the help of our closed formula for the direct calculation of the AFS we were able to generate the expected AFS for a complete, high-resolution, multidimensional grid overlaid on the parameter space that we intended to explore. This direct approach yielded the likelihood distribution, P(data|model), computed at each grid point. Given that there is no sensible way to assign an "informed" prior distribution to the model parameters, the distribution of the likelihood function is equivalent to the posterior distribution and can be used in ranking competing parameters. We point out that an alternative method of achieving the same goal is to use a Markov-chain Monte Carlo (MCMC) technique to obtain the posterior distribution (Griffiths and Tavare 1994a; Kuhner *et al.* 1995). We opted for the direct method because it was simple but computationally feasible, by its nature avoided the convergence issues usually associated with MCMC, and allowed us to evaluate the likelihood function at every grid point, for each of the three population-specific AFS analyzed.

Stepwise constant models of one, two, and three epochs were considered. For each model class defined by the number of epochs, a vector of parameters describing the model was considered, including the effective population size and the duration of the epoch (expressed in terms of generations). We have sampled each effective size parameter, N_i, between 1000 and 150,000 in steps of 1000 up to 30,000 and in steps of 5000 beyond 30,000, and each epoch duration parameter, T_i, between 100 and 50,000 in steps of 100 up to 10,000 and in steps of 500 beyond 10,000. Because of the scaling equivalence of the relative distribution discussed earlier, we fixed the ancestral size (the effective size of the epoch farthest in the past) parameter at 10,000, for each model class. We have generated the unbiased allele frequency spectra by direct calculation using Equation 1, for a sample size of $m + 2$, where $m = 41$ is the (common) sample size after data reduction, and $k = 2$ is the discovery size. We then computed the conditional spectrum using Equation 4. Finally, we folded the spectrum using the definition given in Equation 3. To quantify the degree of fit between a given model and the observations we have used the likelihood of the observed data conditioned on the model:

$$P(\text{data}|\text{model}) = \binom{c}{c_1, \ldots, c_{m-1}} \prod_{i=1}^{m-1} p_i^{c_i}. \quad (9)$$

For generating the likelihood surface for the European bottleneck size *vs.* duration we used the χ^2 metric defined as

$$\chi^2 = \sum_{i=1}^{m-1} \frac{(c_i - c \cdot p_i)^2}{c \cdot p_i}. \quad (10)$$

In the above notations, c_i is the observed number of sites of size i, c is the number of total sites, p_i is the predicted (relative) probability of size i, and m is the common sample size to which all observations were reduced using the equivalence data reduction procedure outlined earlier.

Comparison between models with different epoch numbers: Models within the same structure (same epoch number) could be directly compared on the basis of any of the three goodness-of-fit metrics discussed above. Models with different numbers of epochs were compared using methods of normal hypothesis testing for nested models (Ott 1991), on the basis of the likelihood of the data given each of the two models compared. The quantity $2 \ln(\lambda) = 2 \ln(P(\text{data}|\text{model}_1)/P(\text{data}|\text{model}_2))$ is asymptotically χ^2 distributed, with degrees of freedom equal to the difference in the number of parameters characterizing the models (*i.e.*, adding one extra epoch increases the number of parameters by two). The larger this quantity, the more significant the improvement that was achieved by the introduction of the extra epoch. If the quantity is small, the improvement in data fit does not warrant the introduction of the extra parameters.

RESULTS

Modeling allele frequency: We considered a diploid population whose demographic history was described by a series of epochs such that the effective population size was stepwise constant within each epoch (*e.g.*, Figure 1) and showed that the expected number of samples carrying a mutant allele can be described by a closed, easily computable mathematical formulation (see methods). We derived a method for incorporating the same *frequency ascertainment bias* into AFS models that was introduced into real data by the sampling strategies used during SNP discovery and for revealing the strate-

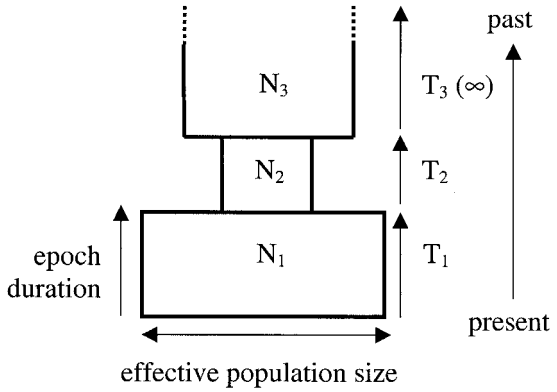

FIGURE 1.—Example of a three-epoch, piecewise constant, bottleneck-shaped population history profile. The ancestral effective population size (N_3) is followed by an instant reduction of effective size (N_2). The duration of this epoch is T_2 generations. This is followed by a stepwise increase of effective population size to N_1, T_1 generations before the present.

gies's consequent effect on SNP population frequency (METHODS). We illustrate the effect of this bias under different values of ascertainment sample size (Figure 2a). As expected, the bias toward sample enrichment for common polymorphisms is strongest when SNPs are discovered in a pair of chromosomes, and it gradually disappears as discovery sample size increases. Under a stationary population history, the folded spectrum under ascertainment in two chromosomes is a constant function of frequency (METHODS), and deviations from a horizontal line signal a nonstationary history that is easy to detect and interpret. In Figure 2b, we contrast the ascertainment bias-corrected, minor allele frequency spectra for notable, competing scenarios of demographic history. When a population expands, an increasing number of chromosomes simultaneously incur new mutations, which results in an overabundance of rare alleles in the spectrum. Conversely, a population collapse is a rapid loss of chromosomes, and the alleles present at high frequency are more likely to be carried by surviving chromosomes than are their rare counterparts. For that reason a collapse generates an overrepresentation of common alleles. Finally, AFS under a bottleneck history (a reduction of effective size followed by a phase of recovery) carries the signature of both the phase of collapse (a valley at intermediate frequencies) and that of growth (elevated signal at low frequencies).

We report a procedure to transform allele counts at a given sample size to a lower, target sample size (METHODS). Using this *equivalence sample size reduction procedure*, allele count observations at all sites can be reduced to the equivalent counts at a lower, "common denominator" sample size, as illustrated in Figure 3. This procedure is useful for analyzing allele counts at sites where the number of available genotypes is variable either because a fraction of attempted genotyping experiments failed or when merging data sets in which the attempted sample sizes are different. In such cases one selects a target sample size and applies the reduction procedure to transform allele counts observed at higher sample sizes to the equivalent counts at this lower target sample size. It is then possible to fit the resulting single AFS containing the contribution of all available data instead of fitting multiple, often sparse spectra, one for each sample size present in the data.

Minor allele frequency spectra observed in samples representing different world populations show differential demographic histories: The SNP Consortium (http://snp.cshl.org), an organization formed primarily for the discovery of a large set of human SNPs, has made well over 1 million polymorphic sites available in the public domain (SACHIDANANDAM *et al.* 2001). Most of these SNPs were discovered by comparing sequencing read fragments from multi-ethnic, anonymous, whole-genome shotgun subclone libraries to the public genome reference sequence (SACHIDANANDAM *et al.* 2001); *i.e.*, the vast majority of the SNPs were found in a discovery size of two chromosomes ($k = 2$). Quasi-random subsets of these candidate sites were then selected for frequency characterization in samples representing European-American, African-American, and East Asian populations (for sample identifiers see http://snp.cshl.org/allele_frequency_project/panels.shtml). In this study, we chose the largest data set of allele frequency counts resulting from genotypes provided by Orchid Biosciences, of 42 individuals (84 chromosomes) drawn from each of the three populations (http://snp.cshl.org/allele_frequency_project/). Experimental results were reported for 33,538 sites. For a significant fraction of the sites genotyping was unsuccessful for one or more of the populations attempted. In some other cases, although genotyping was successful, all samples carried the same allele and hence the site could not be confirmed as polymorphic. For the purpose of our study, we restricted our attention to those sites where (1) genotyping from each of the three sample groups was successful (genotyping for a given population was considered successful if genotype data were obtained for at least half the population samples, *i.e.*, 21 individuals, even if only one of the alternative alleles was seen in that population) and (2) the site was polymorphic within at least one of the three population samples. Of the total 21,407 sites that were successfully genotyped in all three populations the European samples were polymorphic at 18,660 sites, the African samples at 20,587 sites, and the Asian samples at 17,369 sites. At a given site, the total number of alleles counted varied between 42 (the minimum number possible, in case only 21 diploid individuals were successfully genotyped within a population) and 84, the maximum possible if all 42 individuals within a population sample were successfully genotyped. To use all the data available, we have applied our equivalence sample size reduction procedure (METHODS) to convert the allele count data to a common denominator

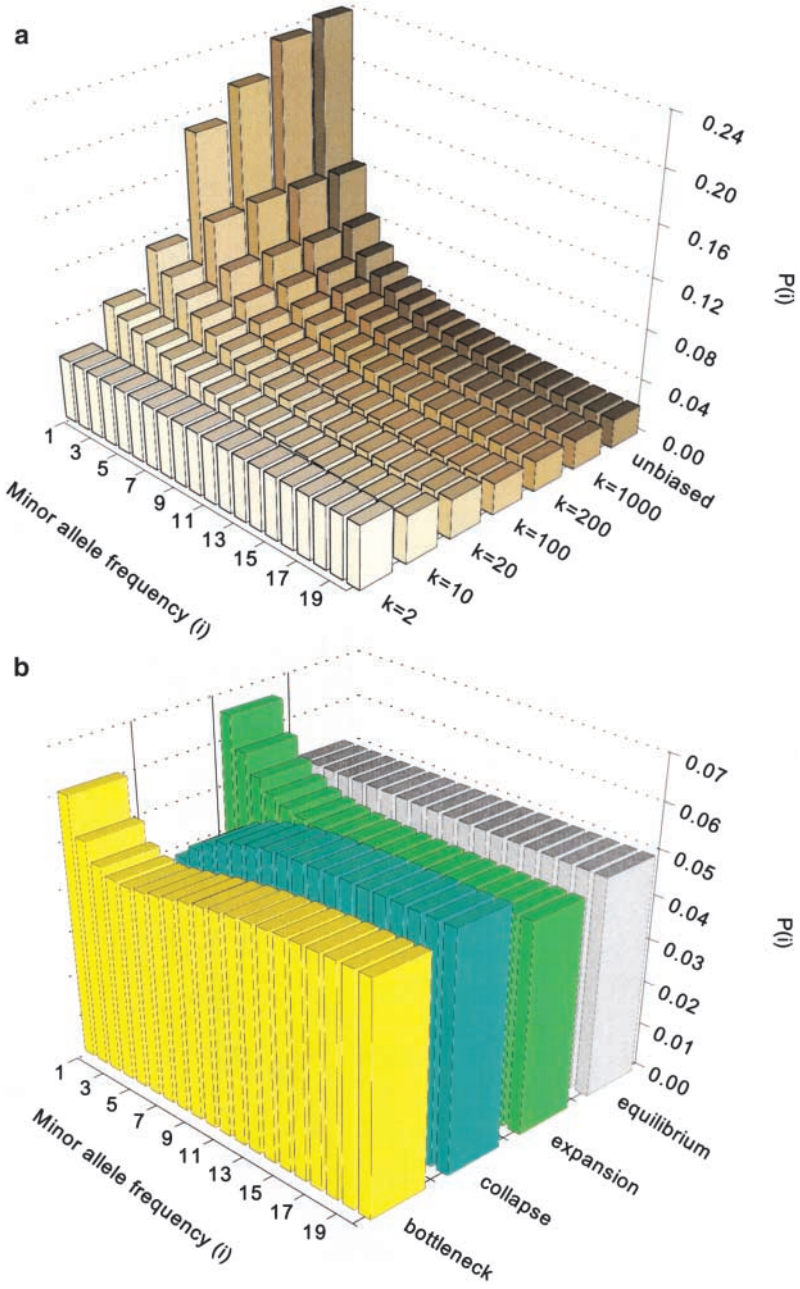

FIGURE 2.—Ascertainment bias. (a) Folded spectra under stationary history, at various values of "discovery sample" size k (METHODS). (b) Allele frequency spectra predicted under competing scenarios of population history (conditioned on pairwise ascertainment $k = 2$). Equilibrium history, $N_1 = 10{,}000$; expansion, $N_1 = 20{,}000$, $T_1 = 3000$, $N_2 = 10{,}000$; collapse, $N_1 = 2000$, $T_1 = 500$, $N_2 = 10{,}000$; bottleneck history, $N_1 = 20{,}000$, $T_1 = 3000$, $N_2 = 2000$, $T_2 = 500$, $N_3 = 10{,}000$. (a and b) Sample size $n = 41$.

sample size. Because the identity of the ancestral and the mutant allele was not known, we used the allele counts of the less frequent (or minor) allele, giving rise to a folded spectrum (METHODS). To avoid the "doubling" effect associated with folding the allele frequency spectrum when the sample size is an even number, as described in METHODS and in particular by Equation 3, we chose the common denominator sample size as $m = 41$, *i.e.*, the first odd number below the (even) sample size 42. The unfolded spectrum hence lies between 1 and 40 (sizes 0 and 41 indicate monomorphisms). Accordingly, the folded spectrum lies between minor allele sizes 1 and 20, for each of the three population-specific sample groups (Figure 4, first column). The allele frequency data used in our analysis are available through our web site: www.ncbi.nlm.nih.gov/IEB/Research/GVWG/AFS-2003/.

To assess the signals of population history within these observed distributions, we generated allele frequency spectra as predicted under competing scenarios of population history of varying complexity: stationary history (one epoch), expansion or collapse (two epoch), and all possible shapes of three-epoch histories (METHODS). For a given set of model parameters, we generated the corresponding theoretically predicted, ascertainment bias-corrected minor allele frequency spectrum and evaluated the degree of fit between the prediction and the observations (METHODS). For each population-specific data set and for each model structure (number of epochs), we determined the best-fitting model parame-

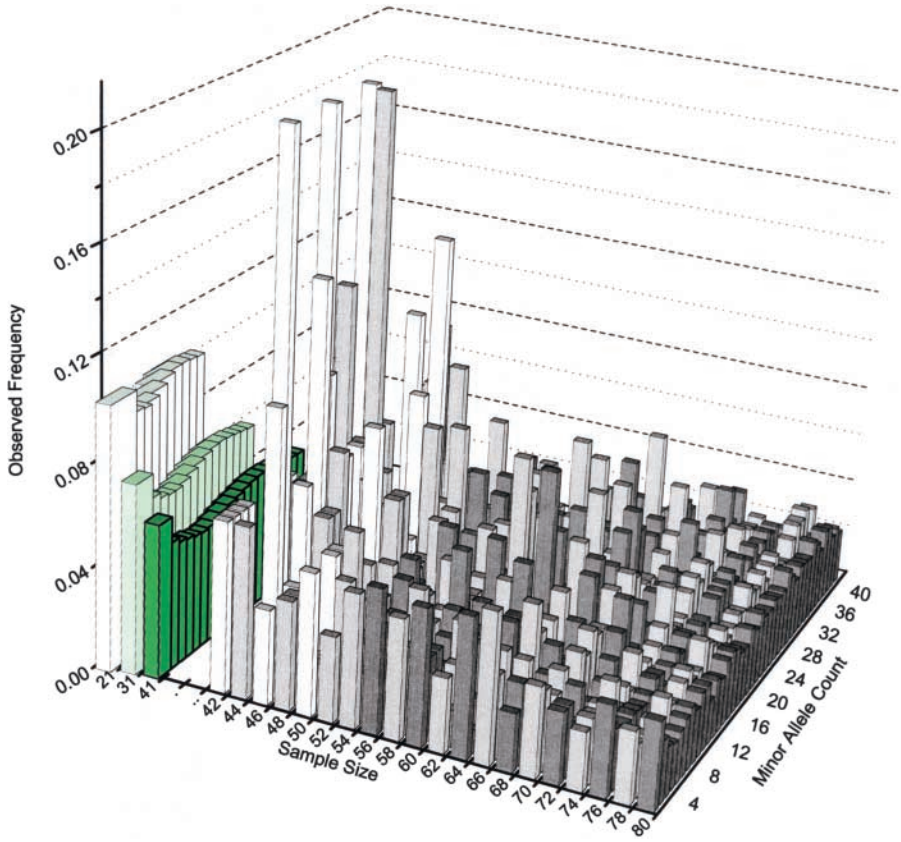

FIGURE 3.—Sample size reduction. Folded, normalized allele frequency distribution for each sample size (n = 42, . . . , 84) present in the European allele count data (gray) is shown. The allele frequency spectra obtained using the equivalence sample size reduction technique (METHODS) are also shown for various equivalence sample sizes (m = 21, 31, and 41; green).

ters and the corresponding measures of goodness of fit. By definition of the likelihood function used for data fitting, the best-fitting model parameters are the maximum-likelihood parameter estimates for that model class (Table 1).

The normalized observed allele frequency distributions for each population group and the corresponding best-performing distributions within each model class are shown in Figure 4. In all three population-specific spectra, stationary history is a poor descriptor of the data, both by visual inspection and by examination of the fit values in Table 1. The best-fitting two-epoch model for all three spectra is that of expansion (Table 1). In the European (Figure 4a) and in the Asian (Figure 4b) samples the best-fitting three-epoch model is one of a bottleneck-shaped history. In the European data, the curve fit produced by the bottleneck profile is a very significant improvement over that produced by histories of expansion. In the Asian data, the improvement is still significant but to a lesser degree. The best-fitting three-epoch models in African-American data (Figure 4c) represent a two-step population increase of moderate size.

In addition to the best-fitting models, a range of parameter values produced comparably good fit to the observations. We have examined parameter sets that produced likelihood values that were at least 90% of the value obtained for the best-fitting three-epoch parameter set. Analysis of these "close to optimal" parameter values in the European data shows that both the size (N, effective number of individuals) and duration (T, generations) of the recovery phase was within a narrow range (N_1 = 19,000–21,000, T_1 = 2700–3000). Parameters of the bottleneck phase were in a wider range (N_2 = 1000–4000 and T_2 = 200–1300), with several alternative pairs available: longer but less severe bottlenecks or shorter, more severe bottlenecks. Given the potential interest in a possible bottleneck in the history of European populations, we further investigated the strength of the bottleneck signal by fixing the recovery size and duration parameters (N_1 = 20,000, T_1 = 3000) and varying the bottleneck size N_2 and duration T_2 in fine increments (20). For each parameter combination, we evaluated the goodness of fit to the European spectrum as measured by the χ^2 statistics and reported the resulting probability surface in Figure 5. The best-fitting parameter combinations (ones not rejected by the χ^2 test even at the 99.8% level) lie on a slightly curved line between the following pairs: effective size of 1040 during the bottleneck for 240 generations and effective size 2320 for 560 generations. The most likely model, at this resolution, is a bottleneck effective size of 1560 for 360 generations. These values and the ratio of effective population size and bottleneck duration being nearly constant in a large region are in good agreement with previous reports (REICH *et al.* 2001). In the Asian data (Figure 4b), all parameters including those characterizing the bottleneck phase were within a tight range: N_2 = 3000–5000, T_2 = 600–1000, N_1 = 24,000–26,000, and T_1 = 3000–

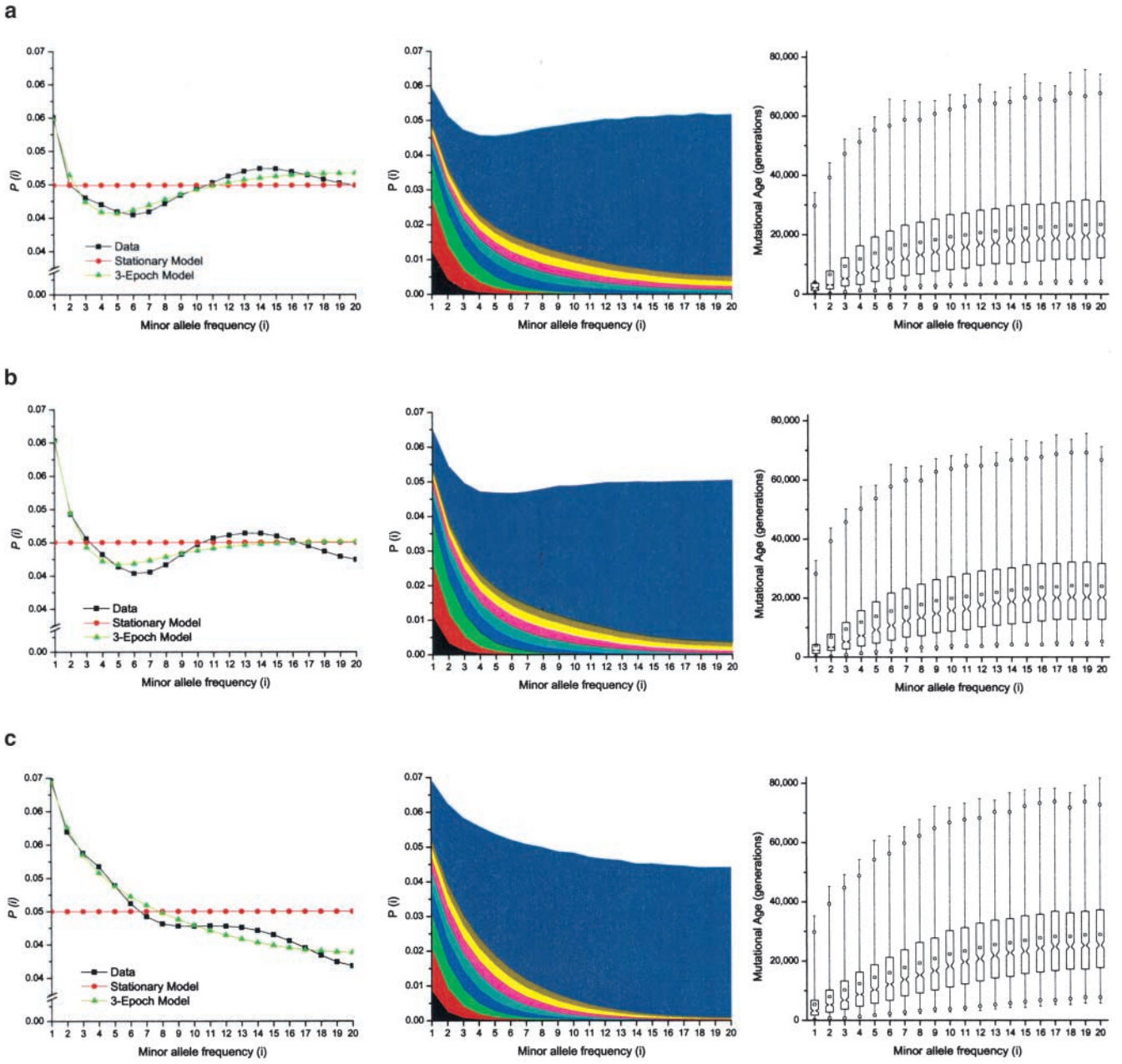

FIGURE 4.—Model fitting to folded AFS observed in population-specific genotype data reduced to common sample size, m = 41. (a) European spectrum. (b) Asian spectrum. (c) African-American spectrum. First column, observed allele frequency spectrum (black), best-fitting three-epoch theoretical model prediction (green), and prediction under stationary effective size (red); second column, breakdown of mutations according to age within each frequency class of the best-fitting model spectra [color bands correspond to a range of 1000 generations (*e.g.*, black band, 1–1000 generations; red band, 1001–2000 generations)]; third column, distribution of mutation times (generations in the past) at each frequency, based on 1 million simulation replicates. Notched box: 25%, median, 75%. Whiskers: min/max values. Open square: mean value. Open circle: 5%, 95% values.

3200. Similarly narrow ranges were observed for the African-American data (Figure 4c): N_2 = 16,000, T_2 = 13,000–15,000, N_1 = 26,000–30,000, and T_1 = 2000–2600.

DISCUSSION

Significance of the allele frequency analysis methods presented here: Equation 1 (METHODS) provides a simple and rapid way to generate expected distributions of allele frequency under stepwise constant models of effective population size history. This procedure is orders of magnitude faster than tabulating simulation replicates, especially for large sample sizes, permitting fast generation of model spectra to explore large parameter spaces at high resolution. The method of ascertainment bias calculation we have presented permits the interpretation of allele frequency spectra measured at polymor-

TABLE 1

Results of fitting multi-epoch models of allele frequency spectrum to population-specific observed allele frequency data

Model structure	Model parameters	Resulting pairwise θ (units of 10^{-4})	ln *P*(data|model)	Improvement over lower-epoch model
		a. European data		
One epoch	$N_1 = 10{,}000$	8.00	−55.98	—
Two epoch	$N_2 = 10{,}000$	8.74	−38.11	$2 \ln \lambda = 35.74$
	$N_1 = 140{,}000$			$P < 10^{-4}$
	($T_1 = 2{,}000$)			Highly significant
Three epoch	$N_3 = 10{,}000$	7.88	−23.72	$2 \ln \lambda = 28.78$
	$N_2 = 2{,}000$			$P < 10^{-4}$
	($T_2 = 500$)			Highly significant
	$N_1 = 20{,}000$			
	($T_1 = 3{,}000$)			
		b. Asian data		
One epoch	$N_1 = 10{,}000$	8.00	−74.26	—
Two epoch	$N_2 = 10{,}000$	8.63	−31.95	$2 \ln \lambda = 84.62$
	$N_1 = 50{,}000$			$P < 10^{-4}$
	($T_1 = 2{,}000$)			Highly significant
Three epoch	$N_3 = 10{,}000$	8.24	−26.39	$2 \ln \lambda = 11.12$
	$N_2 = 3{,}000$			$P = 0.0039$
	($T_2 = 600$)			Significant
	$N_1 = 25{,}000$			
	($T_1 = 3{,}200$)			
		c. African-American data		
One epoch	$N_1 = 10{,}000$	8.00	−197.86	—
Two epoch	$N_2 = 10{,}000$	9.20	−28.69	$2 \ln \lambda = 338.34$
	$N_1 = 18{,}000$			$P < 10^{-4}$
	($T_1 = 7{,}500$)			Highly significant
Three epoch	$N_3 = 10{,}000$	10.29	−26.72	$2 \ln \lambda = 3.94$
	$N_2 = 16{,}000$			$P = 0.1395$
	($T_2 = 15{,}000$)			Not significant
	$N_1 = 26{,}000$			
	($T_1 = 2{,}400$)			

phic sites selected from existing variation resources. Our procedure of equivalence sample size reduction enables the analysis of realistic data sets with genotyping failures. All three of the above procedures are firmly rooted within the coalescent framework. Model calculations directly correspond to experimentally observable quantities, without referencing directly unobservable quantities such as the overall population frequency of alleles. The data-fitting methodology is conceptually simple and allows direct comparison of the degree of fit between each of the three population samples examined, at each grid point (parameter combination).

Differential population histories in the three sample sets: On the basis of the goodness of fit between models and observations (Table 1), a history of stationary population size can be confidently rejected for all three sets of samples. Introduction of even very simple dynamics into the history has dramatically improved data fit. There were large differences among the allele frequency spectra observed in the three populations (Figure 4 and Table 1). Clearly, the shapes of the European and the Asian spectra are closer to each other than either is to the shapes of the African spectra. On the basis of the three-epoch models, both the European and the Asian data are best explained by bottleneck-shaped histories, whereas the best-fitting third-order model for the African-American data is a continued expansion. The results of hierarchical model testing (METHODS) in Table 1 show that the inclusion of the third epoch did not significantly improve the fit to the African-American data. However, the bottleneck history is a dramatic improvement over the best-fitting two-epoch growth models in both the European and Asian data. Considering the range of models that produced close to optimal fit values, but using a fixed, 20-year generation time, the European bottleneck represented a 2.5- to 10-fold decline in population size, lasting 200–1300 generations [4–26 thousand years (KY)]. This was followed by a phase of 5- to 20-fold population expansion, starting 2700–4300 generations (54–86 KY) ago. The Asian bottleneck rep-

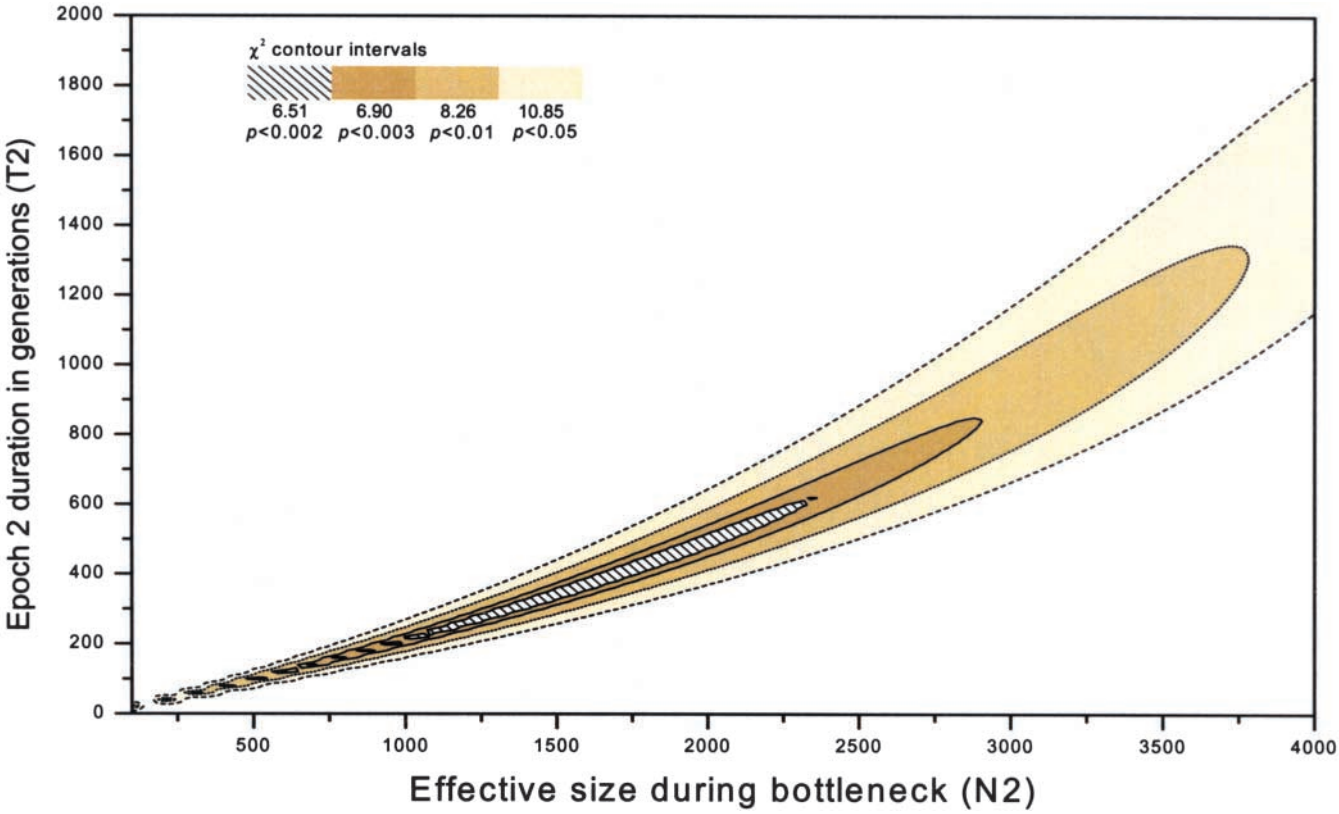

FIGURE 5.—Bottleneck size and duration in the European samples. The probability surface of the effective size and the duration of a bottleneck are shown. Size of the ancestral epoch is fixed at $N_3 = 10{,}000$, size of the present epoch is fixed at 20,000, and the duration of the present epoch is fixed at $T_1 = 3000$. Parameter regions indicated by shading fall into the same bin of significance. Note that the *P* values indicated are the direct χ^2 probabilities (*i.e.*, 1 minus the tail probability).

resented a 2- to 3-fold decline for 600–1000 generations (12–20 KY), followed by 5- to 8-fold growth starting 3000–4200 generations (60–84 KY) ago. The best-fitting models for the African-American data represent uninterrupted growth of effective population size, with the expansion clearly starting earlier than is evident in our European or the Asian data.

Earlier mitochondrial and microsatellite studies report data that are predominantly consistent with expansion-type histories of effective population size. The main evidence that points to expansion is negative values of Tajima's *D* and an excess of low-frequency alleles. The start of such expansion is estimated between 30 and 130 KYA (HARPENDING and ROGERS 2000). Nuclear data, especially in samples of non-African origin, seem to show a different pattern, an excess of common variants (HEY 1997; CLARK *et al.* 1998; REICH *et al.* 2001, 2002). Simulation results have suggested that a bottleneck-shaped history of effective population size consisting of a phase of collapse followed by a recent phase of size recovery can reconcile this seeming contradiction between observations from different mutation systems (FAY and WU 1999; HEY and HARRIS 1999). These studies characterize bottleneck-shaped histories by a size expansion ratio (in our notation N_1/N_2) and a bottleneck severity index (in our notation T_2/N_2) and consider moderate bottlenecks where the expansion ratio is 20 and the severity index is in the range of 0.25 and 4.0. Our own estimates (expansion ratio 5–20 for Europeans, 5–8 for Asians, and severity index of ~0.2 for both populations) are in general agreement with these values and signify bottlenecks on the less severe end of the spectrum. Our estimates for the start of the recovery phase (54–86 KYA for Europeans, 60–84 KYA for Asians) are well within the range of the mitochondrial and microsatellite estimates. The fact that our best-fitting two-epoch models indicate expansion-type histories for all three populations we examined is also consistent with conclusions from mitochondrial and microsatellite data. A valuable reality check of an inferred demographic model is its implied pairwise nucleotide diversity value, θ. Although our data-fitting analysis of the relative spectrum does not provide absolute estimates for θ, these values can be obtained on the basis of the best-fitting models by fixing the ancestral size N_3 and mutation rate μ. For each of the three populations, we use a common ancestral effective size of 10,000 and common mutation rate of 2×10^{-8} [a value that lies between recent, prominent estimates for average per-nucleotide, per-generation human mutation rate (NACHMAN and CROWELL 2000;

Kondrashov 2003)]. This leads to an estimate of $\theta = 7.88 \times 10^{-4}$ for the European model, in good agreement with previously reported values for other genome-wide data sets (Sachidanandam *et al.* 2001; Venter *et al.* 2001; Marth *et al.* 2003). The prediction from the Asian data is slightly higher, 8.24×10^{-4}. The pairwise θ predicted by the best-fitting model for the African-American data is 10.29×10^{-4}, significantly higher than that observed within the European and Asian samples, and in agreement with the general consensus that nucleotide diversity is higher in sub-Saharan samples than in non-African data (Relethford and Jorde 1999; Przeworski *et al.* 2000; Jorde *et al.* 2001; Tishkoff and Williams 2002). All three estimates are well within realistic values, lending further credence to the validity of our model parameters.

A bottleneck-shaped history was also our best-fitting three-epoch model structure for MD distributions observed in overlap fragments of public genome clone data (Marth *et al.* 2003). However, the parameter estimates are significantly different between these two studies. Our estimates from MD data indicated a less severe bottleneck of nearly identical duration and a shorter phase of recovery of more modest size as compared to the AFS in the European samples. Multiple factors may contribute to these differences. First, the DNA samples for the two studies came from different donors. Second, some fraction of the large-insert clones sequenced for the construction of the public genome reference sequence originate from libraries that are not of European origin [although there appears to be an overrepresentation of European sequences (Weber *et al.* 2002), presumably due to the origin of a single bacterial artificial chromosome library with the largest contribution]. If indeed an appreciable fraction of the data represents sub-Saharan DNA, the resultant MD in these mixed data could indicate a less severe bottleneck than would have been evident in a distribution containing only European data.

To understand the consequences of the differential histories that best describe the three population-specific data sets, we have partitioned the corresponding frequency spectra according to the age of the mutations (methods) that gave rise to the polymorphisms (Figure 4, second column). According to these tabulations, 35.9% of the European polymorphisms originated in <10,000 generations, as did a similar fraction, 34.9%, in the Asian model. In contrast, only 29.6% of the African mutation are younger than 10,000 generations. This indicates that the bottleneck events that explain the European and Asian data have eliminated a large fraction of the polymorphisms that predated these events, and a larger fraction of current polymorphisms are of a more recent origin as compared to the African data. This effect is most visible at the common end of the spectrum: only a negligible fraction of the common African SNPs are young, but an appreciable fraction of common European and Asian SNPs have originated <10,000 generations ago and have drifted to high population frequency. Finally, the third column of Figure 4 shows the average age of SNPs at given frequencies, confirming that SNPs at a higher frequency are expected to be older than SNPs at lower frequencies. Also, in each frequency class, the expected age of African SNPs is substantially higher than that of European or Asian SNPs, corroborating earlier observations noting the more ancient origins of African SNPs.

The differential demographic histories of the three populations examined also have important consequences for the extent of allelic association in the human genome, when the different populations are considered. To illustrate this point, we have carried out coalescent simulations, taking into account the individual best-fitting histories, and tabulated the average extent of linkage disequilibrium (LD) between markers separated by different values of recombination fraction (for a fixed value of per-nucleotide, per generation recombination rate, the recombination fraction translates into physical distance), as shown in Figure 6. Similar demographic histories distilled from the Asian and European samples result in similar values of LD at a given marker distance. LD is predicted to decay more rapidly (roughly twice as fast) for the best-fitting demographic history for the African-American samples, in agreement with previous reports (Reich *et al.* 2001). Differences in the extent of allelic association within the genome are expected to have profound consequences for medical association studies.

Caveats and open problems: Clearly, our multi-epoch, stepwise models of demographic history represent simplified versions of the "true" demographic past. Nevertheless, our three-epoch models go beyond the majority of previous studies that explore even simpler models of past population dynamics such as expansion *vs.* collapse or are restricted to the rejection of stationary effective size on the basis of summary statistics. Consideration of the third-order dynamics in this study allowed us to reveal a phase of bottleneck in the history characterizing the European and the Asian samples, permitting reconciliation of the signals of recent population growth apparent in mitochondrial and microsatellite data with realistic, observed values of nucleotide diversity.

Although the signal of differential history is undeniable in the data, the effect is confounded by the fact that the discovery and genotyping data sets were not drawn from a single population. SNP discovery was performed in shotgun sequences from ethnically diverse libraries (with ethnic association of individual reads unknown) aligned to the public genome reference sequence (Sachidanandam *et al.* 2001), presumably representing a mixture of ethnicities, with a bias toward clones from European donors (Weber *et al.* 2002). Polymorphic sites generated by this effort were then selected for genotyping in ethnically well-defined samples. It has

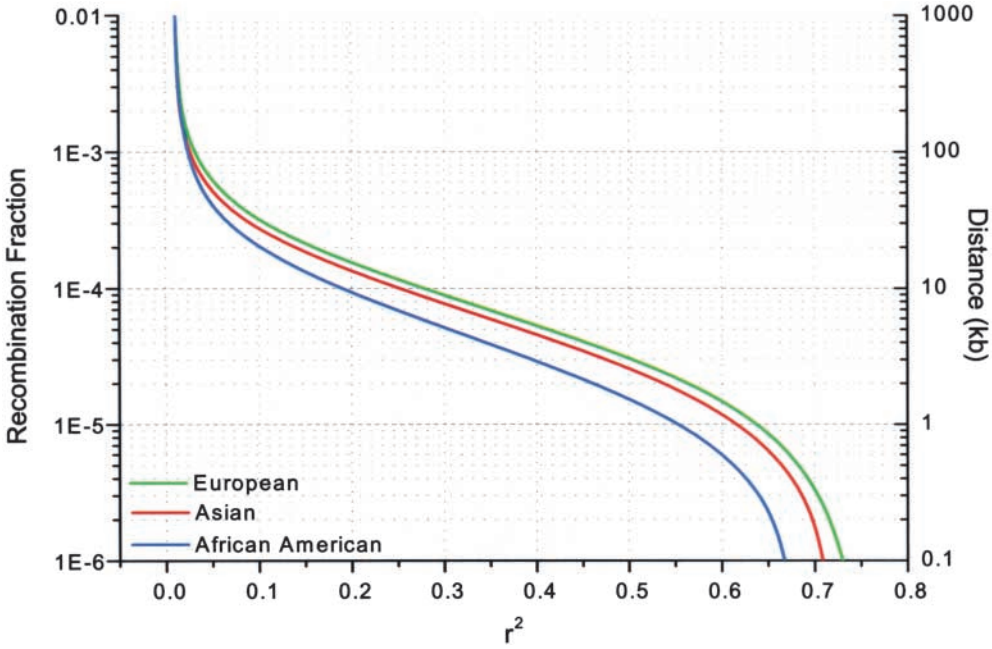

FIGURE 6.—The average extent of linkage disequilibrium, as predicted by the best-fitting, three-epoch demographic models for the three population samples. Values of r^2 and the corresponding values of recombination fraction are shown for each of the three populations. On the right-hand side, we have indicated the equivalent physical distances assuming a genome average per-nucleotide, per-generation recombination rate, $r = 10^{-8}$ (METHODS).

been previously noted that collections of samples from multiple ethnicities contain a surplus of rare SNPs when measured in the same mixed collection (PTAK and PRZEWORSKI 2002). However, it is unclear what the allele frequency of the same SNPs is when measured separately, within subpopulations. If the ethnicity of the discovery and the genotyping samples were known, one could estimate the effect of the ascertainment bias with models of population subdivision using coalescent simulation (PLUZHNIKOV *et al.* 2002). The effect of ascertainment bias between ethnically mismatched or undefined samples is the subject of future investigation.

Additionally, internal population substructure can also distort the frequency spectrum (PRZEWORSKI 2002; PTAK and PRZEWORSKI 2002). Unfortunately, the little amount of information that was available concerning sample origin did not permit incorporation of this effect into our models in a meaningful fashion. Specifically, we did not take into account in our models the effects of recent admixture in the African-American samples. Although the AFS in these samples are best modeled by population growth, it carries a slight but noticeable dip at medium minor allele frequencies, a feature present in a more pronounced form in both the European (Figure 4a) and the Asian (Figure 4b) spectra. This potentially signifies the contribution of European ancestral lineages on the background of African lineages (RYBICKI *et al.* 2002) in the AFS signal.

We must also acknowledge that the current shape of human variation structure is the result of a combination of neutral and nonneutral (selective) forces. The current state of the art in recognizing the effects of selection in variation data has been reviewed recently (BAMSHAD and WOODING 2003). Positive selection resulting in genetic hitchhiking can mimic the effects of population expansion in that it gives rise to an excess of low-frequency alleles (KAPLAN *et al.* 1989; BRAVERMAN *et al.* 1995). Recent efforts have been aimed at detecting loci that exhibit signatures of positive selection (CARGILL *et al.* 1999; SUNYAEV *et al.* 2000; AKEY *et al.* 2002; PAYSEUR *et al.* 2002). However, the exact proportion of genes that have been targets of strong positive selection within our evolutionary past is unclear (BAMSHAD and WOODING 2003). It is also unclear, in general, how far the effects of hitchhiking extend beyond the locus under selection (WIEHE 1998). Given that only a few percent of the human genome represents coding DNA, and that not all genes are expected to be targets of positive selection, we speculate that the distortion due to selective forces on the AFS in our data set of >20,000 randomly selected genomic loci is small when compared to the global effects of drift modulated by long-term demography.

Conclusion: The allele frequency spectrum is an excellent data source for modeling demographic history because of its independence of the effects of recombination and local, or sequence composition-specific variations of mutation rates and because the experimental determination of the allele frequency spectrum requires measurement of allelic states only at single-nucleotide positions, instead of sequencing of long stretches of contiguous DNA. The emergence of population-specific genotype sets on the genome scale provides sufficient data for the direct comparison of model-predicted and observed spectra with great resolution. This permits us to improve on previous conclusions drawn on the strength of summary statistics, on the basis of data from a handful of loci. Recent advances in allele frequency

modeling should provide us with exciting, new tools to explore our demographic past and explain human haplotype structure. Accurate reconstruction of the history of world populations should also help us to detect and interpret differences that must be taken into account during the development of general resources for medical use such as the recently initiated human Haplotype Map Project (Cardon and Abecasis 2003; Clark 2003; Wall and Pritchard 2003).

The authors are indebted to Andrew Clark for useful comments on the manuscript. We also thank Ravi Sachidanandam for kindly providing earlier versions of the allele frequency data set analyzed in this study.

LITERATURE CITED

Akey, J. M., G. Zhang, K. Zhang, L. Jin and M. D. Shriver, 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome Res. **12:** 1805–1814.

Altshuler, D., V. J. Pollara, C. R. Cowles, W. J. Van Etten, J. Baldwin *et al.*, 2000 An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature **407:** 513–516.

Bamshad, M., and S. P. Wooding, 2003 Signatures of natural selection in the human genome. Nat. Rev. Genet. **4:** 99–111.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley and W. Stephan, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783–796.

Cardon, L. R., and G. R. Abecasis, 2003 Using haplotype blocks to map human complex trait loci. Trends Genet. **19:** 135–140.

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.

Clark, A. G., 2003 Finding genes underlying risk of complex disease by linkage disequilibrium mapping. Curr. Opin. Genet. Dev. **13:** 296–302.

Clark, A. G., K. M. Weiss, D. A. Nickerson, S. L. Taylor, A. Buchanan *et al.*, 1998 Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am. J. Hum. Genet. **63:** 595–612.

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetic Theory.* Harper & Row, New York.

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.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

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.

Fu, Y. X., 1995 Statistical properties of segregating sites. Theor. Popul. Biol. **48:** 172–197.

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

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy *et al.*, 2002 The structure of haplotype blocks in the human genome. Science **296:** 2225–2229.

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. Tavare, 1994a Simulating probability distributions in the coalescent. Theor. Popul. Biol. **46:** 131–159.

Griffiths, R. C., and S. Tavare, 1994b Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond. B Biol. Sci. **344:** 403–410.

Harding, R. M., S. M. Fullerton, R. C. Griffiths, J. Bond, M. J. Cox *et al.*, 1997 Archaic African and Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. **60:** 772–789.

Harpending, H., and A. Rogers, 2000 Genetic perspectives on human origins and differentiation. Annu. Rev. Genomics Hum. Genet. **1:** 361–385.

Hey, J., 1997 Mitochondrial and nuclear genes present conflicting portraits of human origins. Mol. Biol. Evol. **14:** 166–172.

Hey, J., and E. Harris, 1999 Population bottlenecks and patterns of human polymorphism. Mol. Biol. Evol. **16:** 1423–1426.

Hudson, R. R., 1991 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology,* edited by D. Futuyama and J. Antonovics. Oxford University Press, London/New York/Oxford.

Ingman, M., H. Kaessmann, S. Paabo and U. Gyllensten, 2000 Mitochondrial genome variation and the origin of modern humans. Nature **408:** 708–713.

Jorde, L. B., W. S. Watkins and M. J. Bamshad, 2001 Population genomics: a bridge from evolutionary history to genetic medicine. Hum. Mol. Genet. **10:** 2199–2207.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

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.

Kondrashov, A. S., 2003 Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum. Mutat. **21:** 12–27.

Kruglyak, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. **22:** 139–144.

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

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860–921.

Li, W. H., 1977 Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. Genetics **85:** 331–337.

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:** 376–381.

Mullikin, J. C., S. E. Hunt, C. G. Cole, B. J. Mortimore, C. M. Rice *et al.*, 2000 An SNP map of human chromosome 22. Nature **407:** 516–520.

Nachman, M. W., and S. L. Crowell, 2000 Estimate of the mutation rate per nucleotide in humans. Genetics **156:** 297–304.

Ott, J., 1991 *Analysis of Human Genetic Linkage.* Johns Hopkins University Press, Baltimore.

Payseur, B. A., A. D. Cutter and M. W. Nachman, 2002 Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. Mol. Biol. Evol. **19:** 1143–1153.

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.

Przeworski, M., 2002 The signature of positive 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.

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.

Reich, D. E., and D. B. Goldstein, 1998 Genetic evidence for a Paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. USA **95:** 8119–8123.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411:** 199–204.

Reich, D. E., S. F. Schaffner, M. J. Daly, G. McVean, J. C. Mullikin *et al.*, 2002 Human genome sequence variation and the influence of gene history, mutation and recombination. Nat. Genet. **32:** 135–142.

Relethford, J. H., and L. B. Jorde, 1999 Genetic evidence for larger African population size during recent human evolution. Am. J. Phys. Anthropol. **108:** 251–260.

Rogers, A. R., 2001 Order emerging from chaos in human evolutionary genetics. Proc. Natl. Acad. Sci. USA **98:** 779–780.

Rogers, A. R., and H. Harpending, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552–569.

Rybicki, B. A., S. K. Iyengar, T. Harris, R. Liptak, R. C. Elston *et al.*, 2002 The distribution of long range admixture linkage disequilibrium in an African-American population. Hum. Hered. **53:** 187–196.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein *et al.*, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409:** 928–933.

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.

Sherry, S. T., H. C. Harpending, M. A. Batzer and M. Stoneking, 1997 Alu evolution in human populations: using the coalescent to estimate effective population size. Genetics **147:** 1977–1982.

Sunyaev, S. R., W. C. Lathe III, V. E. Ramensky and P. Bork, 2000 SNP frequencies in human genes an excess of rare alleles and differing modes of selection. Trends Genet. **16:** 335–337.

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

Tavare, S., D. J. Balding, R. C. Griffiths and P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics **145:** 505–518.

Tishkoff, S. A., and S. M. Williams, 2002 Genetic analysis of African populations: human evolution and complex disease. Nat. Rev. Genet. **3:** 611–621.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural *et al.*, 2001 The sequence of the human genome. Science **291:** 1304–1351.

Wall, J. D., and J. K. Pritchard, 2003 Haplotype blocks and linkage disequilibrium in the human genome. Nat. Rev. Genet. **4:** 587–597.

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

Weber, J. L., D. David, J. Heil, Y. Fan, C. Zhao *et al.*, 2002 Human diallelic insertion/deletion polymorphisms. Am. J. Hum. Genet. **71:** 854–862.

Wiehe, T., 1998 The effect of selective sweeps on the variance of the allele distribution of a linked multiallele locus: hitchhiking of microsatellites. Theor. Popul. Biol. **53:** 272–283.

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

Yu, N., Z. Zhao, Y. X. Fu, N. Sambuughin, M. Ramsay *et al.*, 2001 Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol. Biol. Evol. **18:** 214–222.

Zhao, Z., L. Jin, Y. X. Fu, M. Ramsay, T. Jenkins *et al.*, 2000 Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc. Natl. Acad. Sci. USA **97:** 11354–11358.

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: L. Excoffier

APPENDIX: THE EXPECTED NUMBER OF SEGREGATING SITES IN A SAMPLE DRAWN FROM A POPULATION CHARACTERIZED BY A PIECEWISE CONSTANT, MULTI-EPOCH HISTORY OF EFFECTIVE SIZE

Model: We consider a population of a given organism evolving under the Wright-Fisher model and under selective neutrality. Let us select a specific site in the genome of the organism. Furthermore, let us randomly draw n DNA samples from this population. Without regard to recombination, the samples possess a unique tree-shaped genealogy at the selected site (the site genealogy). Such a genealogy can be described within the framework of the coalescent: starting with n samples in the present and, through a series of coalescent events (pairs of samples finding their common ancestors), this number reduces to 1, the most recent common ancestor (MRCA), or the root of the genealogy at that site (site root). At a given time, the process is said to be in state j, if at that time the current number of samples is j. This process is Markovian, in that the length of time until the next coalescent event depends only on the current state and is independent of the previous states. Due to molecular mutation processes, the nucleotide observed at the site under consideration might be different in different individuals. Let us assume that, at any given site, only two possible nucleotides are observed (diallelic variations). Accordingly, an individual carries either the allele that was present in the site root (also known as the ancestral allele) or a mutant or derived allele. Let us further assume that the mutant allele is the result of a single mutation event (infinite-sites assumption) within an ancestral sample of the site genealogy. Under this assumption, the number of samples that carry the derived allele is identical to the number of descendants of that ancestor within the site genealogy. Conversely, the derived allele is found in exactly i samples if and only if the ancestor in which the mutation occurred gave rise to i descendants. Under the further assumption of a constant-rate mutation process (Hudson 1991), the likelihood that a given mutation is of size i is related to the number of ancestral nodes with i descendants within the site genealogy and to the "life span" of these ancestors. As Fu shows in a seminal work (Fu 1995), this likelihood can be expressed with the length of time the site genealogy spends in *state k*, *i.e.*, while the number of ancestor samples within the genealogy is exactly k. Under the further assumption of constant effective population size N, Fu then derives an explicit formula for the expected length of time in state k, leading to a simple result for the expected number of mutations of a given size within n samples (Fu 1995).

Our final goal is to extend this result from constant to merely piecewise constant population size. To this end, we use a standard continuous approximation according to which the probability density function of the length of time t spent in state k within the genealogy is exponential under a constant population size, and for a diploid population,

$$\frac{\binom{k}{2}}{2N} e^{-\left(\binom{k}{2}/2N\right)t}.$$

Using this approximation, we derive the expectation for the length of time spent in state k, under piecewise constant population history of an arbitrary number of epochs. Under the assumption of a constant-rate mutation process, this allows us to compute the expectation for the number of mutations of size i, denoted by Ψ_i, observed at a single site, at sites having identical site genealogies (DNA without recombination), or at a collection of sites with completely independent site genealogies. Because the distributions are identical for every site, the result is also valid for a collection of sites.

Conventions and useful identities: We use the convention that the value of an empty product is 1 and the value of an empty sum is 0. The probability density function of a random variable X is denoted by f_{X} and its cumulative density function by F_{X}. The variable X conditioned on the event Y is denoted by $\mathsf{X}|Y$. Next, we briefly state three lemmas to aid further derivations. In the following we assume that the a_i are different.

Lemma 1. *For every value of x, for each* $1 \le l \le n$,

$$\sum_{i=l}^{n} \prod_{\substack{m:m\neq i \\ l\le m\le n}} \frac{a_m - x}{a_m - a_i} = 1. \tag{A1}$$

Proof. Let

$$f(x) := -1 + \sum_{j=1}^{n} \left(\prod_{\substack{i:i\neq j, \\ l\le i\le n}} \frac{a_i - x}{a_i - a_j} \right);$$

we need to show that $f(x) \equiv 0$. For r: $l \le r \le n$ we have that

$$f(a_r) = -1 + \prod_{\substack{i:i\neq r; \\ l\le i\le n}} \frac{a_i - a_r}{a_i - a_r} = 0.$$

Since $f(x)$ is of degree at most $n - l$ and it has at least $n - l + 1$ different zeros, necessarily $f(x) \equiv 0$. Q.E.D.

Lemma 2. *For k, i:* $1 \le k < i \le n$ *we have*

$$\sum_{j=k}^{i} \left\{ \frac{a_i}{a_j} \left(\prod_{l:k<l\le j} \frac{a_l}{a_l - a_k} \right) \left(\prod_{m:j\le m<i} \frac{a_m}{a_m - a_i} \right) \right\} = 0. \tag{A2}$$

Proof.

$$\beta_{k,i} := \sum_{j=k}^{i} \left\{ \frac{a_i}{a_j} \left(\prod_{l:k<l\le j} \frac{a_l}{a_l - a_k} \right) \left(\prod_{m:j\le m<i} \frac{a_m}{a_m - a_i} \right) \right\}.$$

$\beta_{k,k+1} = 0$, and for $i > k + 1$

$$\beta_{k,i} = \sum_{j=k}^{i} \left[\frac{a_i}{a_j} \left(\prod_{l:k<l\le j} \frac{a_l}{a_l - a_k} \right) \left(\prod_{m:j\le m<i} \frac{a_m}{a_m - a_i} \right) \right] = \left(\prod_{l:k<l\le i} \frac{a_l}{a_l - a_k} \right) \alpha_{k,i},$$

where

$$\begin{aligned}
\alpha_{k,i} &= 1 + \sum_{j=k}^{i-1} \left\{ \frac{a_i}{a_j} \cdot \frac{a_j}{(a_j - a_i)} \cdot \frac{(a_i - a_k)}{a_i} \left(\prod_{m:j<m<i} \frac{a_m - a_k}{a_m} \right) \left(\prod_{m:j<m<i} \frac{a_m}{a_m - a_i} \right) \right\} \\
&= 1 + \sum_{j=k}^{i-1} \left\{ \frac{a_i - a_k}{a_j - a_i} \left(\prod_{m:j<m<i} \frac{a_m - a_k}{a_m - a_i} \right) \right\} \\
&= \frac{a_{i-1} - a_k}{a_{i-1} - a_i} + \sum_{j=k}^{i-2} \left\{ \left(-1 + \frac{a_j - a_k}{a_j - a_i} \right) \left(\prod_{m:j<m<i} \frac{a_m - a_k}{a_m - a_i} \right) \right\} \\
&= -\left(\sum_{j=k}^{i-2} \left\{ \prod_{m:j<m<i} \frac{a_m - a_k}{a_m - a_i} \right\} \right) + \left(\sum_{j=k+1}^{i-1} \left\{ \frac{a_j - a_k}{a_j - a_i} \left(\prod_{m:j<m<i} \frac{a_m - a_k}{a_m - a_i} \right) \right\} \right)
\end{aligned}$$

$$= -\left(\sum_{j=k}^{i-2}\left\{\prod_{m:j<m<i}\frac{a_m - a_k}{a_m - a_i}\right\}\right) + \left(\sum_{j=k+1}^{i-1}\left\{\prod_{m:j\le m<i}\frac{a_m - a_k}{a_m - a_i}\right\}\right) = 0. \qquad \text{Q.E.D.}$$

Lemma 3. *For* $s < k < i \le n$*:*

$$\sum_{j=k}^{i}\left(\frac{a_i}{a_j}\left\{\prod_{\substack{l:l\ne k;\\ s+1\le l\le j}}\frac{a_l}{a_l - a_k}\right\}\left\{\prod_{\substack{m:m\ne i;\\ j\le m\le n}}\frac{a_m}{a_m - a_i}\right\}\right) = 0. \tag{A3}$$

Proof. From Lemma 2,

$$0 = \left(\prod_{q:s+1\le q<k}\frac{a_q}{a_q - a_k}\right)\left(\prod_{r:i<r\le n}\frac{a_r}{a_r - a_i}\right)\left(\sum_{j=k}^{i}\frac{a_i}{a_j}\prod_{l:k<l\le j}\frac{a_l}{a_l - a_k}\prod_{m:j\le m<i}\frac{a_m}{a_m - a_i}\right)$$

$$= \sum_{j=k}^{i}\left(\frac{a_i}{a_j}\left\{\prod_{\substack{l:l\ne k;\\ s+1\le l\le j}}\frac{a_l}{a_l - a_k}\right\}\left\{\prod_{\substack{m:m\ne i;\\ j\le m\le n}}\frac{a_m}{a_m - a_i}\right\}\right). \qquad \text{Q.E.D.}$$

Lemma 4.

$$\frac{1}{a_s} = \sum_{j=s}^{n}\left\{\frac{1}{a_j}\left(\prod_{\substack{l:i\ne j;\\ s\le i\le n}}\frac{a_i}{a_i - a_j}\right)\right\} - \sum_{j=s+1}^{n}\left\{\frac{1}{a_j}\left(\prod_{\substack{i:i\ne j;\\ s+1\le i\le n}}\frac{a_i}{a_i - a_j}\right)\right\}.$$

Proof. Using Lemma 1,

$$\frac{1}{a_s} = \frac{1}{a_s}\sum_{j=s}^{n}\left(\prod_{\substack{i:i\ne j;\\ s\le i\le n}}\frac{a_i}{a_i - a_j}\right) = \frac{1}{a_s}\left(\prod_{i:s+1\le i\le n}\frac{a_i}{a_i - a_j}\right) + \sum_{j=s+1}^{n}\left\{\frac{1}{a_j}\left(1 - \frac{a_s - a_j}{a_s}\right)\left(\prod_{\substack{i:i\ne j;\\ s\le i\le n}}\frac{a_i}{a_i - a_j}\right)\right\}$$

$$= \sum_{j=s}^{n}\left\{\frac{1}{a_j}\left(\prod_{\substack{i:i\ne j;\\ s\le i\le n}}\frac{a_i}{a_i - a_j}\right)\right\} - \sum_{j=s+1}^{n}\left\{\frac{1}{a_j}\left(\prod_{\substack{i:i\ne j;\\ s+1\le i\le n}}\frac{a_i}{a_i - a_j}\right)\right\}. \qquad \text{Q.E.D.}$$

Constant effective population size: First, we consider a demographic history characterized by a single, constant population size N_1. We introduce the notations $a_j = \binom{j}{2}$ and $a_j^{(1)} = a_j/2N_1$. The length of time spent in state j (after which the number of samples reduces from j to $j - 1$) is denoted by $\mathsf{T}_{j,j-1}$. The random variables $\mathsf{T}_{j,j-1}$ and $\mathsf{T}_{i,i-1}$ are independent for $i \ne j$. The density function of $\mathsf{T}_{j,j-1}$ is $f_{\mathsf{T}_{j,j-1}}(t) = a_j^{(1)}e^{-a_j^{(1)}t}$, according to our model assumptions. The length of time from the present, when the number of samples is n, to the instant when the number of samples reduces to s, is denoted by $\mathsf{T}_{n,s}^{(1)}$. Clearly $\mathsf{T}_{n,s}^{(1)} = \Sigma_{j=s+1}^{n}\mathsf{T}_{j,j-1}$. The probability that, at time t, the genealogy is in state s is $P(\mathsf{T}_{n,s}^{(1)} \le t < \mathsf{T}_{n,s-1}^{(1)})$. Since $\mathsf{T}_{n,l}^{(1)} = \mathsf{T}_{n,l+1}^{(1)} + \mathsf{T}_{l+1,l}$, for l: $1 \le l < n$ we can use the following convolution: $f_{\mathsf{T}_{n,l}^{(1)}}(t) = \int_0^t f_{\mathsf{T}_{n,l+1}^{(1)}}(t - x)f_{\mathsf{T}_{l+1,l}}(x)\,dx$. Using these notations, the following are true:

Theorem 1. *For* s: $1 \le s < n$*:*

$$f_{\mathsf{T}_{n,s}^{(1)}}(t) = \sum_{j=s+1}^{n}\left\{a_j^{(1)}e^{-a_j^{(1)}t}\left(\prod_{\substack{i:i\ne j;\\ s+1\le i\le n}}\frac{a_i}{a_i - a_j}\right)\right\}, \tag{A4}$$

$$F_{\mathsf{T}_{n,s}^{(1)}}(t) = 1 - \sum_{j=s+1}^{n}\left\{e^{-a_j^{(1)}t}\left(\prod_{\substack{i:i\ne j;\\ s+1\le i\le n}}\frac{a_i}{a_i - a_j}\right)\right\}, \tag{A5}$$

$$E\left(\mathsf{T}_{n,s}^{(1)}\right) = \sum_{j=s+1}^{n}\left\{\frac{1}{a_j^{(1)}}\left(\prod_{\substack{i:i\ne j;\\ s+1\le i\le n}}\frac{a_i}{a_i - a_j}\right)\right\} = 2N_1\sum_{j=s+1}^{n}\left\{\frac{1}{a_j}\left(\prod_{\substack{i:i\ne j;\\ s+1\le i\le n}}\frac{a_i}{a_i - a_j}\right)\right\}. \tag{A6}$$

For s: $2 \le s < n$:

$$P\left(\mathsf{T}^{[1]}_{n,s}\right) \le t < \mathsf{T}^{[1]}_{n,s-1}\right) = \sum_{j=s}^{n}\left\{\frac{a_j}{a_s}e^{-a_j^{(1)}t}\left(\prod_{\substack{i:i\neq j;\\ s\le i\le n}}\frac{a_i}{a_i - a_j}\right)\right\} = \frac{f_{T^{[1]}_{n,s-1}}(t)}{a_s^{(1)}}, \tag{A7}$$

$$E\left(\mathsf{T}_{s,s-1}\right) = \frac{1}{a_s^{(1)}}. \tag{A8}$$

For i: $1 \le i < n$:

$$E(\Psi_i) = \frac{4N_1\mu}{i}. \tag{A9}$$

Proof. First we show Equations A4 and A5 by downward induction on *s*. These equations are clearly valid for $s = n - 1$. Assume they are valid for $s: s > k$. Then

$$\begin{aligned}
f_{\mathsf{T}^{[1]}_{n,k}}(t) &= \int_0^t f_{\mathsf{T}^{[1]}_{n,k+1}}(t-x) f_{\mathsf{T}^{[1]}_{k+1,k}}(x)\,dx \\
&= \sum_{j=k+2}^{n}\left(a_{k+1}^{(1)}a_j^{(1)}e^{-a_j^{(1)}t}\prod_{\substack{i:i\neq j;\\ k+2\le i\le n}}\frac{a_i}{(a_i-a_j)}\int_0^t e^{\left(a_j^{(1)}-a_{k+1}^{(1)}\right)x}dx\right) \\
&= \sum_{j=k+2}^{n}\left(a_j^{(1)}e^{-a_j^{(1)}t}\prod_{\substack{i:i\neq j;\\ k+1\le i\le n}}\frac{a_i}{(a_i-a_j)}\left[1 - e^{\left(a_j^{(1)}-a_{k+1}^{(1)}\right)t}\right]\right) \\
&= \left(\sum_{j=k+2}^{n}\left(a_j^{(1)}e^{-a_j^{(1)}t}\prod_{\substack{i:i\neq j;\\ k+1\le i\le n}}\frac{a_i}{a_i-a_j}\right)\right) - e^{-a_{k+1}^{(1)}t}\sum_{j=k+2}^{n}\left\{a_j^{(1)}\left(\prod_{\substack{i:i\neq j;\\ k+1\le i\le n}}\frac{a_i}{a_i-a_j}\right)\right\}.
\end{aligned}$$

For Equation A4 we need to show that

$$-\sum_{j=k+2}^{n}a_j\left(\prod_{\substack{i:i\neq j;\\ k+1\le i\le n}}\frac{a_i}{a_i-a_j}\right) = a_{k+1}\left(\prod_{k+2\le i\le n}\frac{a_i}{a_i-a_{k+1}}\right).$$

This is equivalent to

$$1 = -\frac{\left(\sum_{j=k+2}^{n}a_j\prod_{\substack{i:i\neq j;\\ k+1\le i\le n}}a_i/(a_i-a_j)\right)}{a_{k+1}\prod_{k+2\le i\le n}a_i/(a_i-a_{k+1})} = \sum_{j=k+2}^{n}\left[\prod_{\substack{\iota:\iota\neq\varphi;\\ \kappa+2\le\iota\le\nu}}\left(\frac{\alpha_\iota-\alpha_{\kappa+1}}{\alpha_\iota-\alpha_\varphi}\right)\right],$$

which follows from Lemma 1. Using Lemma 1 with $l = s + 1$ and $x = 0$, we get

$$\begin{aligned}
F_{\mathsf{T}^{[1]}_{n,s}}(t) &= P(\mathsf{T}^{[1]}_{n,s} \le t) = \int_0^t f_{\mathsf{T}^{[1]}_{n,s}}(x)\,dx = \sum_{j=s+1}^{n}\left[\left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}}\frac{a_i}{a_i-a_j}\right)\int_0^t a_j^{(1)}e^{-a_j^{(1)}x}dx\right] \\
&= \sum_{j=s+1}^{n}\left[\left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}}\frac{a_i}{a_i-a_j}\right)\left(1 - e^{-a_j^{(1)}t}\right)\right] \\
&= \left(\sum_{j=s+1}^{n}\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}}\frac{a_i}{a_i-a_j}\right) - \left(\sum_{j=s+1}^{n}e^{-a_j^{(1)}t}\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}}\frac{a_i}{a_i-a_j}\right) \\
&= 1 - \sum_{j=s+1}^{n}e^{-a_j^{(1)}t}\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}}\frac{a_i}{(a_i-a_j)}.
\end{aligned}$$

This completes the proof of Equations A4 and A5. For (A7), note that $P\left(\mathsf{T}^{\{1\}}_{n,s} > t\right) = 1 - F_{\mathsf{T}^{\{1\}}_{n,s}}(t)$ and $P\left(\mathsf{T}^{\{1\}}_{n,s} \le t < \mathsf{T}^{\{1\}}_{n,s-1}\right) = P\left(\mathsf{T}^{\{1\}}_{n,s-1} > t\right) - P\left(\mathsf{T}^{\{1\}}_{n,s} > t\right)$. Then

$$\begin{aligned}
P\left(\mathsf{T}^{\{1\}}_{n,s} \le t < \mathsf{T}^{\{1\}}_{n,s-1}\right) &= \sum_{j=s}^{n}\left\{ e^{-a_j^{(1)}t}\left(\prod_{\substack{i:i\neq j;\\ s\le i\le n}} \frac{a_i}{a_i - a_j}\right)\right\} - \sum_{j=s+1}^{n}\left\{ e^{-a_j^{(1)}t}\left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j}\right)\right\} \\
&= e^{-a_s^{(1)}t}\left(\prod_{s+1\le i\le n} \frac{a_i}{a_i - a_j}\right) + \sum_{j=s+1}^{n}\left\{\left(1 - \frac{a_s - a_j}{a_s}\right)e^{-a_j^{(1)}t}\left(\prod_{\substack{i:i\neq j;\\ s\le i\le n}} \frac{a_i}{a_i - a_j}\right)\right\} \\
&= \frac{a_s e^{-a_s^{(1)}t}}{a_s}\left(\prod_{\substack{i:i\neq s;\\ s\le i\le n}} \frac{a_i}{(a_i - a_{s-1})}\right) + \sum_{j=s+1}^{n}\left\{\frac{a_j e^{-a_j^{(1)}t}}{a_s}\left(\prod_{\substack{i:i\neq j;\\ s\le i\le n}} \frac{a_i}{a_i - a_j}\right)\right\} \\
&= \sum_{j=s}^{n}\left\{\frac{a_j}{a_s} e^{-a_j t}\left(\prod_{\substack{i:i\neq j;\\ s\le i\le n}} \frac{a_i}{(a_i - a_j)}\right)\right\} = \frac{f_{T^{\{1\}}_{n,s-1}}(t)}{a_s^{(1)}}.
\end{aligned}$$

For (A6), since $\mathsf{T}^{\{1\}}_{n,s} \ge 0$,

$$\begin{aligned}
E\left(\mathsf{T}^{\{1\}}_{n,s}\right) &= \int_0^\infty P(\mathsf{T}^{\{1\}}_{n,s} \ge x)dx = \int_0^\infty \sum_{j=s+1}^{n}\left\{ e^{-a_j^{(1)}x}\left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j}\right)\right\}dx \\
&= \sum_{j=s+1}^{n}\left(\frac{1}{a_j^{(1)}}\left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j}\right)\int_0^\infty a_j^{(1)} e^{-a_j^{(1)}x}dx\right) = \sum_{j=s+1}^{n}\left(\frac{1}{a_j^{(1)}}\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j}\right).
\end{aligned}$$

Equation A8 can be easily obtained from $f_{s,s-1}(t)$. Finally, Equation A9 follows from Equation A8, by the argument presented by Fu (1995) to derive Equation 22. Q.E.D.

Piecewise constant effective population size: Consider a demographic history of M distinct epochs indexed by 1, 2, . . . , M, where the ancestral epoch is numbered M. For epoch i, the constant effective population size is N_i, and the duration of this epoch is T_i; in particular, $T_M = \infty$. We define $a_k^{(i)} = \binom{k}{2}/2N_i$. We introduce $\tau_i = \Sigma_{j=1}^{i} T_j$, the time from the present back until the end of the ith epoch (so $\tau_0 = 0$ and $\tau_M = \infty$). At a given time t, the index of the current epoch is denoted by $m(t)$, in formula $m(t) = \min\{k: \tau_k \ge t\}$. In particular, $m(\tau_i) = i$, and $\tau_{m(t)-1} < t \le \tau_{m(t)}$. We also introduce a "normalized" time t^*:

$$t^* = \frac{t - \tau_{m(t)-1}}{2N_{m(t)}} + \sum_{i=1}^{m(t)-1}\frac{T_i}{2N_i}.$$

The proof is based on induction on the number of epochs. To facilitate this, we consider two kinds of partial models with smaller numbers of epochs, as follows:

1. The first model has a single epoch, with effective population size N_i. The random variable $T^{\{i\}}_{n,j}$ denotes the time from the present (state n) to the beginning of state j, under the parameters of the first model.
2. The second model is a truncated version of the original M-epoch model: it consists of i epochs, with parameters that are identical to the parameters of the first i epochs of the original model, except $T_i = \infty$; *i.e.*, the ith of the original model becomes the ancestral epoch of the truncated model. The random variable $T^{[i]}_{n,j}$ denotes the time from the present (state n) to reach state j, under the parameters of the second model.

Note that the two types of models coincide when $i = 1$. The following are true:

Theorem 2. *For s: $1 \le s < n$:*

$$f_{\mathsf{T}^{[M]}_{n,s}}(t) = f_{\mathsf{T}^{[m(t)]}_{n,s}}(t) \quad \textit{and} \quad F_{\mathsf{T}^{[M]}_{n,s}}(t) = F_{\mathsf{T}^{[m(t)]}_{n,s}}(t), \tag{A10}$$

$$f_{\mathsf{T}^{[M]}_{n,s}}(t) = \frac{1}{2N_{m(t)}}\sum_{j=s+1}^{n}\left\{a_j e^{-a_j t^*}\left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j}\right)\right\}, \tag{A11}$$

$$F_{\mathsf{T}_{n,s}^{[M]}}(t) = 1 - \sum_{j=s+1}^{n} \left\{ e^{-a_j t^*} \left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j} \right) \right\}, \tag{A12}$$

$$\begin{aligned} E\left(\mathsf{T}_{\mathsf{T}_{n,s}^{[M]}}\right) &= \sum_{j=s+1}^{n} \left\{ \frac{1}{a_j^{(1)}} \left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j} \right) \right\} + \sum_{m=1}^{M-1} \sum_{j=s+1}^{n} \left\{ e^{-\sum_{l=1}^{m} a_j^{(l)} T_l} \left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j} \right) \left(\frac{1}{a_j^{(m+1)}} - \frac{1}{a_j^{(m)}} \right) \right\} \\ &= 2N_1 \sum_{j=s+1}^{n} \left\{ \frac{1}{a_j} \left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j} \right) \right\} \\ &\quad + \sum_{m=1}^{M-1} \left[2(N_{m+1} - N_m) \sum_{j=s+1}^{n} \left\{ e^{-a_j \tau_m^*} \frac{1}{a_j} \left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j} \right) \right\} \right]. \end{aligned} \tag{A13}$$

For s: $2 \le s < n$:

$$P\left(\mathsf{T}_{n,s}^{[M]} \le t < \mathsf{T}_{n,s-1}^{[M]}\right) = \frac{f_{\mathsf{T}_{n,s-1}^{[M]}}(t)}{a_s^{(m(t))}} = \frac{f_{\mathsf{T}_{n,s-1}^{[m(t)]}}(t)}{a_s^{(m(t))}}, \tag{A14}$$

$$\begin{aligned} E\left(\mathsf{T}_{s,s-1}\right) &= \frac{1}{a_s^{(1)}} + \sum_{m=1}^{M-1} \sum_{j=s}^{n} \left[\left(e^{-\sum_{l=1}^{m} a_j^{(l)} T_l} \prod_{\substack{i:i\neq j;\\ s\le i\le n}} \frac{a_i}{a_i - a_j} \right) \left(\frac{1}{a_s^{(m+1)}} - \frac{1}{a_s^{(m)}} \right) \right] \\ &= \frac{2}{a_s} \left\{ N_1 + \sum_{m=1}^{M-1} \left[(N_{m+1} - N_m) \sum_{j=s}^{n} \left(e^{-a_j \tau_m^*} \prod_{\substack{i:i\neq j;\\ s\le i\le n}} \frac{a_i}{a_i - a_j} \right) \right] \right\}. \end{aligned} \tag{A15}$$

For i: $1 \le i < n$:

$$E(\Psi_i) = 4\mu \left[\frac{N_1}{i} + \sum_{m=1}^{M-1} \left\{ \frac{N_{m+1} - N_m}{i\binom{n-1}{i}} \sum_{k=2}^{n} \left[\binom{n-k}{i-1} \sum_{j=k}^{n} \left(e^{(j(j-1)\tau_m^*)/2} \prod_{\substack{l:l\neq j;\\ k\le l\le n}} \frac{l(l-1)}{l(l-1) - j(j-1)} \right) \right] \right\} \right]. \tag{A16}$$

Proof: (A12) and (A14) are consequences of (A11):

$$\begin{aligned} F_{\mathsf{T}_{n,s}^{[M]}}(t) &= 1 - \int_t^{\infty} f_{\mathsf{T}_{n,s}^{[M]}}(t)\, dt = 1 - \sum_{j=s+1}^{n} \left\{ e^{-a_j^{(M)}(-\tau_{M-1}) - \sum_{l=1}^{M-1} a_j^{(l)} T_l} \left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j} \right) \int_t^{\infty} a_j^{(M)} e^{-a_j^{(M)} t} dt \right\} \\ &= 1 - \sum_{j=s+1}^{n} \left\{ e^{-a_j^{(M)}(t-\tau_{M-1}) - \sum_{l=1}^{M-1} a_j^{(l)} T_l} \left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j} \right) \right\}. \end{aligned}$$

$$\begin{aligned} P\left(\mathsf{T}_{n,s}^{[M]} \le t < \mathsf{T}_{n,s-1}^{[M]}\right) &= F_{\mathsf{T}_{n,s}^{[M]}}(t) - F_{\mathsf{T}_{n,s-1}^{[M]}}(t) = \left(\sum_{j=s}^{n} \left\{ e^{-a_j^{(M)}(t-\tau_{M-1}) - \sum_{l=1}^{M-1} a_j^{(l)} T_l} \left(\prod_{\substack{i:i\neq j;\\ s\le i\le n}} \frac{a_i}{a_i - a_j} \right) \right\} \right) \\ &\quad - \left(\sum_{j=s+1}^{n} \left\{ e^{-a_j^{(M)}(t-\tau_{M-1}) - \sum_{l=1}^{M-1} a_j^{(l)} T_l} \left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j} \right) \right\} \right) \\ &= e^{-a_s^{(M)}(t-\tau_{M-1}) - \sum_{l=1}^{M-1} a_s^{(l)} T_l} \left(\prod_{\substack{i:i\neq j;\\ s\le i\le n}} \frac{a_i}{a_i - a_j} \right) \\ &\quad + \sum_{j=s+1}^{n} \left\{ \left(1 - \frac{a_s - a_j}{a_s} \right) e^{-a_j^{(M)}(t-\tau_{M-1}) - \sum_{l=1}^{M-1} a_j^{(l)} T_l} \left(\prod_{\substack{i:i\neq j;\\ s\le i\le n}} \frac{a_i}{a_i - a_j} \right) \right\} \end{aligned}$$

$$= \sum_{j=s}^{n} \left\{ \frac{a_j}{a_s} e^{-a_j^{(M)}(t-\tau_{M-1}) - \sum_{l=1}^{M-1} a_j^{(l)} T_l} \left(\prod_{\substack{i:i\neq j;\\ s\le i\le n}} \frac{a_i}{a_i - a_j} \right) \right\} = \frac{f_{T_{n,s-1}^{[M]}}(T)}{a_s^{(M)}} = \frac{f_{T_{n,s-1}^{[m(t)]}}(T)}{a_s^{(m(t))}}.$$

We prove (A10) and (A11) by induction on the number of epochs M. The statements are true for $M = 1$ by Theorem 1. For $M > 1$ assume that the statements are true if the number of epochs is less than M. Clearly,

$$\left\{\mathsf{T}_{n,j}^{[M]} = t\right\} = \left\{\mathsf{T}_{n,j}^{[M]} \le \tau_{M-1} \text{ and } \mathsf{T}_{n,j}^{[M]} = t\right\} \cup \left\{ \bigcup_{i=j+1}^{n} \left\{\mathsf{T}_{n,i}^{[M]} \le \tau_{M-1} < \mathsf{T}_{n,i-1}^{[M]} \text{ and } \mathsf{T}_{n,j}^{[M]} = t\right\} \right\}.$$

The right side is a union of disjoint events; therefore (using density functions of conditioned variables) we have

$$\begin{aligned} f_{\mathsf{T}_{n,j}^{[M]}}(t) &= P(\mathsf{T}_{n,j}^{[M]} \le \tau_{M-1}) f_{\mathsf{T}_{n,j}^{[M]} \mid \mathsf{T}_{n,j}^{[M]} \le \tau_{M-1}}(t) \\ &\quad + \sum_{i=j+1}^{n} P(\mathsf{T}_{n,i}^{[M]} \le \tau_{M-1} < \mathsf{T}_{n,i-1}^{[M]}) f_{\mathsf{T}_{n,j}^{[M]} \mid \mathsf{T}_{n,i}^{[M]} \le \tau_{M-1} < \mathsf{T}_{n,i-1}^{[M]}}(t). \end{aligned}$$

Clearly

$$f_{\mathsf{T}_{n,j}^{[M]} \mid \mathsf{T}_{n,j}^{[M]} \le \tau_{M-1}}(t) = \begin{cases} f_{\mathsf{T}_{n,j}^{[M-1]}}(t) / P(\mathsf{T}_{n,j}^{[M]} \le \tau_{M-1}), & t \le \tau_{M-1} \\ 0, & t > \tau_{M-1} \end{cases}$$

and for each $i > j$

$$f_{\mathsf{T}_{n,j}^{[M]} \mid \mathsf{T}_{n,i}^{[M]} \le \tau_{M-1} < \mathsf{T}_{n,i+1}^{[M]}}(t) = \begin{cases} 0, & t \le \tau_{M-1} \\ f_{i,j}^{[M]}(t - \tau_{M-1}), & t > \tau_{M-1} \end{cases}.$$

Therefore for $t \le \tau_{M-1}$ we have $f_{\mathsf{T}_{n,j}^{[M]}}(t) = f_{\mathsf{T}_{n,j}^{[M-1]}}(t)$ and $F_{\mathsf{T}_{n,j}^{[M]}}(t) = F_{\mathsf{T}_{n,j}^{[M-1]}}(t)$, so using the induction hypothesis, for $t \le \tau_{M-1}$, Equations A10, and consequently A11, hold. In particular,

$$P\left(\mathsf{T}_{n,s}^{[M]} \le \tau_{M-1} < \mathsf{T}_{n,s-1}^{[M]}\right) = \frac{f_{T_{n,s-1}^{[M-1]}}(\tau_{M-1})}{a_s^{(M-1)}} = \sum_{j=s}^{n} \left\{ \frac{a_j}{a_s} e^{-\sum_{l=1}^{M-1} a_j^{(l)} T_l} \left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j} \right) \right\}.$$

If $t > \tau_{M-1}$, *i.e.*, $m(t) = M$, then (A10) and (A11) follow from Lemma 3:

$$\begin{aligned} f_{\mathsf{T}_{n,i}^{[M]}}(t) &= \sum_{s=i+1}^{n} \left[\sum_{m=s}^{n} \left\{ \frac{a_m}{a_s} e^{-\sum_{l=1}^{M-1} a_m^{(l)} T_l} \prod_{\substack{p:p\neq m;\\ s\le p\le n}} \frac{a_p}{a_p - a_m} \right\} \sum_{j=i+1}^{s} \left\{ a_j^{(M)} e^{-a_j^{(M)}(t-\tau_{M-1})} \left(\prod_{\substack{q:q\neq j;\\ i+1\le q\le s}} \frac{a_q}{a_q - a_j} \right) \right\} \right] \\ &= \sum_{j=i+1}^{n} \left[a_j^{(M)} e^{-a_j^{(M)}(t-\tau_{M-1})} \sum_{m=j}^{n} \left\{ e^{-\sum_{l=1}^{M-1} a_m^{(l)} T_l} \sum_{s=j}^{m} \left(\frac{a_m}{a_s} \left(\prod_{\substack{q:q\neq j;\\ i+1\le q\le s}} \frac{a_q}{a_q - a_j} \right) \left(\prod_{\substack{p:p\neq m;\\ s\le p\le n}} \frac{a_p}{a_p - a_m} \right) \right) \right\} \right] \\ &= \sum_{j=i+1}^{n} \left[a_j^{(M)} e^{-a_j^{(M)}(t-\tau_{M-1}) - \sum_{l=1}^{M-1} a_j^{(l)} T_l} \left(\prod_{\substack{q:q\neq j;\\ i+1\le q\le n}} \frac{a_q}{a_q - a_j} \right) \right]. \end{aligned}$$

We get Equation A13 in a way similar to the proof of Equation A8:

$$\begin{aligned} E\left(\mathsf{T}_{n,s}^{[M]}\right) &= \int_0^\infty \left(1 - F_{\mathsf{T}_{n,s}^{[M]}}(t)\right) dt = \int_0^\infty \sum_{j=s+1}^{n} \left\{ e^{-a_j^{(m(t))}(t-\tau_{m(t)-1}) - \sum_{l=1}^{m(t)-1} a_j^{(l)} T_l} \left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j} \right) \right\} dt \\ &= \sum_{m=1}^{M} \int_0^{T_m} \sum_{j=s+1}^{n} \left\{ e^{-a_j^{(m)} t - \sum_{l=1}^{m-1} a_j^{(l)} T_l} \left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j} \right) \right\} dt = \sum_{j=s+1}^{n} \left\{ \frac{1}{a_j^{(M)}} e^{-\sum_{l=1}^{M-1} a_j^{(l)} T_l} \left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j} \right) \right\} \\ &\quad + \sum_{m=1}^{M-1} \sum_{j=s+1}^{n} \left\{ \frac{1}{a_j^{(m)}} \left(1 - e^{-a_j^{(m)} T_m}\right) e^{-\sum_{l=1}^{m-1} a_j^{(l)} T_l} \left(\prod_{\substack{i:i\neq j;\\ s+1\le i\le n}} \frac{a_i}{a_i - a_j} \right) \right\} \end{aligned}$$

$$= \sum_{m=1}^{M}\sum_{j=s+1}^{n}\left\{\frac{1}{a_j^{(m)}}e^{-\Sigma_{l=1}^{m-1}a_j^{(l)}T_l}\left(\prod_{\substack{i:\,i\neq j;\\ s+1\le i\le n}}\frac{a_i}{a_i-a_j}\right)\right\} - \sum_{m=1}^{M-1}\sum_{j=s+1}^{n}\left\{\frac{1}{a_j^{(m)}}e^{-\Sigma_{l=1}^{m}a_j^{(l)}T_l}\left(\prod_{\substack{i:\,i\neq j;\\ s+1\le i\le n}}\frac{a_i}{a_i-a_j}\right)\right\}$$

$$= \sum_{j=s+1}^{n}\left\{\frac{1}{a_j^{(1)}}\left(\prod_{\substack{i:\,i\neq j;\\ s+1\le i\le n}}\frac{a_i}{a_i-a_j}\right)\right\} + \sum_{m=1}^{M-1}\sum_{j=s+1}^{n}\left\{e^{-\Sigma_{l=1}^{m}a_j^{(l)}T_l}\left(\prod_{\substack{i:\,i\neq j;\\ s+1\le i\le n}}\frac{a_i}{a_i-a_j}\right)\left(\frac{1}{a_j^{(m+1)}}-\frac{1}{a_j^{(m)}}\right)\right\}.$$

Using Lemma 4,

$$E(\mathsf{T}_{s,s-1}) = E(\mathsf{T}_{n,s-1}^{[M]}) - E(\mathsf{T}_{n,s}^{[M]}) = \frac{1}{a_s^{(1)}} + \sum_{m=1}^{M-1}\sum_{j=s}^{n}\left\{e^{-\Sigma_{l=1}^{m}a_j^{(l)}T_l}\left(\prod_{\substack{i:\,i\neq j;\\ s\le i\le n}}\frac{a_i}{a_i-a_j}\right)\left(\frac{1}{a_j^{(m+1)}}-\frac{1}{a_j^{(m)}}\right)\right\}$$

$$- \sum_{m=1}^{M-1}\sum_{j=s+1}^{n}\left\{e^{-\Sigma_{l=1}^{m}a_j^{(l)}T_l}\left(\prod_{\substack{i:\,i\neq j;\\ s+1\le i\le n}}\frac{a_i}{a_i-a_j}\right)\left(\frac{1}{a_i^{(m+1)}}-\frac{1}{a_j^{(m)}}\right)\right\}$$

$$= \frac{1}{a_s^{(1)}} + \sum_{m=1}^{M-1}\left\{e^{-\Sigma_{l=1}^{m}a_s^{(l)}T_l}\left(\prod_{i=s+1}^{n}\frac{a_i}{a_i-a_s}\right)\left(\frac{1}{a_j^{(m+1)}}-\frac{1}{a_s^{(m)}}\right)\right\}$$

$$+ \sum_{m=1}^{M-1}\sum_{j=s+1}^{n}\left[e^{-\Sigma_{l=1}^{m}a_j^{(l)}T_l}\left(\prod_{\substack{i:\,i\neq j;\\ s\le i\le n}}\frac{a_i}{a_i-a_j}\right)\left(\frac{1}{a_j^{(m+1)}}-\frac{1}{a_j^{(m)}}\right)\left(1-\frac{a_s-a_j}{a_s}\right)\right]$$

$$= \frac{1}{a_s^{(1)}} + \sum_{m=1}^{M-1}\sum_{j=s}^{n}\left[e^{-\Sigma_{l=1}^{m}a_j^{(l)}T_l}\left(\prod_{\substack{i:\,i\neq j;\\ s\le i\le n}}\frac{a_i}{a_i-a_j}\right)\left(\frac{1}{a_j^{(m+1)}}-\frac{1}{a_j^{(m)}}\right)\right].$$

This gives Equation A15. Finally, using manipulations identical to those used by Fu (1995) we derive Equation A16:

$$\frac{E(\Psi_i)}{4\mu} = \sum_{k=2}^{n}\left[\frac{\binom{n-k}{i-1}}{i\binom{n-1}{i}}\left\{N_1 + \sum_{m=1}^{M-1}\left[(N_{m+1}-N_m)\sum_{j=k}^{n}\left(e^{-a_j\tau_m^*}\prod_{\substack{s:\,s\neq j;\\ k\le s\le n}}\frac{a_s}{a_s-a_j}\right)\right]\right\}\right]$$

$$= \frac{N_1}{i\binom{n-1}{i}}\left[\sum_{k=2}^{n}\binom{n-k}{i-1}\right] + \sum_{m=1}^{M-1}\left\{\frac{N_{m+1}-N_m}{i\binom{n-1}{i}}\sum_{k=2}^{n}\left[\binom{n-k}{i-1}\sum_{j=k}^{n}\left(e^{-a_j\tau_m^*}\prod_{\substack{s:\,s\neq j;\\ k\le s\le n}}\frac{a_s}{a_s-a_j}\right)\right]\right\}$$

$$= \frac{N_1}{i} + \sum_{m=1}^{M-1}\left\{\frac{N_{m+1}-N_m}{i}\binom{n-1}{i}^{-1}\sum_{k=2}^{n}\left[\binom{n-k}{i-1}\sum_{j=k}^{n}\left(e^{-a_j\tau_m^*}\prod_{\substack{s:\,s\neq j;\\ k\le s\le n}}\frac{a_s}{a_s-a_j}\right)\right]\right\},$$

where $\tau_m^* = \Sigma_{l=1}^{m}(T_l/2N_l)$. This completes the proof. Q.E.D.