Copyright © 1998 by the Genetics Society of America

Signatures of Population Expansion in Microsatellite Repeat Data

Marek Kimmel,* Ranajit Chakraborty,† J. Patrick King,* Michael Bamshad,‡ W. Scott Watkins‡ and Lynn B. Jorde‡

**Department of Statistics, Rice University, Houston, Texas 77251, †Human Genetics Center, University of Texas Health Science Center, Houston, Texas 77225, and ‡Eccles Institute of Human Genetics, University of Utah Health Sciences Center, Salt Lake City, Utah 84112*

Manuscript received May 30, 1997
Accepted for publication November 24, 1997

ABSTRACT

To examine the signature of population expansion on genetic variability at microsatellite loci, we consider a population that evolves according to the time-continuous Moran model, with growing population size and mutations that follow a general asymmetric stepwise mutation model. We present calculations of expected allele-size variance and homozygosity at a locus in such a model for several variants of growth, including stepwise, exponential, and logistic growth. These calculations in particular prove that population bottleneck followed by growth in size causes an imbalance between allele size variance and heterozygosity, characterized by the variance being transiently higher than expected under equilibrium conditions. This effect is, in a sense, analogous to that demonstrated before for the infinite allele model, where the number of alleles transiently increases after a stepwise growth of population. We analyze a set of data on tetranucleotide repeats that reveals the imbalance expected under the assumption of bottleneck followed by population growth in two out of three major racial groups. The imbalance is strongest in Asians, intermediate in Europeans, and absent in Africans. This finding is consistent with previous findings by others concerning the population expansion of modern humans, with the bottleneck event being most ancient in Africans, most recent in Asians, and intermediate in Europeans. Nevertheless, the imbalance index alone cannot reliably estimate the time of initiation of population expansion.

TANDEM repeat loci, with repeat motifs 2–6 nucleotides long, called microsatellites (Tautz 1993), have been shown to be extremely helpful in evolutionary studies (Chakraborty and Jin 1993a; Bowcock *et al.* 1994; Deka *et al.* 1995), forensic identification of individuals (National Research Council 1996), determination of parentage and relatedness of individuals (Chakraborty and Jin 1993b; Pena and Chakraborty 1994), and mapping genes in the genome (Cox-Matise *et al.* 1994; Hanis *et al.* 1996). This is because of their abundant distribution in the genome (Gyapay *et al.* 1994) and ease and automated procedure of typing (Lin *et al.* 1996). The relative efficiency of microsatellites in comparison to the classical genetic markers for all of the above applications mainly arises because of their high heterozygosity (Weissenbach *et al.* 1992), as well as ubiquity of polymorphism, even in inbred populations or species (Gilbert *et al.* 1990).

The use of microsatellite loci for evolutionary purposes, however, has been a subject of intense research in recent studies because the mechanisms that produce new variation at such loci are unusual in comparison to those of classical loci. While the exact mechanism of mutations at such loci is still not characterized at a molecular level (*e.g.*, Jeffreys *et al.* 1994), it is generally believed that the processes and the patterns of mutations at different tandem repeat loci may differ from locus to locus, depending on the motif as well as the size of alleles at each locus (Weber 1990; Weber and Wong 1993; Jin *et al.* 1996; Chakraborty *et al.* 1997). Empirical and theoretical studies indicate that for most microsatellite loci, mutations lead to stepwise changes of the repeat size of alleles although the relative frequencies of mutations leading to expansion may not be equal to those of contraction of allele sizes (Di Rienzo *et al.* 1994; Valdes *et al.* 1993; Shriver *et al.* 1993; Rubinsztein *et al.* 1995).

Therefore, we recently developed a general stepwise mutation model to study the population dynamics of microsatellite loci in which mutations may change the allele size in any arbitrary specified manner that is not necessarily symmetric (Kimmel *et al.* 1996; Kimmel and Chakraborty 1996). Under such models, in accordance with the previous results of simple stepwise mutation models (Moran 1975), even though the allele size distributions may be fluctuating, mutations and genetic drift will produce a stationary distribution of size differences among randomly chosen alleles from the population, and consequently, the population will have a steady-state value of homozygosity (heterozygosity) that is specified by a composite parameter, θ, the product of the effective size of the population and the rate of mutation at the locus (Kimmel and Chakraborty 1996).

In such formulations, it is assumed that the population

Corresponding author: Ranajit Chakraborty, Human Genetics Center, University of Texas Health Science Center, P.O. Box 20334, Houston, TX 77225. E-mail: rc@hgc9.sph.uth.tmc.edu

maintains a constant effective size during evolution. In contrast, through the analysis of distributions of nucleotide differences in pairwise comparison of mitochondrial DNA sequences from human populations, Rogers and Harpending (1992), Harpending *et al.* (1993), and Rogers (1995) have concluded that most human populations have experienced recent expansions. Several authors, however, have argued that natural selection (Di Rienzo and Wilson 1991), high levels of homoplasy associated with hypervariable nucleotide sites (Lundstrom *et al.* 1992), and population structure (Marjoram and Donnelly 1994) may also mimic the signature of population expansion on the distribution of nucleotide differences in pairwise comparisons of mtDNA sequence data. More recently, Bertorelle and Slatkin (1995) showed that when recurrent mutations at the same site (a more realistic mutation model for the mtDNA sequence data) are considered, the observed number of segregating sites does not always support the population expansion theory from the analysis of the mtDNA sequence data. In other words, specific assumptions of a mutation model may differentially affect different measures of genetic variation, and thus, inference regarding population history from different measures of genetic variation may not always be the same. Thus, because the mutation model for microsatellite loci is different from that of nucleotide sequence variation, it is important to examine the signature of population expansion on the genetic variance at microsatellie loci and to evaluate the effect of population expansion on different measures (*e.g.*, heterozygosity vs. variance of allele sizes) of variability at microsatellite loci.

The purpose of this research is to investigate such problems. Specifically, we present calculations of genetic variance (variance of allele sizes) and homozygosity (probability of size identity of alleles) at a microsatellite locus using a time-continuous Moran model (Moran 1975) for several variants of population growth possibly preceded by a bottleneck. From the expected variance of allele sizes and homozygosity in the population, we show that if the population growth model is ignored and these population measures are used to estimate the equilibrium value of θ, the variance-based estimator deviates from that based on homozygosity.

To quantify this imbalance of variance- and homozygosity-based estimates of θ, we define their ratio $\hat{\beta}$ as the imbalance index. Under the assumptions of our model, the parametric value of this imbalance index, β, when >1, is a signature of population expansion preceded by a bottleneck. Under different scenarios of population growth, we provide numerical calculations of such a ratio over time and apply the theory to data on 60 tetranucleotide loci surveyed in three major groups of human populations. Our results indicate that the tetranucleotide loci generally provide evidence of recent population expansion preceded by a bottleneck in all major human populations.

DYNAMICS OF MICROSATELLITE LOCI ACCORDING TO THE TIME-CONTINUOUS MORAN MODEL

Statistics used to describe a sample of alleles: Consider a sample of n haploid individuals or chromosomes and a locus with a denumerable set of alleles indexed by integer numbers. The expectation of the estimator of the within-population component of genetic variance,

$$\hat{V}/2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1), \tag{1}$$

where X_i is the size of the allele at the locus in the ith chromosome present and $\overline{X}$ is the mean of the X_i, is equal to $V(t)/2$, where

$$V(t) = E(\hat{V}) = E[(X_i - X_j)^2], \tag{2}$$

and X_i and X_j are the sizes of two alleles from the population (Kimmel *et al.* 1996). X_i and X_j are time-dependent random variables, *i.e.*, $X_i = X_i(t)$ and $X_j = X_j(t)$, but for notational simplicity, the argument t is suppressed frequently because the time dependence is always clear from the context.

If p_k denotes the relative frequency of allele k in the sample, then an estimator of homozygosity has the form

$$\hat{P}_0 = \left(n \sum_{k=1}^{n} p_k^2 - 1\right)/(n-1). \tag{3}$$

Note that the random variables X_i are not independent by only exchangeable. The expected value of $\hat{P}_0$, however, is the true homozygosity; *i.e.*,

$$P_0(t) = E(\hat{P}_0) = \sum_k \Pr[X_i = X_j = k]. \tag{4}$$

The latter equation can be demonstrated by using the definition of p_k as the fraction of chromosomes with allele of size k, *i.e.*, $p_k = n_k/n$, and further representing n_k as the sum of indicator variables δ_{kX_i} ($= 1$ when $X_i = k$; and $= 0$ otherwise), *i.e.*, $n_k = \Sigma_i \delta_{kX_i}$, substituting into Equation 3, and taking expectation.

The time-continuous Moran model: We consider the evolution of joint distributions of allele sizes in a stepwise mutation model with sampling from the finite allele pool. We assume the following:

1. The population is composed of a constant number of $2N$ haploid individuals. Each individual undergoes death/birth events according to a Poisson process with intensity 1 (mean length of life of each individual is equal to 1). Upon a death/birth event, a genotype for the individual is sampled with replacement from the $2N$ chromosomes present at this moment, including the chromosome of the just-deceased individual (time-continuous Moran model, Ewens 1979).
2. Each individual is independently subjected to a mutation that replaces an allele of size X with an allele of size $X + U$, where U is an integer-valued random variable with probability generating function (pgf)

$$\phi(s) = \sum_{u=-\infty}^{\infty} s^u \Pr[U = u] = \mathrm{E}(s^u), \tag{5}$$

defined for s on the unit circle of the complex plane or in its neighborhood. Mutations occur according to a Poisson process with intensity v.

Suppose that we follow the evolution of the distribution of allele sizes $X_1(t)$ and $X_2(t)$ of two individuals in the population. We are interested in the distribution of the difference between these two allele sizes. The respective pgf is denoted as follows:

$$R(s,t) = E[s^{X_1(t) - X_2(t)}].$$

$R(s,t)$ is a pgf of an integer-valued random variable. It is generally defined on the unit circle of the complex plane $|s| = 1$, or in its neighborhood. Consequently, it might be more appropriate to consider only $\tilde{R}(\phi,t) = R(e^{\iota\phi},t), \phi \in (-\infty,\infty)$, which is the characteristic function of the same random variable. For notational simplicity, however, it seems better to adhere to the pgf formalism and to use the characteristic function $\tilde{R}$ only when required.

In the next paragraphs, we consider the dynamics of $R(s,t)$ when the population size is changing according to various patterns.

The assumptions above can be used to derive a differential equation for studying the dynamics of the function $R(s,t)$ (our Equations 6 and 17). We omit these calculations, however, in favor of a derivation based on the coalescent representation of the model. This has an advantage of proving that our calculations also are valid for a diffusion approximation of the Wright-Fisher model.

Stepwise change in population size and the disequilibrium index: The ordinary differential equation that describes the dynamics of the pgf $R(s,t)$ is given by

$$\dot{R}(s,t) = -\{1/(2N) + 2v[1 - \psi(s)]\}\, R(s,t) + 1/(2N), \tag{6}$$

where $\dot{R}(s,t)$ is the derivative of $R(s,t)$ with respect to t and $\psi(s) = [\phi(s) + \phi(1/s)]/2$ is the symmetrized version of the pgf $\phi(s)$ of U. This differential equation is analogous to the one used in the analysis of genetic variation at electrophoretically determined protein loci (Wehrhahn 1975; Chakraborty and Nei 1982; Li 1976) under the stepwise mutation model (SMM; Ohta and Kimura 1973). In the present formulation, however, the distribution of allele size change caused by mutation (represented by the random variable U) can be general, multistep, and asymmetric.

A formal solution of this differential equation can be obtained,

$$R(s,t) = R(s,0)\exp[-a(s)t] + \frac{1 - \exp[-a(s)t]}{2Na(s)}, \tag{7}$$

where

$$a(s) = 1/(2N) + 2v[1 - \psi(s)]. \tag{8}$$

For $|s| = 1$, the solution tends to the equilibrium value

$$R(s,\infty,N) = [2Na(s)]^{-1}$$

as $t \to \infty$.

The stepwise change of population size is described as

$$N(t) = \begin{cases} N_0; & -\infty < t \le 0, \\ N; & t > 0. \end{cases}$$

Under this condition, Equation 7 assumes the form

$$R(s,t) = R(s,\infty,N_0)\exp[-a(s)t] + R(s,\infty,N)\{1 - \exp[-a(s)t]\}. \tag{9}$$

Based on Equation 9, it is possible to derive expressions for the genetic variance and homozygosity at a given repeat locus. The variance is equal to $V(t)/2$, where $V(t) = E\{[X_1(t) - X_2(t)]^2\} = \partial^2 R(1,t)/\partial s^2$, because $E[X_1(t) - X_2(t)] = 0$. Consequently,

$$V(t) = 4v\psi''(1)\,\{N_0\exp[-t/(2N)] + N\{1 - \exp[-t/(2N)]\}\}, \tag{10}$$

in which $\psi''(1)$ is the second derivative of $\psi(s)$ evaluated at $s = 1$. $V(t)$ clearly converges to $V(\infty) = 4vN\psi''(1) = \theta\psi''(1)$ as $t \to \infty$. If the single-step SMM is assumed, *i.e.*, if $\psi(s) = (s + s^{-1})/2$ and consequently $\psi''(1) = 1$, we obtain

$$V(\infty) = 4vN = \theta. \tag{11}$$

The expression for homozygosity requires evaluation of the zero-order (constant) term in the Laurent series expansion of $R(s, t)$, *i.e.*,

$$P_0(t) = \frac{1}{2\pi\iota}\oint \frac{R(s,t)}{s} ds,$$

with the integration path being a closed contour around the singularity at $s = 0$. It is convenient to choose the unit circle around the origin with the parameterization $s = \exp(\iota\phi)$. If the single-step SMM is assumed, *i.e.*, if $\psi(s) = (s + s^{-1})/2$, using the symmetry properties of the integrand, we obtain

$$\begin{aligned} P_0(t) = {} & \pi^{-1}\int_0^\pi \exp\{-\{1/(2N) + 2v[1 - \cos(\phi)]\}t\}/ \\ & \{1 + 4N_0v(1 - \cos(\phi))]\}d\phi \\ & + \pi^{-1}\int_0^\pi \{1 - \exp\{-\{1/(2N) + 2v[1 - \cos(\phi)]\}t\}\}/ \\ & \{1 + 4Nv\,[1 - \cos(\phi)]\}d\phi. \end{aligned} \tag{12}$$

As $t \to \infty$, $P_0(t)$ converges to a limit value that can be explicitly written as

$$P_0(\infty) = (1 + 8Nv)^{-1/2} = (1 + 2\theta)^{-1/2}. \tag{13}$$

Equations 11 and 13 provide two intuitive estimators of the composite parameter θ,

$$\hat{\theta}_V = \hat{V}, \tag{14}$$

called the (allele size) variance estimator of v, and

$$\hat{\theta}_{P_0} = (1/\hat{P}_0^2 - 1)/2, \tag{15}$$

the homozygosity (heterozygosity) estimator of θ. At equilibrium,

$$\frac{E(\hat{\theta}_V)}{E(\hat{\theta}_{P_0})} \approx \frac{V(\infty)}{[1/P_0(\infty)^2 - 1]/2} = 1,$$

which leads to a parametric definition of an index $\beta(t)$, given by

$$\beta(t) = \frac{V(t)}{[1/P_0(t)^2 - 1]/2}, \tag{16}$$

which represents an imbalance (caused by population size changes) at a microsatellite locus.

Arbitrary pattern of population size change: Formal substitution of $N(t)$ for N in Equation 6 yields

$$\dot{R}(s,t) = -a(s,t)R(s,t) + 1/[2N(t)], \tag{17}$$

where

$$a(s,t) = 1/[2N(t)] + 2v[1 - \psi(s)].$$

The solution obtained from the variation of constants is

$$R(s,t) = R(s,0)e^{-\int_0^t a(s,t)d\tau} + \int_0^t \frac{1}{2N(\tau)} e^{-\int_\tau^t a(s,u)du} d\tau. \tag{18}$$

As demonstrated in the appendix, Equations 17 and 18 can be obtained using the coalescent-based approach. Similarly as before, we derive expressions for variance and homozygosity,

$$V(t) = V(0)e^{-\int_0^t \frac{d\tau}{2N(\tau)}} + 2v\psi''(1) \int_0^t e^{-\int_\tau^t \frac{du}{2N(u)}} d\tau \tag{19}$$

and

$$P_0(t) = \pi^{-1} \int_0^\pi \tilde{R}(\phi,0) e^{-\int_0^t \tilde{a}(\phi,\tau)d\tau} d\phi + \pi^{-1} \int_0^\pi \int_0^t \frac{1}{2N(\tau)} e^{-\int_\tau^t \tilde{a}(\tau,u)du} d\tau d\phi, \tag{20}$$

where

$$\tilde{R}(\phi,0) = R(e^{i\phi},0)$$

and

$$\tilde{a}(\phi,t) = a(e^{i\phi},t).$$

If a mutation-drift equilibrium is assumed at time $t = 0$, we obtain

$$R(s,t) = R(s,\infty,N_0)e^{-\int_0^t a(s,\tau)d\tau} + \int_0^t \frac{1}{2N(\tau)} e^{-\int_\tau^t a(s,u)du} d\tau, \tag{21}$$

and $V(0) = 4vN_0\psi''(1)$. In this latter case,

$$V(t) = v[2\psi''(1)] \left[2N_0 e^{-\int_0^t \frac{d\tau}{2N(\tau)}} + \int_0^t e^{-\int_\tau^t \frac{du}{2N(u)}} d\tau \right]. \tag{22}$$

NUMERICAL EXAMPLES

Modeling of imbalance index $\beta(t)$ under different population growth patterns and initial conditions: We modeled the imbalance index $\beta(t)$, as defined in Equation 16, as a function of time (number of generations) for several patterns of population growth:

1. Stepwise population growth: $N(t) = N_0$, $t = 0$, and $N(t) = N$, $t > 0$.
2. Exponential population growth: $N(t) = N_0 \exp(\alpha t)$, $t \geq 0$, where the growth rate $\alpha = [\ln(N/N_0)]\}/T$ has been selected so that $N(t) = N$ if $t = T$.
3. Logistic population growth: $N(t) = K/[1 + (K/N_0 - 1)\exp(-\alpha t)]$, $t \geq 0$, where the growth rate α and the carrying capacity K have been selected so that $N(t) = N$ if $t = T$, and $N(t) = N/2$ if $t = T/2$.

Three types of initial conditions selected are as follows:

1. Mutation-drift equilibrium: $V(0) = 4vN_0$, $R(s,0) = R(s,\infty,N_0)$.
2. Initial population monomorphic: only a single allele present, hence $V(0) = 0$, $R(s,0) = 1$.
3. Initial population carrying two alleles: uniform mixture of two alleles differing in size by k repeats, with respective frequencies p and $q = 1 - p$, hence $V(0) = 2k^2pq$, $R(s,0) = (1 - 2pq) + pq(s^k + s^{-k})$.

Finally, one more complex growth pattern was contemplated, with population initially of large size N_{00}, dropping instantly to a smaller size N_0, and then regrowing exponentially to a final size N, *i.e.*,

$$N(t) = \begin{cases} N_{00}; & t < 0, \\ N_0 e^{\alpha t}; & t \geq 0, \end{cases} \tag{23}$$

where $\alpha = \ln(N/N_0)/T$ has been selected so that $N(t) = N$ if $t = T$. Technically, this variant can be computed for $t > 0$ as exponential growth starting from size N_0 but from equilibrium $R(s, \infty, N_{00})$ corresponding to N_{00}.

Population increase with parameters estimated from data on human populations: We used the numerical values obtained by Rogers and Harpending (1992), who fitted distributions of pairwise differences of numbers of segragating sites in mitochondrial DNA to the data of Cann *et al.* (1987). The second row of Table 1 in Rogers and Harpending (1992) contains estimates concerning the world's population expansion. Correcting the fact that Rogers and Harpending (1992) considered only females while we consider both genders, *i.e.*, multiplying all effective sizes by 2, we obtain expansion from $N_0 = 3{,}254$ to $N = 547{,}586$ within 120,000 yr or $T = 4{,}800$ generations, assuming generation times roughly equivalent to 25 yr. We combined these values with mutation rates $v = 10^{-4}$ and 5×10^{-4} typical for microsatellite loci (Weber and Wong 1993).

Figure 1, a and b, presents the $\beta(t)$ index values for the stepwise and exponential population growth, with equilibrium initial conditions. The index falls with time to values <1, the deviation increasing with the mutation rate v. The logistic growth (not shown) leads to an effect that is intermediate between those caused by the stepwise and exponential growth.

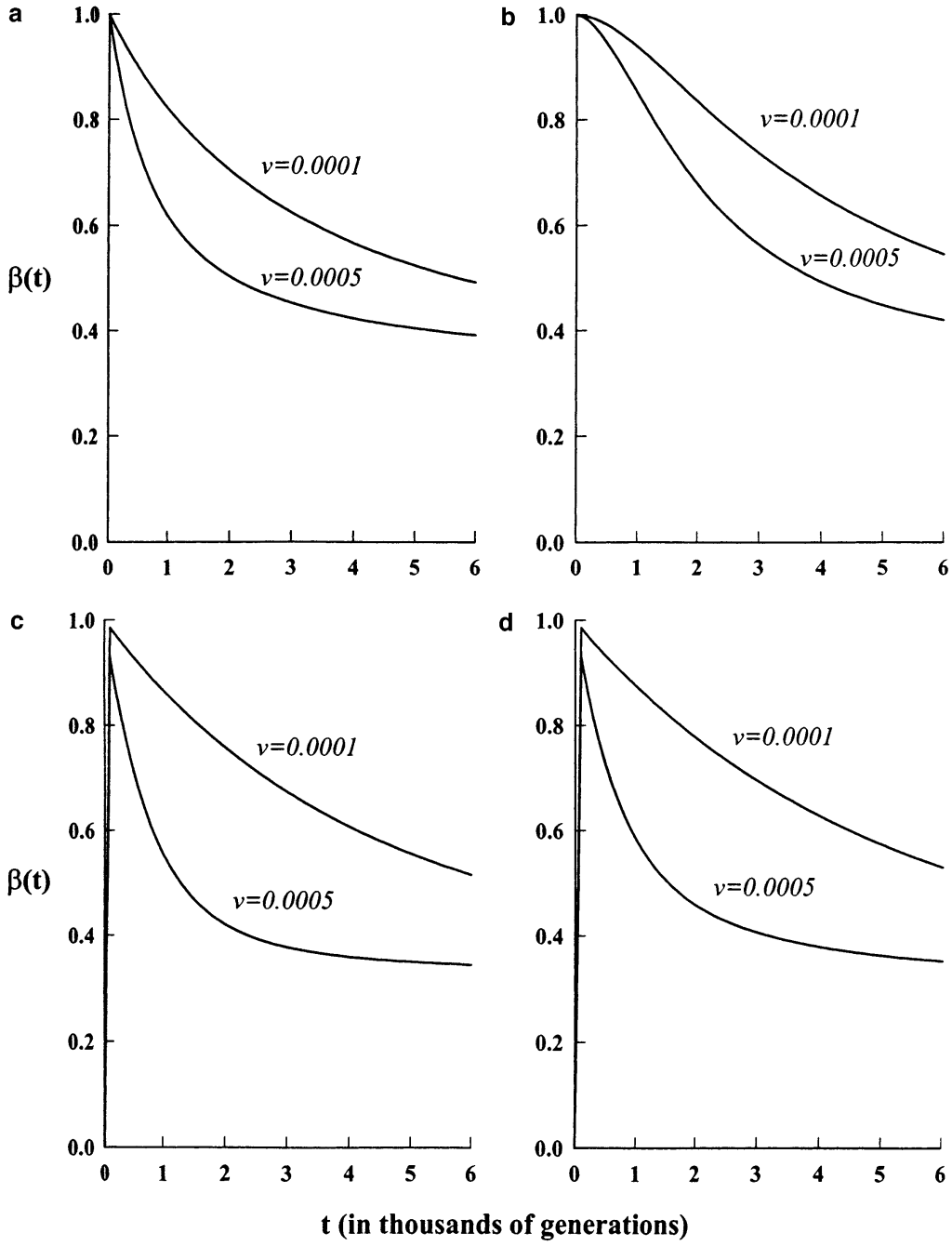

Figure 1.—Values of the $\beta(t)$ index for stepwise and exponential population growth corresponding to population expansion from N_0 = 3,254 to N = 547,586, within 120,000 yr or T = 4,800 generations, with mutation rates $v = 10^{-4}$ and 5×10^{-4}. Equilibrium initial conditions: (a) stepwise growth and (b) exponential growth. Monomorphic initial conditions: (c) stepwise growth and (d) exponential growth.

Figure 1, c and d, presents the $\beta(t)$ index values for stepwise and exponential population growth, with initial conditions corresponding to a monomorphic population. The index is initially close to 0, but then rapidly, during ~100 generations, increases to a value close to 1 and subsequently follows almost the same trajectory as the case of equilibrium initial conditions.

Figure 2, a and b, presents the $\beta(t)$ index values for stepwise and exponential population growth, with initial conditions corresponding to a mixture of two alleles with parameters $k = 5$, $p = q = \frac{1}{2}$. An interesting effect is observed: The index is initially much greater than 1 but falls to values between 1 and 2. Higher mutation rates yield lower values of the index.

Figure 3 presents the $\beta(t)$ index values for the bottleneck patterns of Equation 23, with the prebottleneck population size N_{00} = 40,000, N_0 = 3,254, and N = 547,586, as described above. Again, for an initial period, the index increases from 1 to values higher than 1, the increase being greater for greater mutation rates. After that initial period, an imbalance as in simple exponential growth is restored.

To examine the impact of the initial population size (N_0) on the imbalance index $\beta(t)$, in Figure 4, we pres-

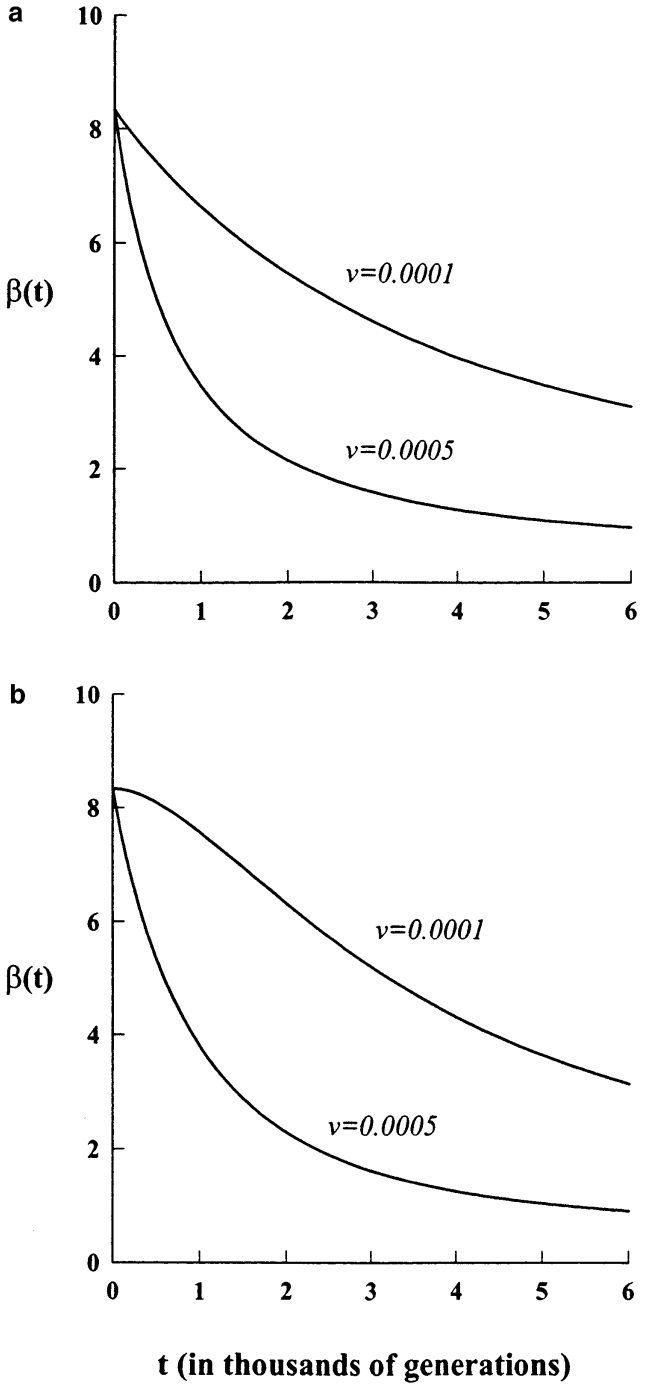

Figure 2.—Values of the $\beta(t)$ index for stepwise and exponential population growth, corresponding to population expansion from $N_0 = 3{,}254$ to $N = 547{,}586$, within 120,000 yr or $T = 4{,}800$ generations, with mutation rates $v = 10^{-4}$ and 5×10^{-4}. Initial conditions corresponding to a mixture of two alleles, with parameters $k = 5$, $p = q = \frac{1}{2}$. (a) Stepwise growth, (b) exponential growth.

ent the values of $\beta(t)$ as a function of t for three values of the initial population size: $N_0 = 10{,}000$, 20,000, and 50,000. As expected, larger N_0 diminishes the deviation of $\beta(t)$ from 1. Nevertheless, the signature of expansion [namely, $\beta(t) < 1$] is present for all initial sizes and for

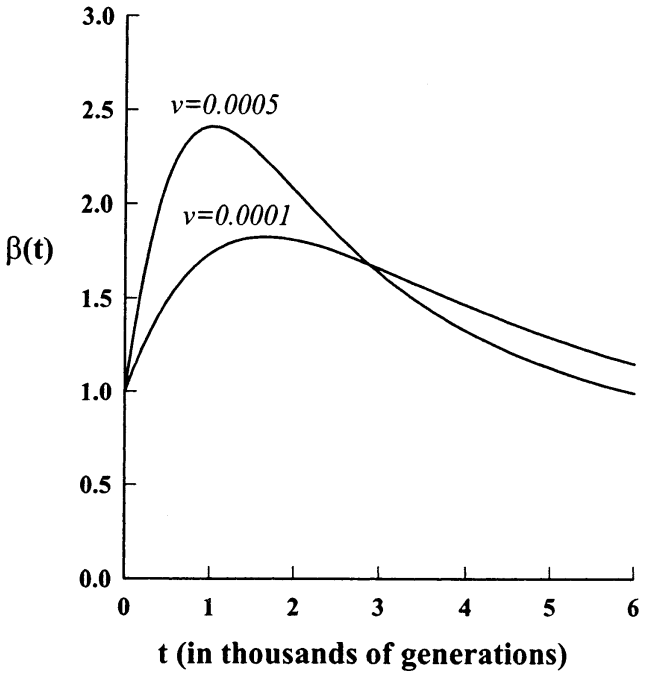

Figure 3.—Values of the $\beta(t)$ index for the bottleneck pattern of Equation 23 with the pre-bottleneck population size $N_{00} = 40{,}000$, $N_0 = 3{,}254$, $N = 547{,}586$, and $T = 4{,}800$ generations, with mutation rates $v = 10^{-4}$ and 5×10^{-4}.

both models of population growth (stepwise or exponential). Similar sensitivity studies demonstrate robustness of the bottleneck pattern of Equation 23.

In summary, if before expansion the population is at a mutation-drift equilibrium, the imbalance index deviates downwards from 1 [*i.e.*, $\beta(t) < 1$]. In contrast, if the population experiences a bottleneck preceding expansion, there will be a long (*e.g.*, several thousand generations) transient time period during which $\beta(t) > 1$ before showing the signature of expansion alone [$\beta(t) < 1$]. Figure 1, c and d, shows an obvious exception to this general rule, when the bottleneck is severe enough to make the population monomorphic before expansion, in which case $\beta(t) < 1$ for all times.

ANALYSIS OF DATA ON TETRANUCLEOTIDE LOCI

Jorde *et al.* (1995, 1997) recently analyzed allele frequency distributions at 60 tetranucleotide loci in a worldwide survey of human populations. These authors also describe the details of the loci surveyed, as well as the various characteristics of the allele frequency distributions at these loci. In this section, we investigate whether there is any imbalance between allele size variances and heterozygosity (homozygosity) observed in these data, as analyzed by the imbalance index $\beta(t)$ defined above. The purpose is to examine if such an imbalance, if it exists, is in accordance with the population expansion model of human

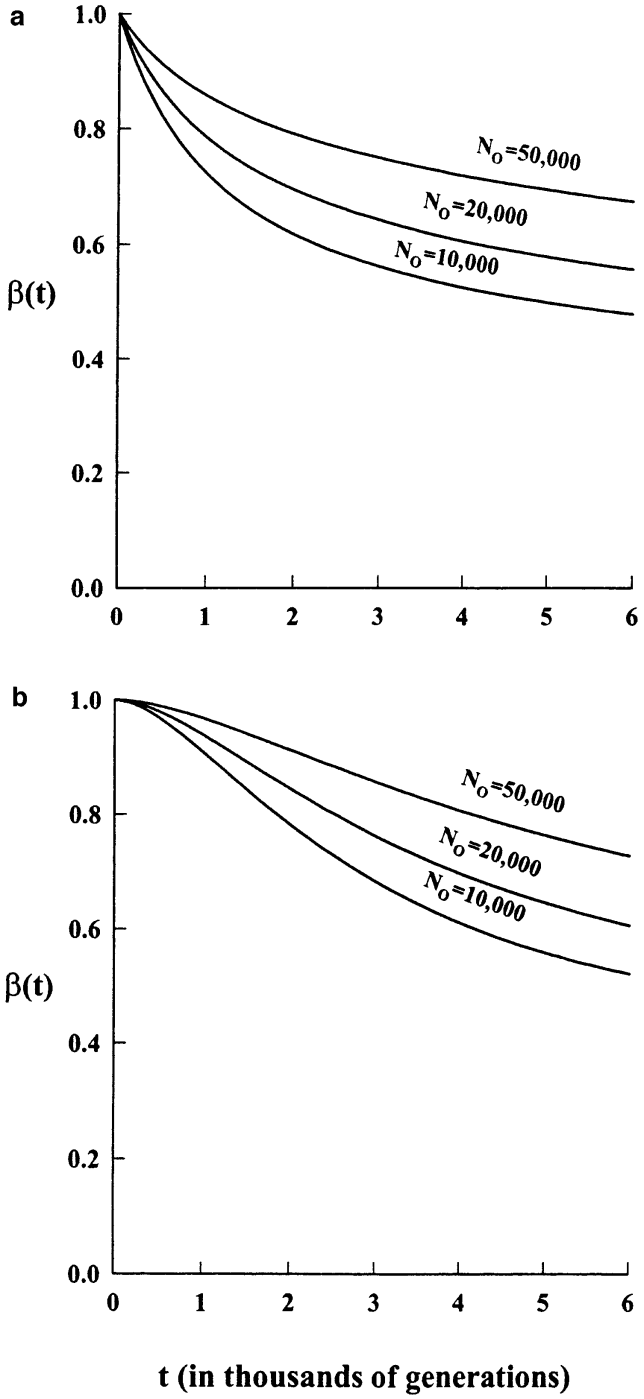

Figure 4.—Values of the $\beta(t)$ index for stepwise (a) and exponential (b) population growth, corresponding to population expansion from equilibrium condition with N_0 (= 10,000, 20,000, and 50,000) to N = 547,586, within 120,000 yr or T = 4,800 generations, with mutation rate $v = 5 \times 10^{-4}$.

populations suggested from the analysis of mtDNA variation reported by Rogers *et al.* (1992).

Three major groups of population, Asians, Africans, and Europeans, are considered for this purpose. For each population, the allele size variance and homozygosity at each locus were calculated from the distributions of allele frequencies within each of these population groups. Estimators $\hat{V}/2$ and $\hat{P}_0$ in Equations 1 and 3, respectively, averaged over the 60 loci were used for these computations for the respective parameters. The variance estimator $\hat{\theta}_V$ is obtained by equating $\theta = V$, while the homozygosity estimator $\hat{\theta}_{P_0}$ is obtained by equating $\theta = (P_0^{-2} - 1)/2$.

Finally, the estimator used has the form

$$\ln\hat{\beta} = \ln\hat{\theta}_{\bar{V}} - \ln\hat{\theta}_{\bar{P}_0} = \ln(\bar{\hat{V}}) - \ln[\bar{\hat{P}}_0^{-2} - 1)/2],$$

where $\bar{\hat{V}}$ and $\bar{\hat{P}}_0$ are estimates averaged over 60 loci.

Simulation studies were carried out to determine the statistical properties of the estimator $\ln\hat{\beta} = \ln\hat{\theta}_{\bar{V}} - \ln\hat{\theta}_{\bar{P}_0}$ under the null hypothesis of constant population size and mutation-drift equilibrium.

Figure 5 depicts histograms of $\ln\hat{\beta}$ based on coalescent simulations with different values of $\theta = 4Nv$. The estimator has an almost symmetric distribution centered around 0. For example, for $\theta = 10$, the 0.05 and 0.95 quantiles of the empirical distribution of $\ln v$ are $q_{0.05} = -0.24$ and $q_{0.95} = 0.21$, respectively.

Table 1 contains the values estimated from the data on three major groups of populations. The values of $\ln\hat{\beta}$ for Asians, Europeans, and Africans are equal to 0.60, 0.29, and 0.11, respectively.

Figure 6 depicts a comparison of the sample values of $\ln\hat{\beta}$ with the simulation-based quantiles (with 500 replications of coalescent simulations of 60 loci each) of the distribution of $\ln\hat{\beta}$ under the null hypothesis of constant population size and mutation-drift equilibrium. The value for Asians exceeds the 0.99 quantile. The value for Europeans is located between the 0.95 and the 0.99 quantiles. The value for Africans, residing around the 0.70 quantile, is not significantly different from 0.

The behavior of $\ln\hat{\beta}$ obtained from the data is consistent with the growth scenarios depicted in Figures 2 and 3, *i.e.*, $\hat{\beta} > 1$ or $\ln\hat{\beta} > 0$. Both of these scenarios assume a reduced diversity of the population at the time when population expansion begins ($t = 0$), representing the consequences of a pre-expansion bottleneck.

The gradation of sample values of $\ln\hat{\beta}$ is consistent with the bottleneck being most ancient in Africans, most recent in Asians, and of intermediate age in Europeans.

In general, this is in agreement with a population growth scenario with pre-expansion and the present effective sizes, as estimated by Rogers *et al.* (1992), although these authors do not explicitly model a bottleneck. Of course, from β indices alone, the exact pattern of population growth (stepwise *vs.* logistic or exponential) or the time of initiation of the expansion cannot be predicted reliably.

Another technical remark concerns alternative estimators of $\ln\beta$. For example, if $(\ln\hat{\beta})_i = (\ln v_{\hat{V}})_i - (\ln v_{\hat{P}0})_i$ is calculated for each individual locus and these individual estimators are averaged, one obtains an estimator that is seriously downward biased, although it has a

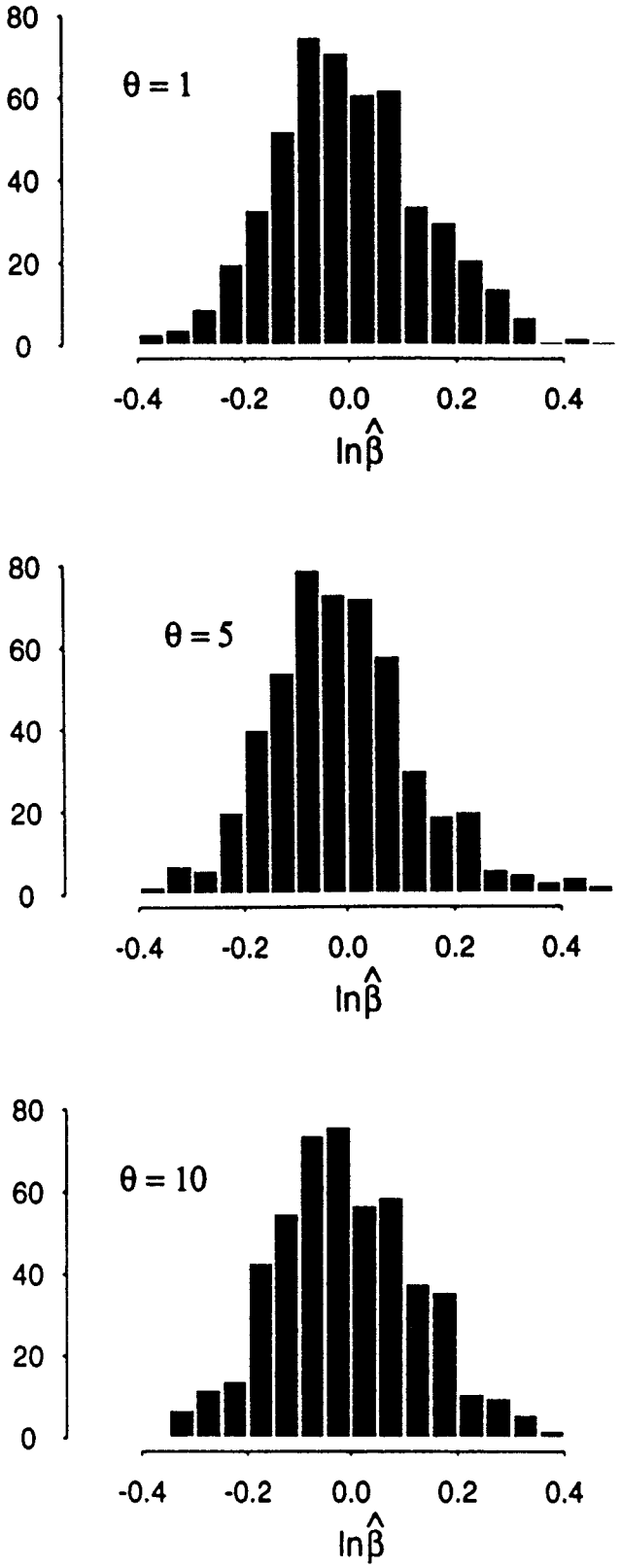

Figure 5.—Empirical distribution of $\ln\hat{\beta}$, from coalescent-based simulations, under the null hypothesis of constant population size and mutation-drift equilibrium and under the single-step stepwise mutation model. Estimates of $\ln\beta$ are based on averages of variance and homozygosity over 60 loci. Five hundred simulations were run for each assumed value of parameter θ.

TABLE 1

Estimates of parameter θ and of disequilibrium index β based on data for three major human populations

	Africans	Europeans	Asians
$\hat{\theta}_{\bar{V}} = \bar{V}e$	9.19	8.56	8.97
$\bar{P}_0$	0.24	0.27	0.30
$\hat{\theta}_{\bar{P}_0}$	8.20	6.44	4.94
$\hat{\beta}$	1.12	1.33	1.82
$\ln\hat{\beta}$	0.11	0.29	0.60

lower variance than the one we used (based on simulations, not shown). For our purposes, it is more appropriate to have a less biased estimator. Furthermore, the estimator we used also has a lower mean square error than the one mentioned above.

DISCUSSION

Our theory indicates that population expansion leaves a strong signature on allele size distributions, and the signature is specific for different major human populations. The departure from the equilibrium value of $\ln\hat{\beta}$ is strongest in Asians, weakest in Africans, and intermediate in Europeans. This can be translated into the bottleneck being most ancient in Africans, least ancient in Asians, and of intermediate age in Europeans. This, in turn, is consistent with a scenario in which a small subpopulation emerges from Africa and moves via Europe to Asia, with some of its descendants settling en route and expanding, possibly replacing the preexisting populations.

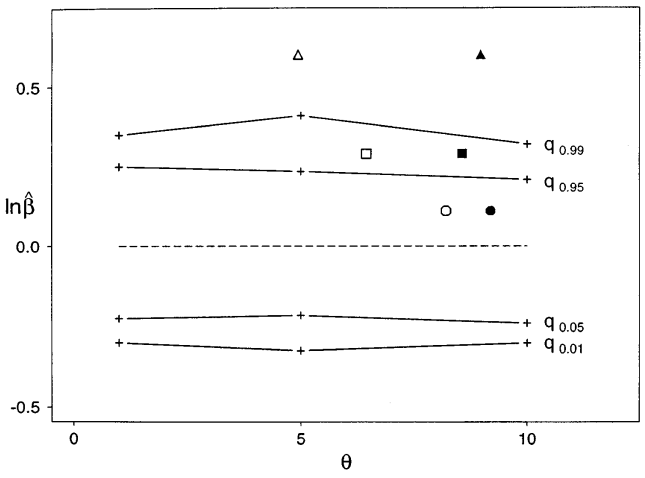

Figure 6.—Continuous lines: Simulation-based 0.01, 0.05, 0.95, and 0.99 quantiles of the distribution of $\ln\hat{\beta}$, under the hypothesis of constant size and mutation-drift equilibrium and plotted against the assumed values of parameter θ. Symbols: Data-based estimates of $\ln\hat{\beta}$ for the three major human populations—△, Asians; □, Europeans; ○, Africans—plotted against estimates of θ based on variance (solid symbols) and on homozygosity (empty symbols).

Before considering the implications of these findings, recall that any signature of past population size changes through the imbalance index β requires unbiased estimation of the index. We adopted the estimation procedure where $\ln\theta_V$ and $\ln\theta_{P_0}$ were estimated from average (over loci) estimates of V and P_0 to obtain $\ln\hat{\beta} = \ln\hat{\theta}_{\bar{V}} - \ln\hat{\theta}_{\bar{P}_0}$. While, in theory, locus-specific estimates of lnβ can be obtained, our simulations (not shown) indicate that $\hat{\beta}$, estimated in this fashion, is severely biased downwards (*i.e.*, in the direction $\beta < 1$), even when population size is constant and the population remains in mutation-drift equilibrium throughout time.

The theory described above also indicates that the deviation from $\beta = 1$ is of a qualitatively different pattern for different scenarios of past population size changes. For example, a population at a mutation-drift equilibrium, when it suddenly or gradually increases in size, will produce $\beta < 1$, while if it experiences a bottleneck followed by expansion, it will produce β transiently >1 and subsequently falling <1. With realistic values of parameters (Figure 3), the transient values of $\beta > 1$ can persist for several thousand generations. These patterns, which are due to fluctuations of population sizes, these patterns are valid for a general stepwise mutation model. Because any general form of $\psi(s)$ (Equations 22 and 20) can yield $\beta \neq 1$, we argue that the specificity of mutation pattern is not the critical determinant of the signature of population expansion preceded by bottlenecks at different time points, as noted in the present work.

The importance of the implications of our findings is worth discussing. Expansion of population size, preceded by bottleneck events that appear to have occurred at different points in time for the three major human populations, is consistent with a replacement model (Stringer 1989) of the origin of modern humans, while it tends to argue against the multiregional model (Weidenreich 1939), which maintains that humans probably have not experienced a major bottleneck. We also argue that should the recent expansion of size apply to most human populations, evolutionary inference based on summary statistics of microsatellite variation should be viewed with caution. For example, when genetic distances are based on indices related to heterozygosity alone (as in the case of Nei's distance D_a; Nei *et al.* 1983), the branch lengths and topology may be grossly misspecified. So will be the case of allele size variance–based measures of genetic distance (Goldstein *et al.* 1995; Slatkin 1995; Kimmel *et al.* 1996).

Second, deviation from mutation-drift equilibrium is not necessarily an indicator of selective forces operating on the microsatellites. Demographic history of populations, as shown in our analysis, can produce deviation that cannot always be distinguished from certain types of selection (see Bertorelle and Slatkin 1995).

Third, note that the present analysis indicates that the within-population variance of allele size is different from its mutation-drift equilibrium value for a growing population, and this departure is dependent on the mutation rate at the locus, as well as the growth pattern of the population. Although in the present work we used data on tetranucleotide loci alone, the impact of these findings on the estimates of relative mutation rates of different motif types of microsatellites is also important. We argue that although Chakraborty *et al.* (1997) used a mutation-drift equilibrium model to estimate the relative mutation rates of di-, tri-, and tetranucleotide loci, their conclusions are consistent with the analyses of the present set of data. This is so because Equation 20 clearly shows that even in the nonequilibrium case (caused by population size change), the ratio of expected variances between loci is simply given by the respective ratio of their mutation rates.

Finally, we note that an observed imbalance such as the one noted in the present analysis is not necessarily caused by population expansion alone. There could be possible effects of population structure superimposed on this factor (data considered here are in fact from a number of different national populations within each group), and even the different loci may be subject to differential allele size constraints.

This work was supported by grants GM 41399 (R.C.), GM 58545 (R.C. and M.K.), and RR 00064 (L.B.J., W.S.W., and M.B.) from the National Institutes of Health, as well as grants DMS 9409909 (M.K.), DBS 9310105 (L.B.J., W.S.W., and M.B.), and DBS 9514733 (L.B.J., W.S.W., and M.B.). The authors also acknowledge support from the National Science Foundation, grant 1T15LM07093-04 from the National Library of Medicine (J.P.K.), and the Keck's Center for Computational Biology at Rice University (M.K. and J.P.K.).

LITERATURE CITED

Bertorelle, G., and M. Slatkin, 1995 Number of segregating sites in expanding human populations, with implications for estimates of demographic parameters. Mol. Biol. Evol. **12:** 887–892.

Bowcock, A. M., R.-A. Linares, J. Tomfohrde, E. Minch, J. R. Kidd and L. L. Cavalli-Sforza, 1994 High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368:** 455–457.

Cann, R., M. Stoneking and A. C. Wilson, 1987 Mitochondrial DNA and human evolution. Nature **325:** 31–36.

Cakraborty, R., and M. Nei, 1982 Genetic differentiation of quantitative characters between populations of species. I. Mutation and random genetic drift. Genet. Res. Camb. **39:** 303–314.

Chakraborty, R., and L. Jin, 1993a A unified approach to study hypervariable polymorphisms: statistical considerations of determining relatedness and population distances, pp. 153–175 in *DNA Fingerprinting: State of the Science*, edited by S. D. J. Pena, R. Chakraborty, J. T. Epplen and A. J. Jeffreys. Birkhauser, Basel.

Chakraborty, R., and L. Jin, 1993b Determination of relatedness between individuals by DNA fingerprinting. Hum. Biol. **65:** 875–895.

Chakraborty, R., M. Kimmel, D. N. Stivers, R. Deka and L. J. Davison, 1997 Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc. Natl. Acad. Sci. USA **94:** 1041–1046.

Cox-Matise, T., M. Perlin and A. Chakravarti, 1994 Automated construction of genetic linkage maps using an expert system (multiMap): A human genome linkage map. Nature Genet. **6:** 384–390.

Deka, R., M. D. Shriver, L. M. Yu, R. E. Ferrell and R. Chakraborty, 1995 Intra- and inter-population diversity at short tandem repeat loci in diverse populations of the world. Electrophoresis **16:** 1659–1664.

Di Rienzo, A., and A. C. Wilson, 1991 Branching pattern in the

evolutionary tree from human mitochondrial DNA. Proc. Natl. Acad. Sci. USA **88:** 1597–1601.
Di Rienzo, A., A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin *et al.*, 1994 Mutational process of simple-sequence repeat loci in human populations journal. Proc. Natl. Acad. Sci. USA **91:** 3166–3170.
Ewens, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, New York.
Gilbert, D. A., N. Lehman, S. J. O'Brien and R. K. Wayne, 1990 Genetic fingerprinting reflects population differentiation in California Channel Island fox. Nature **344:** 764–767.
Goldstein, D. B., A. R. Linares, M. W. Feldman and L. L. Cavalli-Sforza, 1995b An evaluation of genetic distances for use with microsatellite loci. Genetics **139:** 463–471.
Gyapay, G., J. Morissette, A. Vignal, C. Dib, C. Fizames *et al.*, 1994 The 1993–1994 Genethon human genetic linkage map. Nature Genet. **7:** 246–339.
Hanis, C. L., E. Boerwinkle, R. Chakraborty, D. L. Ellsworth, P. Concannon *et al.*, 1996 A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. Nature Genet. **13:** 161–166.
Harpending, H. C., C. S. T. Sherry, A. R. Rogers and M. Stoneking, 1993 The genetic structure of ancient human populations. Curr. Anthropol. **34:** 483–496.
Jeffreys, A. J., K. Tamaki, A. MacLeod, D. G. Monckton, D. L. Neil *et al.*, 1994 Complex gene conversion events in germline mutation at human minisatellites. Nature Genet. **6:** 136–145.
Jin, L., C. Macaubas, J. Mallmayar, A. Kimura and E. Mignot, 1996 Mutation rate varies among alleles at a microsatellite locus: phenotypic evidence. Proc. Nat. Acad. Sci. USA **93:** 15285–15288.
Jorde, L. B., M. J. Bamshad, W. S. Watkins, R. Zenger, A. E. Fraley *et al.*, 1995 Origins and affinities of modern humans: A comparison of mitochondrial and nuclear genetic data. Am. J. Hum. Genet. **57:** 523–538.
Jorde, L. B., A. R. Rogers, W. S. Watkins, P. Krakowiak, S. Sung *et al.*, 1997 Microsatellite diversity and the demographic history of modern humans. Proc. Natl. Acad. Sci. USA **94:** 3100–3103.
Kimmel, M., and R. Chakraborty, 1996 Measures of variation at DNA repeat loci under a general stepwise mutation model. Theor. Pop. Biol. **50:** 345–367.
Kimmel, M., R. Chakraborty, D. N. Stivers and R. Deka, 1996 Dynamics of repeat polymorphisms under forward-backward mutation model: Within- and between-population variability at microsatellite loci. Genetics **143:** 549–555.
Li, W.-H., 1976 Electrophoretic identity of proteins in a finite population and genetic distance between taxa. Genet. Res. Camb. **28:** 119–127.
Lin, Z., X. Cui and H. Li, 1996 Multiplex genotype determination at a large number of gene loci. Proc. Natl. Acad. Sci. USA **93:** 2582–2587.
Lundstrom, R., S. Tavare and R. H. Ward, 1992 Modelling evolution of the human mitochondrial genome. Math. Biosci. **112:** 319–335.
Marjoram, P., and P. Donnelly, 1994 Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human populations. Genetics **136:** 673–683.
Moran, P. A. P., 1975 Wandering distributions and the electrophoretic profile. Theor. Pop. Biol. **8:** 318–330.
Nei, M., F. Tajima and Y. Tateno, 1983 Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. J. Mol. Evol. **19:** 153–170.
National Research Council, 1996 *The Evaluation of Forensic DNA Evidence by National Research Council*, National Academy Press, Washington DC.
Ohta, T., and M. Kimura, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22:** 201–204.
Pena, S. D. J., and R. Chakraborty, 1994 Paternity testing in the DNA era. Trends Genet. **10:** 204–209.
Rogers, A. R., 1995 Genetic evidence for a Pleistocene population explosion. Evolution **49:** 608–615.
Rogers, A. R., and H. C. Harpending, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552–569.
Rubinsztein, D. C., W. Amos, J. Leggo, S. Goodburn, S. Jain *et al.*, 1995 Microsatellite evolution—evidence for directionality and variation in rate between species. Nature Genet. **10:** 337–343.
Shriver, M. D., L. Jin, R. Chakraborty and E. Boerwinkle, 1993 VNTR allele frequency distributions under the stepwise mutation model—a computer simulation approach. Genetics **134:** 983–993.
Slatkin, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics **139:** 457–462.
Stringer, C. B., 1989 Neanderthals, their contemporaries and modern human origin, pp. 351–355 in *Hominidae*, edited by G. Giacobini. Jaca Book, Milan.
Tautz, D., 1993 Notes on the definition and nomenclature of tandemly repetitive DNA sequence, pp. 21–28 in *DNA Fingerprinting: State of the Science*, edited by S. D. J. Pena, R. Chakraborty, J. T. Epplen and A. J. Jeffreys. Birkhauser, Basel.
Vales, A. M., M. Slatkin and N. B. Freimer, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133:** 737–749.
Weber, J. L., 1990 Informativeness of human $(dC\text{-}dA)_n \cdot (dG\text{-}dT)_n$ polymorphisms. Genomics **7:** 524–530.
Weber, J. L., and C. Wong, 1993 Mutation of human short tandem repeats. Hum. Mol. Genet. **2:** 1123–1128.
Wehrhahn, C. F., 1975 The evolution of selectively similar electrophoretically detectable alleles in finite natural populations. Genetics **80:** 375–394.
Weidenreich, F., 1939 Six lectures on *Sinanthropus pekinensis* and related problems. Bull. Geol. Soc. China **19:** 1–110.
Weissenbach, J., G. Gyapay, C. Dib, A. Vignal, P. Morrissette *et al.*, 1992 A second-generation linkage map of the human genome. Nature **359:** 794–801.

Communicating editor: N. Takahata

APPENDIX

Coalescent-based derivation of expression (Equation 17): Let us consider the present time (t) as a reference point, and let us introduce the reverse time τ^* such that $\tau^* = t - \tau$, where τ is the chronological time assuming value $\tau = t$ at the present. Let us further denote $N^*(\tau^*) = N(t - \tau^*)$ and $R^*(s,\tau^*) = R(s,t - \tau^*)$. Suppose that lineages of two chromosomes from the population coalesce at the reverse time $T = \tau^*$. Then, under the SMM,

$$R^*(s,0|T = \tau^*) = e^{2\nu\tau^*[1-\psi(s)]}.$$

The distribution of the nonnegative random variable T has hazard rate $[2N^*(\tau^*)]^{-1}$, $\tau^* \geq 0$, equal to the coalescence intensity. T is proper if $\int_0^\infty [2N^*(\tau^*)]^{-1} d\tau^* = \infty$. Therefore,

$$\begin{aligned} R^*(s,0) &= \int_0^\infty R^*(s,0|T = \tau^*) f_T(\tau^*) d\tau^* \\ &= \int_0^\infty e^{2\nu\tau^*[1-\psi(s)]} [2N^*(\tau^*)]^{-1} e^{-\int_0^{\tau^*} [2N^*(u^*)]^{-1} du^*} d\tau^*. \end{aligned}$$

Passing to the usual time, we obtain

$$\begin{aligned} R(s,t) &= \int_{-\infty}^0 e^{2\nu(t-\tau)[1-\psi(s)]} [2N(\tau)]^{-1} e^{-\int_\tau^t [2N(u)]^{-1} du} d\tau \\ &= \int_{-\infty}^t [2N(\tau)]^{-1} e^{-\int_\tau^t a(s,u) du} d\tau. \end{aligned}$$

But this is equal to

$$\begin{aligned} R(s,t) &= \left[\int_{-\infty}^0 [2N(\tau)]^{-1} e^{-\int_\tau^0 a(s,u) du} d\tau \right] e^{-\int_0^t a(s,u) du} \\ &\quad + \int_0^t [2N(\tau)]^{-1} e^{-\int_\tau^t a(s,u) du} d\tau, \end{aligned}$$

which is identical as Equation 18, considering that

$$R(s,0) = \int_{-\infty}^0 [2N(\tau)]^{-1} e^{-\int_\tau^0 a(s,u) du} d\tau.$$