The Impact of Population Expansion and Mutation Rate Heterogeneity on DNA Sequence Polymorphism

Stéphane Aris-Brosou and Laurent Excoffier†*

*Institut National Agronomique Paris-Grignon, France; and †Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Switzerland

In order to study the effect of mutation rate heterogeneity on patterns of DNA polymorphism, we simulated samples of DNA sequences with gamma-distributed nucleotide substitution rates in stationary and expanding populations. We find that recent population expansions and mutation rate heterogeneity have similar effects on several polymorphism indicators, like the shape and the mean of the observed pairwise difference distribution, or the number of segregating sites. The inferred size of population expansion thus appears overestimated if nucleotides have dissimilar substitution rates. Interestingly, population expansion and uneven mutation rates have contrasting effects on Tajima's D statistic when acting separately, and the consequence on the associated test of selective neutrality is investigated. The patterns of polymorphism of several human populations analyzed for the mitochondrial control region are examined, mainly showing the difficulty in quantifying the respective contribution of past demographic history and uneven mutation rates from a single sampled evolutionary process. However, substitution rates appear more heterogeneous in the second hypervariable segment of the control region than in the first segment.

Introduction

New molecular data are providing information that is qualitatively different from that previously available, as the number of evolutionary steps separating a given pair of genes can be estimated. The number of observed differences between two gene copies is indicative of the amount of time since these genes diverged (Watterson 1975; Tajima 1983; Hudson 1990; Slatkin 1995), and can be used to infer different aspects of the natural history of extant populations (Avise 1994). For instance, the genetic structure of populations can be inferred from the distribution of observed pairwise differences (Excoffier, Smouse, and Quattro 1992), the amount of gene flow between populations can be estimated from molecular phylogenies (Slatkin and Maddison 1989), and a star phylogeny is consistent with a population expansion after a bottleneck (Slatkin and Hudson 1991). More recently, a series of papers have investigated the possibility of inferring the occurrence, the timing, and the level of past demographic expansions in finite populations by examining the shape of the distribution of observed pairwise differences (also called the mismatch distribution) (Rogers and Harpending 1992; Harpending et al. 1993; Rogers and Jorde 1995). This new methodology was introduced after the observation that many such distributions were unimodal in the mitochondrial control region of human populations (Di Rienzo and Wilson 1991). Such a distribution shape is indeed very unlikely under population stationarity, because the distribution of pairwise differences among genes reflects the shape of the gene genealogy, which may considerably differ between several outcomes of the same evolutionary process (Slatkin and Hudson 1991, and see figure 2, column 1). A unimodal distribution is, however, expected after a sudden and large demographic expansion (Slatkin and Hudson 1991; Rogers and Harpending 1992; Marjoram and Donnelly 1994). The analytical model developed by Rogers and Harpending (1992) to infer past population bottlenecks assumes a random substitution process and equal mutation rates for all nucleotide positions. Both assumptions have been shown to be violated for mitochondrial DNA (mtDNA). First, there is a well-known substitution bias in favor of transitions in mammalian mtDNA (Brown et al. 1982). Second, the uniform mutation rate model has been recently challenged. Several studies have shown that the number of substitutions per site was not Poisson distributed as expected under the uniform model in the mitochondrial control regions (Kocher and Wilson 1991; Hasegawa et al. 1993; Wakeley 1993). It rather followed a negative binomial distribution, suggesting (but not necessarily implying) that the mutation rates varied from site to site according to a gamma distribution (Uzzel and Corbin 1971). Although the exact underlying molecular mechanisms are unknown, it means that a few sites are the target of most mutation events and therefore act as mutational hot spots, whereas a majority of sites are not variable. Mutation rate heterogeneity and transition bias should therefore result in a reduced number of observed differences between pairs of genes. The effects of gamma-distributed mutation rates have been previously studied when estimating the amount of transition bias (Wakeley

Key words: gamma-distributed mutation rates, mutation rate heterogeneity, pairwise difference distribution, DNA polymorphism, human molecular evolution, Monte Carlo simulation, demographic expansion, mitochondrial DNA.

Address for correspondence and reprints: Stéphane Aris-Brosou and Laurent Excoffier, Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, CP 511, 1211 Geneva 24, Switzerland.

Mol. Biol. Evol. 13(3):494–504. 1996

1994) and nucleotide substitution rates (Tamura and Nei 1993; Rzhetsky and Nei 1994), but their effect on either the shape of the distribution of observed pairwise differences or on other indices of DNA polymorphism at the population level are largely unknown. Although it has been claimed that gamma-distributed rates would not significantly affect the number of differences between random pairs of genes (Rogers 1992), analytical work has shown that unimodal distributions of pairwise differences may be obtained when a given fraction of nucleotide sites are mutating much faster than the other sites (Lundstrom, Tavaré, and Ward 1992). Bertorelle and Slatkin (1995) have also recently shown that other indices of polymorphism, such as the number of segregating sites or Tajima's D statistic (Tajima 1989*a*), could be also affected if nucleotide sites were discretely distributed between slow- and fast-evolving sites.

In this paper, we use a Monte Carlo approach to simulate uneven substitution rates in stationary and expanding populations. The effects of sudden demographic expansions and uneven mutation rates on the pattern of DNA polymorphism are then compared. The problem of distinguishing between these two effects, and thus of inferring past demographic events from molecular data, is discussed in the context of human mtDNA sequence polymorphism.

Material and Methods

Simulation Model

Samples of 100 DNA sequences with a length of 300 base pairs (bp) were generated in stationary and expanding populations, with or without mutation rate heterogeneity and with transitional bias using a Monte Carlo approach. Gene genealogies (topologies and branch lengths) were first built using a coalescent simulation model based on a continuous time approximation and described in Hudson (1990). Under this model, the time t during which the sample has j distinct lineages is approximately exponentially distributed with mean $(j(j - 1)/2)^{-1}$. Time is expressed here in units of M generations, where M is equal to the present population inbreeding effective size N for haploids, or $2N$ for diploids. Episodes of rapid population growth (sudden expansions) were reflected on the genealogy by rescaling the length of the branches by a factor equal to M_1/M, where M_1 is the size before the expansion (Hudson 1990). In our simulations, all expansions occurred $t = 0.1M$ generations ago. As our application deals with human mtDNA, the time of expansion was chosen in order to approximately match the known demographic history of human populations. The human effective female population size has been estimated from mtDNA sequence polymorphism to be around 5,000 females (Vigilant et al. 1989; Graven et al. 1995). In this case, a time of 0.1 M generations is equal to 10,000 years, assuming a generation time of 20 years. This date corresponds approximately to the beginning of the Neolithic expansion.

We have simulated an independent substitution process occurring on the branches of the genealogy, starting from an arbitrary ancestor sequence of DNA, as a two-step process. The number of substitutions occurring on each branch of the genealogy was first assumed to follow a Poisson distribution of parameter $\theta t/2$, where $\theta = 2Mu$, u is the mean substitution rate for the whole sequence, and t is the branch length expressed in units of M generations. Substitutions were then randomly allocated to sites using the conditional gamma probability distribution described below and shown in figure 1, which allows for uneven substitution rates. Upon mutation, the target nucleotide was randomly chosen depending on a given transition/transversion ratio. The probability of a mutation being a transition was then set to 95%, in agreement with estimates obtained from human mtDNA (Brown et al. 1982; Horai and Hayasaka 1990; Graven et al. 1995). The final states of the sequences in the sample were recorded, as well as the actual number of mutations having occurred between all pairs of sequences.

Generation of Gamma-distributed Mutation Probabilities

Analytical studies of uneven mutation rates become extremely complex when there are more than two classes of mutation rates (Lundstrom, Tavaré, and Ward 1992). We have therefore chosen to study the effect of gamma-distributed mutation rates on the observed pattern of sequence polymorphism by simulating gamma-distributed hit probabilities. A hit probability is defined as the conditional probability for a given site to be the target of a substitution knowing that a mutation event has occurred. It is obtained by drawing random gamma-distributed numbers, a procedure described below.

The probability density of a gamma-distributed variate X is described by

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \cdot e^{-\beta x} \qquad (1)$$

where $\Gamma(\cdot)$ is the standard gamma function, $\alpha = \bar{X}^2/V(X)$, $\beta = \bar{X}/V(X)$, $\bar{X}$ is the mean and $V(X)$ is the variance. For simplification purposes, we arbitrarily set the parameter β to 1, a value very close to those estimated by Wakeley (1993) for human mtDNA control region sequences. This simplification has no major consequence on the shape of the gamma distribution as it is primarily controlled by the parameter α, but it implies that the mean is equal to the variance, as in the Poisson distribution. A gamma-distributed variate y_i is then found by numer-

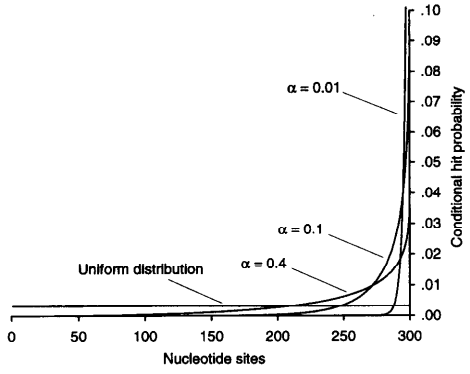


FIG. 1.—Gamma-distributed hit probabilities for different amounts of mutation rate heterogeneities (see text).

ically solving the incomplete gamma distribution (see Press et al. 1992)

$$P(y_i, \alpha) = \frac{1}{\Gamma(\alpha)} \int_0^{y_i} e^{-x} x^{\alpha-1}\, dx. \qquad (2)$$

Thus, a gamma-distributed variate y_i is defined such that the area under the incomplete gamma distribution curve before y_i is equal to a given number $P(y_i, \alpha)$. The values of $P(y_i, \alpha)$ can be randomly drawn from a uniform distribution between zero and one, but in order to uniformly cover the variation range of the y_i's, we have set the $P(y_i, \alpha)$ equal to $i/(m + 1)$, i varying between 1 and m, m being the length of the sequence. The conditional gamma probability distributions shown in figure 1 were finally obtained for different α values by normalizing the y_i as $z_i = y_i/\Sigma^m y_i$. Thus, for each α value, the mutation rates are different for all sites. They are, however, assumed to be constant throughout time at each site, and unchanged over all simulated gene genealogies. These deterministic mutation rates introduce less variation than a completely random mutation process that would require much more computing time under our simulation scheme, but both mutation models should lead to identical expectations.

The shape parameter α has been previously estimated from different human mtDNA sequence data sets by fitting the distribution of the number of substitutions per site estimated on most-parsimonious phylogenetic trees to a negative binomial (Johnson and Kotz 1973, p. 131). It was found to vary between 0.11 (Kocher and Wilson 1991; Tamura and Nei 1993) and 0.47 (Wakeley 1993). The smaller α values indicate larger departures from uniform mutation rates (fig. 1). Note that these estimated values may be overestimated since the principles of parsimony or minimum evolution used in most phylogenetic reconstruction techniques are violated when variable mutation rates are allowed. In order to study the effect of increasing departure from the uniform model, we have used four different α values (0.4, 0.1, 0.05, and 0.01) in our simulations.

Molecular Polymorphism Statistics

The consequence of sudden population expansion and variable mutation rates has been studied for several statistics of population sequence polymorphism, such as the number of observed sequences in the sample (k), the observed number of polymorphic sites (S), the mean number of nucleotide differences between pairs of sequences (π), and Tajima's (1989*a*) D statistic, defined as

$$D = \frac{\theta_a - \theta_b}{\sqrt{\mathrm{var}(\theta_a - \theta_b)}}, \qquad (3)$$

where $\theta_a = \pi$ and $\theta_b = S/\Sigma_i^{n-1} i^{-1}$. Note that Tajima's D is commonly used as a selective neutrality test statistic under the infinite-site model (Watterson 1975). Its confidence intervals may be found in table 2 of Tajima's paper (Tajima 1989*a*) for different sample sizes. The distribution of each statistic under each condition was obtained from 500 simulations. In the following, the estimators will be differentiated from the parameters by a hat sign (e.g., $\hat{D}$), and the averages will be indicated by a bar above the estimators (e.g., $\bar{\hat{D}}$).

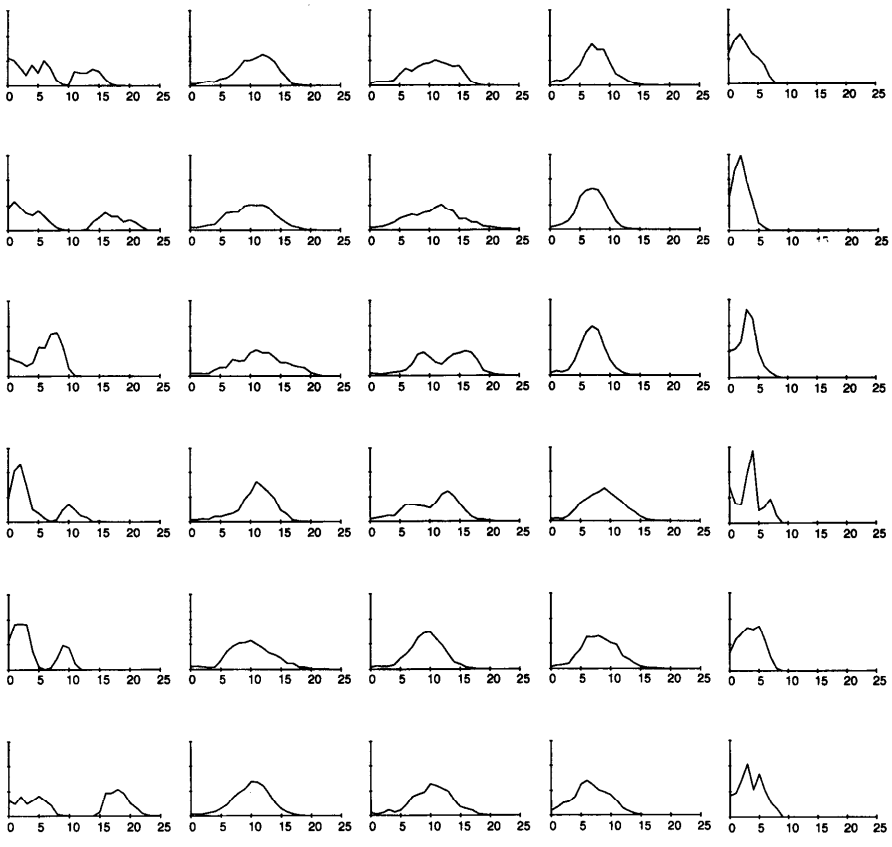
Results

Heterogeneity of Mutation Rates

Hit probability distributions are shown in figure 1 for the uniform model and for different amounts of mutation rate heterogeneity. With an α value of 0.4, slightly more than 200 out of 300 sites have a non-negligible probability of being the target of a mutation, whereas for lower α values of 0.1 and 0.01 the numbers of potential polymorphic sites are considerably smaller and are approximately equal to 100 and 20, respectively. In the latter case, it means that all mutations will practically occur on less than 10% of the sites. Among these only two or three sites will be hit most of the time and thus act like mutational hot spots. Note that the sites are arranged on the x axis in figure 1 by order of increasing conditional probability. This will not affect our interpretations since no population statistic depends on site locations in the 300-bp sequence.

Distribution of Observed Pairwise Differences

Six random examples of distributions of observed pairwise differences are presented on figure 2 for different combinations of population expansion and mutation rate heterogeneities. Each distribution represents a single outcome of a simulated evolutionary process. Figure 2 confirms that the uniform model of mutations in stationary population may lead to very different distribution shapes (Slatkin and Hudson 1991), whereas un-



Uniform rate, stationarity	Uniform rate, 100-fold expansion	α=0.1, 10-fold expansion	α=0.1, 100-fold expansion	α=0.01, stationarity
θ=10	*θ=100 ; t=0.1*	*θ=100 ; t=0.1*	*θ=100 ; t=0.1*	*θ=10*

FIG. 2.—Distributions of observed pairwise differences obtained under different simulation conditions. Each curve represents a single evolutionary outcome. Six random outcomes are presented for each simulation condition. The x-axis represents the number of pairwise differences and ranges from 0 to 25. The empirical probability of observing a pair of sequences differing by a given number is shown on the y-axis. Its range varies between 0 and 0.3.

even mutation rates or population expansions lead to predominantly unimodal distributions. Heterogeneity of mutation rates appears to have basically the same effect on the observed pairwise difference distribution as a sudden population expansion, by shifting the mode of the distribution toward lower values and reducing the variance. Moreover, their effect is synergetic when acting together. It thus appears difficult to infer correctly the demography of a population from the distribution of pairwise differences alone without taking into account a possible heterogeneity of mutation rates. For instance, a sudden 100-fold expansion leads to distributions having essentially the same shape and the same mode as those resulting from a mere 10-fold expansion with a mutation rate heterogeneity defined by an alpha value of 0.1 (fig. 2).

Mean Number of Nucleotide Differences between Pairs of Sequences (π)

The averages of 500 simulated distributions of observed mean pairwise differences ($\bar{\hat{\pi}}$) are reported in table 1, and compared to the present generation mutation parameter θ for different simulation conditions. Under the infinite-site model and population stationarity, $E(\pi)$ is equal to θ (Tajima 1989). This relationship approximately holds for the stationary finite-site model in our simulations ($\bar{\hat{\pi}}/\theta = 0.932$, $\bar{\hat{\pi}} = 9.32$, SD $[\hat{\pi}] = 13.19$), but with increasing mutation rate heterogeneity, the mode and the mean of the distributions move toward much lower values (table 1 and fig. 2). The same trend may be observed in all cases when a population has been expanding. The present day population parameter

Table 1
Average Sequence Diversity ($\bar{\hat{\pi}}$) in a Sample of 100 Sequences under Different Simulation Conditions

Demographic Model	Finite-Site Model	Mutation Rate Heterogeneity			
		$\alpha = 0.4$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1. Stationarity:					
$\bar{\hat{\pi}}$	9.324	8.277	6.536	5.473	2.913
c.v.	1.415	1.348	1.293	1.264	1.216
$\bar{\hat{\pi}}/\theta$ ($\theta = 10$)	0.932	0.828	0.654	0.547	0.291
2. Population Expansion ($t = 0.1M$ generations ago):					
×2:					
$\bar{\hat{\pi}}$	10.017	8.943	7.043	5.938	3.071
c.v.	1.351	1.306	1.249	1.213	1.173
$\bar{\hat{\pi}}/\theta$ ($\theta = 20$)	0.501	0.447	0.352	0.297	0.154
×5:					
$\bar{\hat{\pi}}$	13.015	11.300	9.029	7.255	3.614
c.v.	1.251	1.202	1.163	1.135	1.122
$\bar{\hat{\pi}}/\theta$ ($\theta = 50$)	0.260	0.226	0.181	0.145	0.072
×10:					
$\bar{\hat{\pi}}$	17.151	14.880	11.244	9.057	4.263
c.v.	1.162	1.122	1.092	1.087	1.090
$\bar{\hat{\pi}}/\theta$ ($\theta = 100$)	0.172	0.149	0.112	0.091	0.043
×100:					
$\bar{\hat{\pi}}$	10.126	9.376	7.914	6.677	3.501
c.v.	1.068	1.066	1.066	1.069	1.093
$\bar{\hat{\pi}}/\theta$ ($\theta = 100$)	0.101	0.094	0.079	0.067	0.035
×1,000:					
$\bar{\hat{\pi}}$	9.293	8.695	7.430	6.340	3.380
c.v.	1.066	1.066	1.067	1.071	1.096
$\bar{\hat{\pi}}/\theta$ ($\theta = 100$)	0.093	0.087	0.074	0.063	0.034

θ is thus increasingly underestimated from the distribution of observed pairwise differences for smaller α values (more uneven mutation rates). As expected, the effect of sudden population growth on $\bar{\hat{\pi}}$ is very similar under the finite site model. We see that $\bar{\hat{\pi}}$ values are shifted toward lower values for larger expansion factors, reaching a value of approximately $\pi = \theta t$ for very large expansion factors ($\geq$100) (Rogers and Harpending 1992; Slatkin and Hudson 1991).

Number of Alleles (k)

We also compared the average number of alleles obtained for the simulations to their expectation under the infinite-allele model for the present generation. Interestingly, the expected number of alleles was not found very sensitive to either population expansion or mutation rate heterogeneity, unless the latter took rather extreme values (α less than 0.05). Although $\hat{\theta}_k$ estimated from the number of alleles (Ewens 1972) can be considered as an estimator of the long-term mutation parameter θ (Ewens 1983), our simulations showed that it is essentially unbiased when population expansions have occurred as recently as $0.1M$ generations ago. It confirms that the population diversity in terms of the number of alleles is rapidly reaching its equilibrium value in large samples (Griffiths 1979).

Number of Polymorphic Sites (S)

In table 2, we show the average number of polymorphic sites obtained over 500 simulated samples ($\bar{\hat{S}}$), compared to those expected under the infinite-site model and population stationarity. As expected, $\bar{\hat{S}}$ decreases rapidly with α, because mutations are restricted to a smaller number of potentially polymorphic sites for smaller α values (see fig. 1). Even under stationarity and an even mutation process, the observed number of polymorphic sites is only 91% of its expectation under the infinite-site model. Here again, population expansion and mutation rate heterogeneity act synergistically on S. Note, however, that for the same present mutation parameter $\theta = 100$, the number of polymorphic sites decreases with increasing expansion factors. For instance, in the finite site model $\bar{\hat{S}} = 152.162$ for a 10-fold expansion, whereas it is only 134.73 after a 1,000-fold expansion. This is due to the fact that, given the same present population size, the ancestral population size

Table 2
Average Number of Polymorphic Sites ($\bar{\hat{S}}$) in a Sample of 100 Sequences under Different Simulation Conditions

Demographic Model[a]	Finite-Site Model	Mutation Rate Heterogeneity			
		$\alpha = 0.4$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1. Stationarity:					
$\bar{\hat{S}}$	47.296	39.262	28.346	21.752	9.030
c.v.	0.254	0.206	0.175	0.169	0.177
$\bar{\hat{S}}/S_\theta$*	0.914	0.758	0.548	0.420	0.174
2. Population Expansion:					
×2:					
$\bar{\hat{S}}$	61.062	49.264	34.18	25.074	9.908
c.v.	0.200	0.172	0.137	0.131	0.148
$\bar{\hat{S}}/S_\theta$*	0.590	0.476	0.330	0.242	0.096
×5:					
$\bar{\hat{S}}$	101.426	73.944	46.026	32.262	11.642
c.v.	0.392	0.286	0.178	0.106	0.045
$\bar{\hat{S}}/S_\theta$*	0.392	0.286	0.178	0.125	0.045
×10:					
$\bar{\hat{S}}$	152.162	101.244	57.238	38.984	13.208
c.v.	0.077	0.069	0.077	0.076	0.111
$\bar{\hat{S}}/S_\theta$*	0.294	0.196	0.116	0.075	0.026
×100:					
$\bar{\hat{S}}$	136.866	93.866	54.364	37.338	12.712
c.v.	0.080	0.070	0.077	0.081	0.109
$\bar{\hat{S}}/S_\theta$*	0.264	0.181	0.105	0.072	0.025
×1,000:					
$\bar{\hat{S}}$	134.73	92.680	54.280	37.138	12.852
c.v.	0.072	0.072	0.075	0.085	0.113
$\bar{\hat{S}}/S_\theta$*	0.260	0.179	0.105	0.072	0.025

[a] θ values and expansion times are those described in table 1.
* S_θ is the expected number of polymorphic sites under the infinite-site model (Watterson 1975).

was much smaller for the largest expansions. In this case, it implies that most polymorphic sites have emerged after the expansion, whereas an important fraction of the polymorphic sites did accumulate before the smallest expansions. Compared to its expectation under the infinite site model, $\hat{S}$ is, however, increasingly biased with expansion size. As shown by the coefficient of variation, the number of polymorphic sites is more constrained for larger expansions and more uneven mutation rates.

Tajima's *D*

The average *D* values ($\bar{D}$) over 500 simulations tabulated in table 3 agree with Tajima's (1989*a*, 1989*b*) suggestion that a sudden expansion leads to negative *D* values. Under the finite-site model, a 100-fold population expansion is indeed sufficient to move the observed *D* value outside the 95% confidence interval derived for a neutral locus and population stationarity. Interestingly, mutation rate heterogeneity has the opposite effect on *D*, shifting it toward more positive values for more uneven mutation rates. This effect had already been recognized by Bertorelle and Slatkin (1995) in the case of a finite-site two-rate mutation model. Note also that population expansions have opposite effects on the variance of *D*, as large population expansions lead to smaller variances, whereas more uneven mutation rates increase the variance. These contrasting effects of population expansion and rate heterogeneity have important consequences on the test of selective neutrality proposed by Tajima (1989*a*). In table 4, we show the probability of rejecting the hypothesis of selective neutrality and population stationarity under different simulation conditions. Under the finite-site model and with even mutation rates, selective neutrality is rejected at the 5% level in more than 95% of the cases for population expansions larger than 100-fold. It is, however, rejected in much less than 5% of the cases at stationarity for several combinations of small population expansions and reasonable α values, as well as for large expansions and extreme α values. This seemingly complex behavior can be understood by considering figure 3, where we have plotted the distributions of Tajima's *D* for a few cases, as well as the limits of the neutrality 95% confidence interval

Table 3
Average Tajima's $\bar{\bar{D}}$ Statistic in a Sample of 100 Sequences under Different Simulation Conditions

Demographic Model[a]	Finite-Site Model	Mutation Rate Heterogeneity			
		$\alpha = 0.4$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
1. Stationarity:					
$\bar{\bar{D}}$	−0.012	0.227	0.545	0.869	1.681
SD	0.915	0.890	0.884	0.889	0.898
2. Population Expansion:					
×2:					
$\bar{\bar{D}}$	−0.531	−0.235	0.189	0.667	1.564
SD	0.758	0.840	0.734	0.798	0.818
×5:					
$\bar{\bar{D}}$	−1.130	−0.684	0.045	0.511	1.625
SD	0.559	0.600	0.681	0.627	0.679
×10:					
$\bar{\bar{D}}$	−1.393	−0.786	0.065	0.643	1.845
SD	0.449	0.477	0.474	0.530	0.620
×100:					
$\bar{\bar{D}}$	−2.041	−1.575	−0.774	−0.214	1.171
SD	0.158	0.185	0.242	0.290	0.436
×1,000:					
$\bar{\bar{D}}$	−2.126	−1.677	−0.917	−0.344	1.002
SD	0.143	0.164	0.214	0.275	0.430

[a] θ values and expansion times are those described in table 1.

given by Tajima (1989*a*, table 2). Under stationarity and even mutation rates, the distribution is fitting well to the theoretical confidence interval. However, a 1,000-fold expansion both shifts the distribution toward large negative values, well outside the neutrality range, and considerably reduces its variance. Adding mutation rate heterogeneity will both shift the distribution rightward toward less negative values and also increase its variance. However, in the latter case and for $\alpha = 0.1$, the distribution is completely within the neutrality range. Tajima's test cannot reject the null hypothesis of neutrality and is thus overly conservative after a large population expansion combined with mutation rate heterogeneity.

Application to Human mtDNA Control Region Polymorphism

Four human population samples chosen because they had characteristics similar to those used in our simulations (large sample size and about 300 bp sequenced in a noncoding region) were examined for their polymorphism in the mtDNA control region. A summary of their sequence polymorphism is shown in table 5 and figure 4. All four population samples analyzed for the first hypervariable segment (HVS I) of the control region show negative D statistics, whereas the Mandenka sample analyzed for the second hypervariable segment (HVS II) shows a positive D statistic (table 5). All five

Table 4
Empirical Probability of Rejecting Tajima's Test of Neutrality in a Sample of 100 Sequences at the 5% Level under Different Simulation Conditions

Demographic Model	Finite-Site Model	Mutation Rate Heterogeneity			
		$\alpha = 0.4$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
Stationarity	0.030	0.020	0.046	0.096	0.310
Population Expansion:					
×2	0.018	0.014	0.008	0.048	0.266
×5	0.072	0.006	0.006	0.016	0.252
×10	0.164	0	0	0.008	0.346
×100	0.966	0.11	0	0	0.02
×1,000	0.996	0.236	0	0	0.014

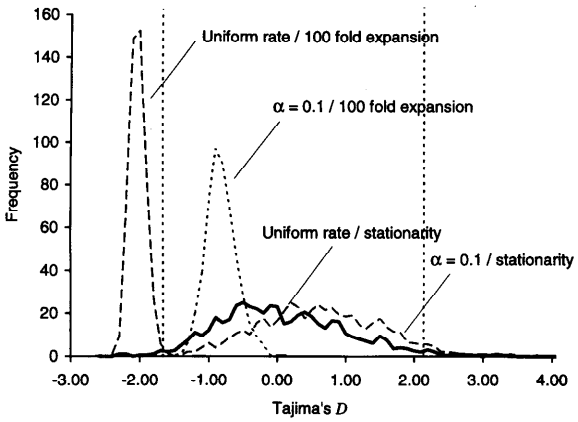


FIG. 3.—Distribution of Tajima's D under different simulation conditions.

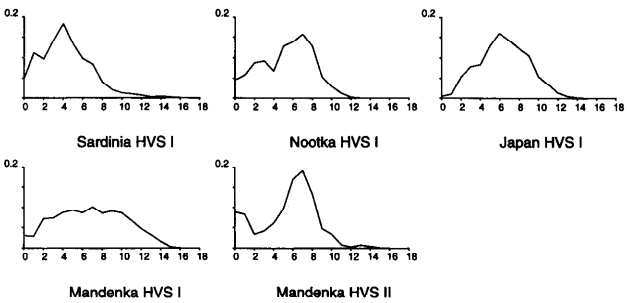


FIG. 4.—Distributions of observed mtDNA control region pairwise differences in four human populations. The x-axis represents the number of pairwise differences and ranges from 0 to 18.

samples show a major peak in their distributions of pairwise differences (fig. 4), although only the Japanese sample presents a strict unimodal distribution. The shapes of the distributions of pairwise differences were analyzed using the program "mmest" written by A. Rogers to infer the size and the age (τ) of possible demographic expansions, as well as the present day mutation parameter (θ_1) under the infinite-site model (see Rogers and Harpending 1992). Considering only HVS I, Rogers and Harpending's procedure suggests that the Sardinian, the Nootka, and the Mandenka populations have recently expanded by about a 10-fold factor, whereas the Japanese population would have undergone a much larger 1,000-fold expansion. Note that the mutation parameters estimated by Rogers and Harpending's (1992) method ($\hat{\theta}_1$) are in quite close agreement with those estimated from the mere number of alleles ($\hat{\theta}_k$), except for the Sardinians. The large expansion factor, the significant negative D value, and the unimodal distribution of pairwise differences are in good agreement with the occurrence of a sudden and large expansion in the Japanese population and a homogeneous mutation process in HVS I. The picture is less obvious for the other three populations. The Nootka and the Mandenka populations present a mild expansion size in agreement with negative, but nonsignificant, $\hat{D}$ statistics and bimodal pairwise difference distributions. The slightly negative $\hat{D}$ value and the small number of polymorphic sites in the Nootka population is compatible with the joint action of a mild expansion and mutation rate heterogeneity. For HVS I, the polymorphism of the Mandenka population seems to have been mainly affected by a mild demographic expansion rather than by high amounts of mutation rate heterogeneity. The Sardinian population presents a highly significant $\hat{D}$ value quite incompatible both with the 11-fold inferred expansion factor and with mutation rate heterogeneity. Its mean number of nucleotide differences ($\hat{\pi}$) is smaller than in other populations, although having approximately the

Table 5
Description of the Molecular Polymorphism Detected in the mtDNA Control Region of Four Human Populations

Population	Sample Size	$\hat{k}$	$\hat{S}$	$\hat{\pi}$	$\hat{\theta}_k$	$\hat{\theta}_1$	$\hat{\tau}$	$(t)^e$	Expansion Size	$\hat{D}$
HVS I:										
Japan[a]	61	53	65	6.41	189.34	227.75	6.21	(0.31)	1,172.3	−1.84*
Sardinia[b]	69	46	53	4.28	59.15	19.05	2.60	(0.13)	11.3	−2.04*
Nootka[c]	63	28	26	5.32	18.75	20.94	3.92	(0.20)	15.0	−0.11
Mandenka[d] ...	119	53	58	6.81	36.08	32.43	4.45	(0.22)	13.8	−1.17
HVS II:										
Mandenka[d] ...	119	58	27	5.48	44.02	10.36	3.61	(0.18)	5.6	0.25

NOTE.—The statistics are defined in the text.
[a] Horai and Hayasaka (1990).
[b] Di Rienzo and Wilson (1991).
[c] Ward et al. (1991).
[d] Graven et al. (1995).
[e] τ is expressed in arbitrary units of sequence mutation rate u. t is expressed in arbitrary units of M.
* $P < 0.05$.

same number of segregating sites ($\hat{S}$) (fig. 4, table 5). As S is more dependent on the present day population size and π is more strongly influenced by the size of the original population (Tajima 1989*b*), the Sardinian population could have grown to an approximately similar effective size to the Japanese and the Mandenka, but from a much smaller ancestral population. Alternatively, the expansion could have been much more recent, as suggested by Rogers and Harpending's (1992) parameters (table 4, $\hat{\tau} = 2.6$), but the effect of varying amount of time since the expansion was not studied here.

The pattern of polymorphism in Mandenka's HVS II is strikingly different from that inferred from HVS I. Although having slightly more alleles (58 vs. 53), the second segment shows less than half the segregating sites observed in HVS I (27 vs. 58). This result combined with the smaller $\hat{\pi}$ value and the positive $\hat{D}$ value suggests the occurrence of a stronger heterogeneity of mutation rates in HVS II than in HVS I.

Discussion

We have simulated a heterogeneous mutation rate process by using gamma-distributed conditional probabilities of a site being hit by a mutation (fig. 1), which are a close approximation of gamma-distributed mutation rates. The main goal of this study was to investigate the consequences of uneven mutation rates on patterns of DNA sequence polymorphisms compared to the uniform rate model and to contrast them with the effects of sudden population expansions. We have shown that both processes were leading to unimodal distributions of pairwise differences and were reducing the number of segregating sites in the sample. Although they have opposite effects on Tajima's D statistic, their respective contributions to the observed pattern of DNA polymorphism appear difficult to quantify. Only significantly negative $\hat{D}$ values can be tentatively interpreted as the signature of large expansions alone, whereas nonsignificant values (either positive or negative) are expected either in the absence or in the simultaneous presence of population expansion and rate heterogeneity (table 4, fig. 3). We therefore recommend that the use of Tajima's test, as a means to assess selective neutrality, should be restricted to loci with even mutation rates and to stationary populations.

If only a few sites can be polymorphic, Rogers and Harpending's (1992) procedure will tend on average to severely overestimate the size of the expansion and underestimate its age, as the distribution of observed pairwise differences will have a smaller mean and variance than in the case of a pure expansion model. Figure 2 indeed shows that the size of an expansion could be underestimated by an order of magnitude with a plausible α value of 0.1, such as found by Tamura and Nei (1993) for the human mtDNA control region. This results is in clear contrast with Rogers' (1992) study considering the possible effect of Gamma-distributed mutation rates on the shape of the pairwise difference distribution. Comparing the observed number of differences between pairs of sequences to the actual number of mutations having occurred since their divergence, Rogers (1992) concluded that the relative error on the modal values of the distribution would be only about 3% with an α parameter of the gamma distribution equal to 0.11. However, the effect of uneven mutation rates will be larger for nonmodal, more divergent pairs of sequences in the distribution. The relative error for a large number of differences should therefore be larger as well. In fact, as one of the main consequences of mutation rate heterogeneity is to reduce the number of possible polymorphic sites, this process will affect the mode and the shape of the distribution of observed pairwise differences whenever the polymorphic sites are saturated, which can occur in several circumstances (e.g., for a very low α value or for a moderate α value in a large population at or close to equilibrium).

Inferring past demographic events from DNA sequence polymorphism in the presence of mutation rate heterogeneity remains challenging. No single statistic reviewed here seems sufficient to reveal the respective contribution of the mutation pattern and the demographic history of a population sample. Only the Japanese sample shown in table 4 presented a pattern of polymorphism compatible with a large expansion model. The pattern of polymorphism in HVS I for the other population samples was best explained by the joint action of both mild expansions and mild mutation rate heterogeneity, whereas the pattern of polymorphism in HVS II could be attributed to strong uneven mutation rates in the Mandenka population. The observation of variable amounts of mutation rate heterogeneities between physically adjacent loci underlines the necessity to study several loci before drawing any conclusion on the history of any population. This added level of variation combines with the large stochasticity of the genealogical process (fig. 2) (Slatkin and Hudson 1991; Marjoram and Donnelly 1994). If a single cause may lead to different outcomes, one must also be aware that a given observation can be the outcome of quite different evolutionary processes as many parameters (e.g., the level of mutation rate heterogeneity, the time and the factor of population expansion, the population size, or the mutation rate) can affect the pattern of DNA polymorphism. Our present simulation study has tackled only a small amount of the vast array of different possible situations, and more realistic conditions could be envisioned. For instance, a completely stochastic muta-

tion process, where mutation rates would vary between generations, could increase the variance of our estimates of sample polymorphism. Population expansions taking the form of a more realistic exponential growth would certainly lead to results similar to ours for sufficiently large population size increases (at least two orders of magnitude, Rogers and Harpending 1992), but smoother demographic transitions of lower magnitude would certainly have less impact on the sample polymorphism than those presented here. Further studies involving analytical developments such as those initiated by Lundstrom, Tavaré, and Ward (1992) are needed to fully understand the complex process of DNA change in nonstationary populations with mutation rate heterogeneities.

Acknowledgments

We thank Giorgio Bertorelle, Monty Slatkin, Peter Smouse, and two anonymous reviewers for their comments on the manuscript, André Langaney for his continuous support throughout this work, and A. Rogers for sharing his computer program mmest. S.A.-B. was supported by a European Science Foundation grant, and L.E. by a Swiss National Science Foundation grant No. 32-37 821-93.

LITERATURE CITED

AVISE, J. C. 1994. Molecular markers, natural history and evolution. Chapman & Hall, New York.

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

BROWN, W. M., E. M. PRAGER, A. WANG, and A. C. WILSON. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. **18**:225–239.

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.

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

EWENS, W. J. 1983. The role of models in analysis of molecular genetic data with particular reference to restriction fragment data. Pp. 45–73 *in* B. S. WEIR, ed. Statistical analysis of DNA sequence data. Marcel Dekker, New York and Basel.

EXCOFFIER, L., P. E. SMOUSE, and J. M. QUATTRO. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131**:479–491.

FITCH, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**:406–416.

GRAVEN, L., G. PASSARINO, O. SEMINO, P. BOURSOT, S. SANTACHIARA-BENERECETTI, A. LANGANEY, and L. EXCOFFIER. 1995. Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Senegalese Mandenka sample. Mol. Biol. Evol. **12**:334–345.

GRIFFITHS, R. C. 1979. Exact sampling distributions from the infinite neutral alleles model. Adv. Appl. Prob. **11**:326–354.

HARPENDING, H., S. T. SHERRY, A. R. ROGERS, and M. STONEKING. 1993. The genetic structure of ancient human populations. Curr. Anthropol. **34**:483–496.

HASEGAWA, M., A. DI RIENZO, T. D. KOCHER, and A. C. WILSON. 1993. Toward a more accurate time scale for the human mitochondrial DNA tree. J. Mol. Evol. **37**:347–354.

HORAI, S., and K. HAYASAKA. 1990. Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. Am. J. Hum. Genet. **46**:828–842.

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

JOHNSON, N. I., and S. KOTZ. 1973. Discrete distributions. Houghton Mifflin, Boston.

KOCHER, T. D., and A. C. WILSON. 1991. Sequence evolution of mitochondrial DNA in humans and chimpanzees: control region and protein-coding regions. Pp. 391–413 *in* S. OSAWA and T. HONIO, eds. Evolution of life: fossils, molecules and culture. Springer Verlag, Tokyo.

LUNDSTROM, R., S. TAVARÉ, and R. H. WARD. 1992. Modeling the 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 evolution. Genetics **136**: 673–683.

PRESS, W. H., B. P. FLANNERY, S. A. TEUKOLSKY, and W. T. VETTERLING. 1992. Numerical recipes in Pascal: the art of scientific computing. Cambridge University Press, Cambridge.

ROGERS, A. R. 1992. Error introduced by the infinite sites model. Mol. Biol. Evol. **9**:1181–1184.

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.

ROGERS, A. R., and L. B. JORDE. 1995. Genetic evidence on modern human origins. Hum. Biol. **67**:1–36.

RZETSKY, A., and M. NEI. 1994. Unbiased estimates of the number of nucleotide substitutions when substitution rate varies among different sites. J. Mol. Evol. **38**:295–299.

SLATKIN, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics **139**:457–462.

SLATKIN, M., and R. R. HUDSON. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129**:555–562.

SLATKIN, M., and W. P. MADDISON. 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics **123**:603–613.

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

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

TAJIMA, F. 1989*b*. The effect of change in population size on DNA polymorphism. Genetics **123**:597–601.

TAMURA, K., and M. NEI. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10**:512–526.

UZZEL, T., and K. W. CORBIN. 1971. Fitting discrete probability distributions to evolutionary events. Science **172**:1089–1096.

VIGILANT, L., R. PENNINGTON, H. HARPENDING, T. D. KOCHER, and A. C. WILSON. 1989. Mitochondrial DNA sequences in single hairs from a southern African population. Proc. Natl. Acad. Sci. USA **86**:9350–9354.

WAKELEY, J. 1993. Substitution-rate variation among sites in hypervariable region I of human mitochondrial DNA. J. Mol. Evol. **37**:613–623.

WAKELEY, J. 1994. Substitution-rate variation among sites and the estimation of transition bias. Mol. Biol. Evol. **11**:436–442.

WARD, R. H., B. L. FRAZIER, K. DEW-JAGER, and S. PÄÄBO. 1991. Extensive mitochondrial diversity within a single Amerindian tribe. Proc. Natl. Acad. Sci. USA **88**:8720–8724.

WATTERSON, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **10**:256–276.

JULIAN P. ADAMS, reviewing editor

Accepted November 21, 1995