Allele Frequencies at Microsatellite Loci: The Stepwise Mutation Model Revisited

Ana Maria Valdes,* Montgomery Slatkin* and Nelson B. Freimer†

**Department of Integrative Biology, University of California, Berkeley California 94720, and †Department of Psychiatry, University of California, San Francisco, California 94143*

Manuscript received July 15, 1992
Accepted for publication November 24, 1992

ABSTRACT

We summarize available data on the frequencies of alleles at microsatellite loci in human populations and compare observed distributions of allele frequencies to those generated by a simulation of the stepwise mutation model. We show that observed frequency distributions at 108 loci are consistent with the results of the model under the assumption that mutations cause an increase or decrease in repeat number by one and under the condition that the product Nu, where N is the effective population size and u is the mutation rate, is larger than one. We show that the variance of the distribution of allele sizes is a useful estimator of Nu and performs much better than previously suggested estimators for the stepwise mutation model. In the data, there is no correlation between the mean and variance in allele size at a locus or between the number of alleles and mean allele size, which suggests that the mutation rate at these loci is independent of allele size.

MICROSATELLITE loci are widely dispersed in the human genome. Alleles at these loci are distinguished by different numbers of repeats of a few nucleotides. Because of the potential utility of microsatellite loci in genome mapping, there have been numerous surveys of them. In this paper we will examine distributions of allele frequencies at 108 microsatellite loci in humans to consider what population genetic processes could account for those distributions. We will show that, although observed distributions are often quite irregular, the observed patterns are consistent with a relatively simple model of the generation of new alleles. The relevant model, called the "stepwise mutation" model was introduced to population genetics in the 1970s, but we will see that the published results for this model require modification and reinterpretation before they can be used. We will begin by describing the relevant features of microsatellite loci and the particular data we examined, then review the analytic theory of the stepwise mutation model and finally present some simulation results for that model.

MICROSATELLITES

Simple sequence repeats, or microsatellite sequences, are ubiquitous in eukaryotic genomes, in particular those of mammals (LEVINSON and GUTMAN 1987a). These sequences are currently of great interest to mammalian geneticists for three reasons: (a) they are central to efforts to construct whole genome genetic maps for several species including humans; (b) instability in certain repeats has recently been implicated in the inheritance of at least three human diseases; and (c) they may be extremely useful in studies of variation at the level of populations.

Microsatellite segments consist of runs of several repeats of 2–5 nucleotides. The most thoroughly studied to date have been the CA repeats (or GT on the other strand); it has been estimated that there are at least 35,000 of these sequences in the haploid human genome (WEBER 1990), that is they occur at least every 100,000 base pairs (bp). The majority of currently identified microsatellite segments are highly polymorphic in most mammalian species. Also, it appears that microsatellite segments are relatively evenly dispersed throughout mammalian genomes. This characteristic differentiates them from minisatellite segments, also called variable number tandem repeats or VNTRs, which consist of tandem blocks of repeats of at least 20 bp, and are located predominantly in subtelomeric regions of chromosomes (ROYLE *et al.* 1988).

Because they are highly polymorphic and so well distributed, simple sequence repeats have become the mainstay of intensive international efforts to develop genetic maps of a variety of organisms, including humans. For example, DIETRICH *et al.* (1992) recently reported the development of a genetic map of the entire mouse genome consisting entirely of microsatellite markers. Because of their importance in genetic mapping, there is a rapidly expanding database about the distribution of these repeats and about the number and frequency of alleles in several populations. Attempts to characterize these segments must take into account diverse and nonrandom ways by which they

have been identified: (a) through searches of sequence databases (*e.g.*, GenBank and EMBL) for these types of repeats; (b) incidentally, through the sequencing of some fragment of DNA; or (c) by screening genomic DNA libraries by hybridization with oligonucleotide probes that are composed of the particular repeat sequence. Unique sequences are then identified that flank the repeat. Primers from the unique sequences are used to enable amplification of the repeat through the polymerase chain reaction which then permits detection of polymorphisms through gel electrophoresis.

The variability in microsatellite segments is not generally dependent on specific sequence composition of the repeats. Although initial investigations focused on the CA repeats, it is now apparent that virtually every 2-, 3- and 4-base sequence is represented among microsatellites (BECKMANN and WEBER 1992). It does appear, however, that, at least in humans, there are limitations on the numbers of repeats in polymorphic alleles. Loci with fewer than about five repeat units are almost never polymorphic. At the same time, from analysis of human CA sequences from GenBank and EMBL, only 45 of 383 had a length of 40 bp or more, with the vast majority being approximately 20 bp (by contrast, almost half of the rat sequences in the databases are 40 bp or longer).

The basis for limitation in repeat length of alleles at microsatellite sequences has become the focus of intense interest in human genetics. Recently, it has become clear that instability in repeat length of trinucleotide repeats is an important mechanism in the causation of some inherited diseases, as exemplified in the identification of the mutations that result in myotonic dystrophy (DM), spino-bulbo-muscular dystrophy, and fragile X mental retardation syndrome. In each case, individuals from the normal population possess alleles with a strict upper limit in the number of repeat units. Through unknown mechanisms, some individuals develop alleles that are beyond this limit, forming a so called premutational state, with some of their descendants developing an extreme expansion which, for DM and fragile X gene (FMR-1) can run to several thousand copies. It is of interest to separately discuss the situations in these two cases. In relation to the FMR-1 gene repeat (which has the sequence CGG), several hundred unrelated chromosomes from unaffected families have been studied in four distinct ethnic populations (FU *et al.* 1991). They identified 31 distinct alleles with variation in repeat number from 6 to 54. In all populations there is a single most common allele of 29 repeat units (about 30%), while in fragile X families one sees a premutation range from about 50–200 repeat units. There was almost no overlap with the "normal" population except that in one "normal" family there were 54–60 repeat units among members. The full mutation individuals display between 200 and 1500 repeat units. The premutation alleles are clearly meiotically unstable, in that alterations in repeat number have been seen from all offspring, including those in the "normal" family. In those who possess the unstable allele, there is considerable somatic mosaicism, indicating mitotic instability of both the pre- and full-mutation size alleles. In instability there is a much greater trend toward observable increases than decreases (although there could be a selection bias at work because families are generally identified through disease cases and one might miss many cases of premutations going down to normal range). There appears to be a meiotic stability threshold operating somewhere between 46 and 52 repeats, although the number of chromosomes assessed is still relatively small.

The CTG repeat sequence of DM is also highly polymorphic in the general population (BROOK *et al.* 1992). A repeat number of 5 is most frequent in the normal population, with common alleles in the range between 10 and 16 units, up to a maximum of 27 (in 282 normal chromosomes). No alleles have been detected with between 6 and 9 repeat units. Those who are minimally affected with DM have alleles with at least 50 repeat units. Although one sees clear expansion of the repeat over generations in affected families, with DM it has also been observed that stable transmission of abnormally large alleles (*e.g.*, 60 repeat units) can occur over several generations. Mitotic instability is also observed as evidenced by somatic mosaicism.

The mechanism for instability of repeats is unknown. For each of the disease associated triplets, allele sizes in the premutation range ($50–100n$) vary slightly across generations. It has been suggested that these changes result from polymerase slippage (CASKEY *et al.* 1992). It has also been hypothesized that the instability in the premutation alleles which leads to the extraordinary expansions observed in DM and fragile X mental retardation patients results from the presumed difficulty of replicating long GC-rich sequences (CASKEY *et al.* 1992). In this scenario, unequal rates of DNA synthesis lead to multiple incomplete single strands of complementary, triplet, reinitiated sequences. At this point, the occurrence of strand switching during replication between multiple incomplete strands (leading and lagging) could lead to dramatic alteration in the copy number of the final double stranded product. It is possible that expansion of GC-rich triplet repeats beyond 50 units may interfere with packaging of the entire segment into nucleosomes.

It is of interest that there is evidence that in each of these examples the instability that produced the premutation is an ancient event (CASKEY *et al.* 1992). In the case of DM, this evidence is provided by strong

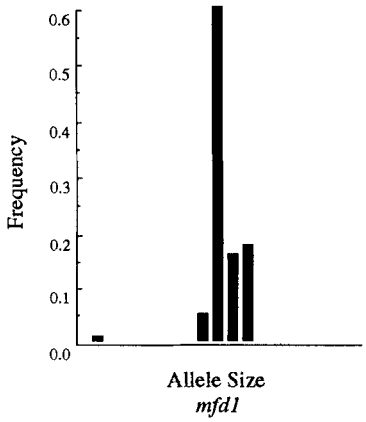
linkage disequilibrium in several ethnic populations between the DM allele and nearby DNA (BROOK *et al.* 1992). For fragile X mental retardation, there is also strong evidence for a few founder alleles, based on associations found between the disease genotype and haplotypes from two microsatellite loci that flank FMR-1 (RICHARDS *et al.* 1992). In the general population, disease-associated haplotypes are more frequently observed in individuals whose copy number for the FMR-1 repeat is at the high end of the normal range. This finding suggests that, at this locus, instability increases with allele size. However, the evidence is indirect and, as discussed previously, it is possible that deviations in either direction from the normal range enhance instability, with the expansion in repeat numbers being more frequently observed because of the fragile X mental retardation phenotype. Two explanatory models for the initiation of instability remain possible: (1) rare increases in perfect repeat copy number are sufficient and (2) mutations in sequences outside of the repeat predispose to increased instability (RICHARDS and SUTHERLAND 1992).

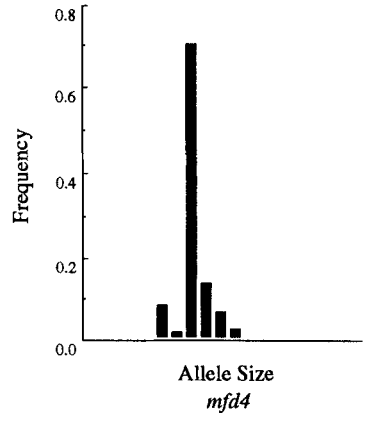
There is relatively little empirical evidence about the time of origin of microsatellite segments, in general, and about the rate and nature of mutational changes. In analyses of the GenBank and EMBL databases, it has been observed that dinucleotide repeats are conserved by chromosomal location in closely related species (*e.g.*, mouse and rat) but not in more widely diverged species (*e.g.*, humans and rodents) (STALLINGS *et al.* 1991). Although no large scale allele frequency comparisons between populations have yet been reported for microsatellites, EDWARDS *et al.* (1992) recently reported on the distribution of alleles at five loci highly polymorphic for trimeric and tetrameric repeat units in U.S. racial populations that had been defined, in part, for forensic purposes (Caucasian, Blacks, Mexican-Americans and Asians). There was considerable variation noted between populations for four of the five markers, with the greatest difference being between Blacks and the other four groups.

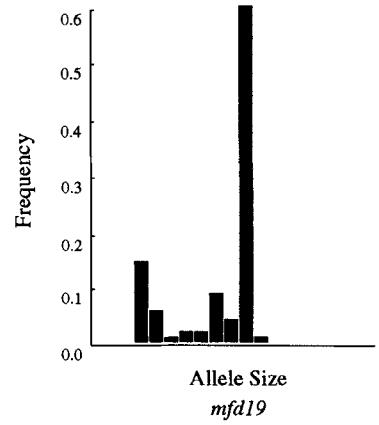
New mutations have been reported for at least four dinucleotide microsatellite markers in humans. In each case, a new allele was observed in a single individual in Caucasian pedigrees that have been widely used as standards in human genetic mapping studies (DAUSSET *et al.* 1990). HUANG *et al.* (1992) identified new alleles at the *DXS453* and *DXS454* loci, evaluating 214 and 182 independent meioses respectively. At *DXS453* the new allele was 2 bp larger than the parental alleles, and at *DXS454* it was 4 bp shorter. In a separate study of 50 chromosomes evaluated with several markers, new alleles were observed at the *D9S58* and *D9S63* loci, representing respectively a gain of 2 bp and a loss of 4 bp (KWIATKOWSKI *et al.* 1992). An absence of new alleles in surveys of tri- and tetranucleotide repeats from 860 chromosomes suggests that these repeats may have a lower mutation rate than dinucleotide segments (EDWARDS *et al.* 1992).

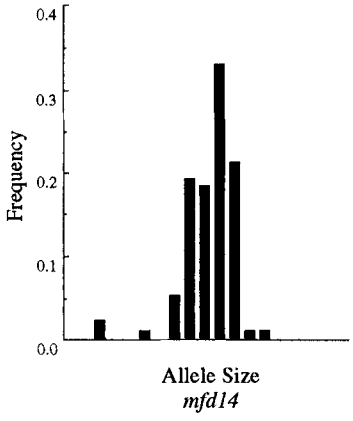
Two mechanisms have been proposed for new allele formation for repeat sequences; unequal exchange in meiosis and strand slippage in replication. For VNTRs, slippage appears to be the predominant mode of new allele formation (WOLFF *et al.* 1989), however sequence analysis of a VNTR locus with a particularly high mutation rate has also revealed mutations resulting from recombination (JEFFREYS *et al.* 1991). Slippage implies displacement of the strands of a denatured fragment followed by mispairing of complementary bases at the site of an existing short repeat sequence. As repeats gain more units they theoretically provide a more efficient substrate for slippage and therefore for further expansion (LEVINSON and GUTMAN 1987a). *In vitro* experiments using synthetic oligonucleotides and a variety of polymerases indicate that the rate of slippage is dependent on the size of the repeat unit (greatest for dinucleotides) and on its sequence (slowest for GC-rich repeats) (SCHLOTTERER and TAUTZ 1992). In these experiments the rate of slippage was independent of the length of the repeat fragment; however, experiments in prokaryotes have suggested that longer fragments show a greater rate of slippage (LEVINSON and GUTMAN 1987b). Although it has been suggested that, in mammals, longer repeat units are more polymorphic, the data remain equivocal (WEBER 1990; HUDSON *et al.* 1992).

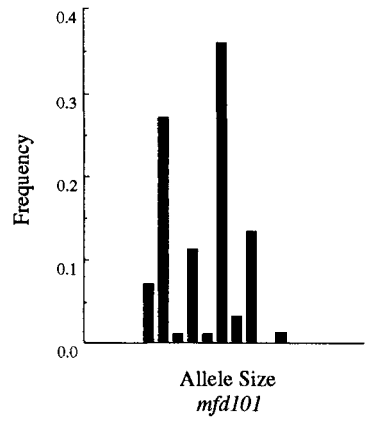
The data set: The data were obtained from the Genome Database at Johns Hopkins University, and from published information (KWIATKOWSKI *et al.* 1992; FU *et al.* 1991): 108 dinucleotide repeat markers were analyzed covering all human chromosomes except *Y*. Although a few of the sequences were identified because they reside in proximity to known coding sequences, most were identified at random in genome mapping efforts. The information on polymorphisms for these markers is derived from analysis of pedigrees drawn from Caucasian populations (mostly from France, Utah and Venezuela). Most of the genotyping data is contained in the database of the CEPH collaboration (DAUSSET *et al.* 1990). Review of the genotyping data, by family, suggested that the frequency of the common alleles does not vary appreciably between the three populations represented. The number of unrelated chromosomes analyzed to estimate allele frequencies ranges from approximately 50 to 250. The polymorphic alleles in these families have been detected based on size variation on acrylamide gels following amplification of the repeats and single copy flanking sequences using the polymerase chain reaction (WEBER and MAY 1989; LITT and LUTY 1989). The use of sequencing size standards enables detec-

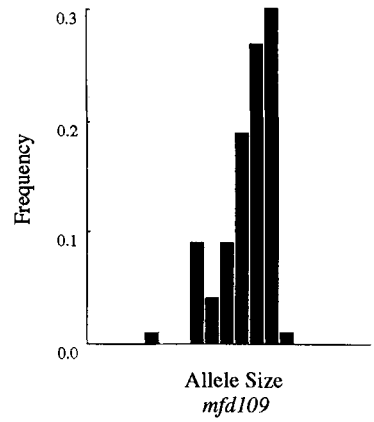

Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Allele Size
mfd1

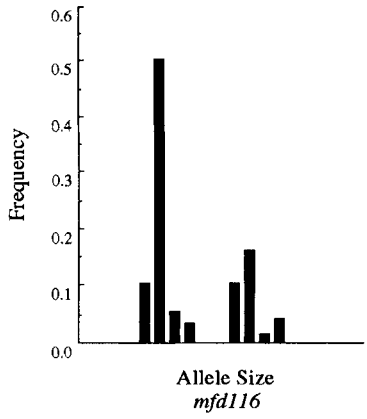

Frequency
0.0
0.2
0.4
0.6
0.8
Allele Size
mfd4

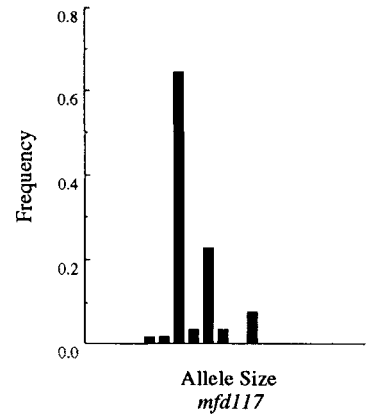

Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Allele Size
mfd19

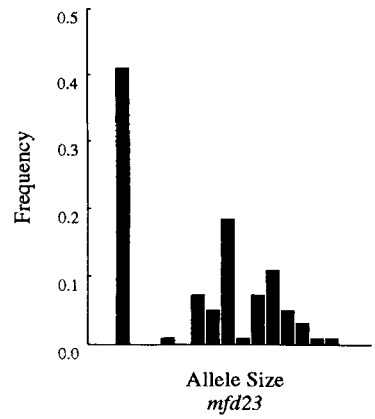

Frequency
0.0
0.1
0.2
0.3
0.4
Allele Size
mfd14

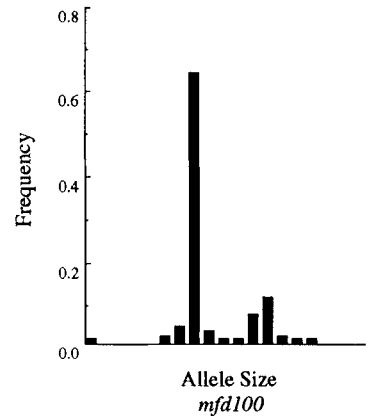

Frequency
0.0
0.1
0.2
0.3
0.4
Allele Size
mfd101


Frequency
0.0
0.1
0.2
0.3
Allele Size
mfd109


Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Allele Size
mfd116


Frequency
0.0
0.2
0.4
0.6
0.8
Allele Size
mfd117


Frequency
0.0
0.1
0.2
0.3
0.4
0.5
Allele Size
mfd23


Frequency
0.0
0.2
0.4
0.6
0.8
Allele Size
mfd100

tion of allele size at the level of a single base pair. The allele sizes reported later include flanking regions of DNA that were amplified along with the microsatellite loci themselves. Figure 1 shows 10 typical distributions of allele frequencies at 10 of these loci. The distributions are quite irregular, showing bimodal or even trimodal distributions, and there is considerable variation from locus to locus in these distributions.

STEPWISE MUTATION MODELS

Background: The stepwise mutation model was introduced to population genetics theory by OHTA and KIMURA (1973) and WEHRHAHN (1975). Both of these papers followed the suggestion of BULMER (1971) that there were regularities in the distributions of the frequencies of alleles that could be distinguished by protein electrophoresis. BULMER noted that at many loci there was one common allele with intermediate mobility and several less common alleles roughly symmetrically distributed in mobility about the common allele. BULMER suggested that these observations were consistent with a model of mutation in which alleles would increase or decrease their net charge by one unit, which would move the allele from one mobility class to another. Under this hypothesis, alleles in the same mobility class would not necessarily be identical by descent but only identical in state.

These initial papers were followed by a rapid development of the mathematical theory by MORAN (1975), KINGMAN (1977), CHAKRABORTY and NEI (1977) and others, and by the development of more refined statistical methods for parameter estimation (BROWN, MARSHALL and ALBERCHT 1975; WEIR, BROWN and MARSHALL 1978; KIMURA and OHTA 1978). Although the stepwise mutation model yielded results that were often consistent with observed distributions of allele frequencies, interest in the model declined with the discovery that adjacent electrophoretic alleles did not in general differ by a single charge state, as assumed by the stepwise model (RAMSHAW, COYNE and LEWONTIN 1979; FUERST and FERRELL 1980).

Although the stepwise mutation model is apparently not applicable to electrophoretically distinguishable alleles, it might well be applicable to alleles at microsatellite loci. The question is whether the assumption underlying the stepwise mutation model, namely that repeat number changes by only one or two as a result of mutation, is applicable to microsatellite loci. EDWARDS *et al.* (1992) have also applied some results from the stepwise mutation model to microsatellite allele frequencies. DEKA, CHAKRABORTY and FERRELL (1991) have applied the same model to VNTR allele frequencies, and ROE (1992) has extended the stepwise model to allow for more general mutation schemes and has also applied his results to VNTR data.

The model: Assume that there is a population containing N diploid individuals and consider a single autosomal locus at which alleles are distinguished by an integer, i, that indicates the number of repeat units. We assume that all alleles at this locus are selectively equivalent and that there is random mating in this population. In each generation, assume that each allele can mutate to another allelic class. In the simplest case, which we call the "one-step model," we assume that i can increase or decrease by one with probability $u/2$ per generation. We did some additional simulations and analysis using a two-step model, in which case i increases or decreases by 2 with probability $v/2$.

Let p_i be the frequency of allele i in the population. MORAN (1975) showed that p_i does not have a limiting distribution under these assumptions because the mean value of i is not constrained in any way. The distribution will "wander" on the i axis indefinitely. MORAN showed that the moments

$$C_j = E[\Sigma_i p_i p_{i+j}], \tag{1}$$

where $E[\ldots]$ denotes mathematical expectation, do have limiting distributions. At equilibrium the C_j can be expressed in the form

$$C_j = a_1\lambda_1^j + a_2\lambda_2^j, \tag{2}$$

where the a_i and λ_i are constants that depend on Nu and Nv, with the constraint that $|\lambda_i| < 1$ (BROWN, MARSHALL and ALBERCH 1975). The mean value of i is not specified but the distribution about the mean decreases smoothly and symmetrically from the maximum at $\bar{i}$, as shown in Figure 2 which obviously differs from the distributions shown in Figure 1.

Another way to interpret the C_j will help us to understand later results. The value of C_j is the probability that two genes that are randomly drawn from the population will differ by exactly j repeats. That is the distribution of differences that would be expected if we could generate a sample by drawing two genes from a population, recording the difference in repeat number and then continue drawing pairs of genes from *different* replicates of the population until a large sample were obtained. Such a sampling process is of course impossible. Instead, values of C_j are estimated

FIGURE 1.—Frequency distributions for 10 of the microsatellite loci described in the text. The most frequent allelic size is 192 bp for *mfd1*, 167 bp for *mdfd4*, 96 bp for *mfd14*, 152 bp for *mfd19*, 269 bp for *mfd101*, 83 bp for *mfd109*, 198 bp for *mfd116*, 142 for *mfd117*, 89 bp *mfd23*, and 131 for *mfd100*. These are all dinucleotide repeat loci; adjacent bars in the histograms indicate frequencies of alleles that differ in size by one repeat unit (2 bp).

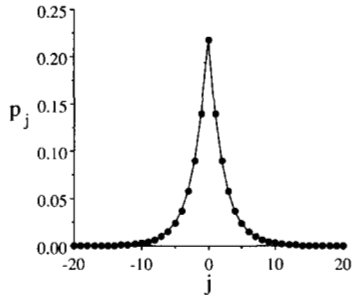


FIGURE 2.—A plot of the expected distribution of p_j under the one-step mutation model with $4Nu = 10$. This is a symmetric exponential distribution with the mean of j arbitrarily set at 0.

by making comparisons of all pairs of genes in a single sample. In a single sample, there is a correlation in the allelic states caused by gene genealogy of the sample. Our simulations show that this correlation can be sufficiently important that the values of the C_j in a sample differ substantially from the expectation given by (2). This situation is very similar to that discussed by SLATKIN and HUDSON (1991) for the distribution of pairwise differences in DNA sequences in a sample.

Estimation of parameters: Given the assumptions of the stepwise mutation model there are several ways to estimate its parameters based on various ways to fit the observed values of the C_j to the expectation given by Equation 2. We used the method of WEIR, BROWN and MARSHALL (1976) who developed a minimum χ^2 criterion. We concentrated on the one-step model, in which $a_2 = 0$ and for which the value of λ_1 is found by solving the equation

$$\lambda_1 = \frac{x(1 + 2\lambda_1 + 2\lambda_1^2 + \ldots + 2\lambda_1^{m-1})}{1 + 2\lambda_1 + 3\lambda_1^2 + \ldots + (m-1)\lambda_1^{m-1})} \quad (3)$$

where $2x = \Sigma_{j=1}^{m-1} jD_j$, D_j are estimates of the C_j in a sample and $2m + 1$ is the number of classes in the sample (WEIR, BROWN and MARSHALL 1976, pp. 646–647).

A second estimator is the variance of i in the population, σ_i^2. WEHRHAHN (1975) used generating functions to show that for two alleles sampled at random, the expectation of $(i_1 - i_2)^2$ is $4N(u + 4v)$. By writing

$$E[(i_1 - i_2)^2] = E[(i_1 - \bar{i} + \bar{i} - i_2)^2]$$
$$= E[(i_1 - \bar{i})^2] + E[(i_2 - \bar{i})^2] + 2E[(i_1 - \bar{i})(i_2 - \bar{i})]$$

and noting that the last term is zero because mutations are assumed to be independent on different lineages, we conclude that the expected variance in copy number is

$$\sigma_i^2 = 2N(u + 4v), \quad (4)$$

a result also obtained by CHAKRABORTY and NEI (1982) and by ROE (1992). Note that (4) is just the variance in the change in i in one generation, $u + 4v$, multiplied by the expected number of generations since two randomly drawn genes had their most recent common ancestor, $2N$. If $v = 0$, the one-step model, then σ_i^2 is an estimator of $2Nu$. ROE (1992) has used coalescent theory to derive higher moments of i, including the expected variance of σ_i^2.

Simulation model: To find the distribution of i in a sample of alleles from a randomly mating population, we modified the simulation program of HUDSON (1990). We assumed a panmictic diploid population of constant size N. The simulation method was based on the coalescent process for a neutral locus without recombination. As described by HUDSON (1990) each sample is obtained by first producing a genealogy of the sample under the assumption of a large constant population size. Once the genealogy is produced, mutations are randomly placed on the genealogy. The allelic state of the most recent common ancestor of all genes in the sample was set to be 0. In general the allelic state, x, of a gene (that is, the number of repeats) if its ancestor in the previous node had y repeats is given by $x = y + u_+ - u_- u_- + 2v_-$. Here u_+ and u_- are the numbers of +1 and −1 mutations and both are Poisson distributed with mean $ut/2$; v_+ and v_- are the numbers of +2 and −2 mutations and both are Poisson distributed with mean $vt/2$, where t is the branch length. This approach ensures that a stochastic equilibrium distribution of samples will be obtained and it avoids the potential problems of the simulation approach of WEIR, BROWN and MARSHALL (1976) who started their simulations with a distribution of i similar to the distribution expected from the analytic theory.

Some results from the simulation model are shown in Figure 3, which shows that even for the one-step model, quite irregular distributions of i can be obtained when $4Nu = 10$. These results are similar to those of SLATKIN and HUDSON (1991) (Figure 3) and have the same explanation. Some gene genealogies will be roughly symmetric and have a deep root, with the result that the distribution of i will be bimodal. A visual comparison of the observed distributions in Figure 1 and the simulated distributions in Figure 2 shows that even the simple one step model is consistent with the observations.

Table 1 shows the results for the one-step and two-step mutation models for a range of values of $\sigma_m^2 = 2N(u + 4v)$. Our analytic results predict that the expected value of m_2, the observed variance in repeat number, will be equal to σ_m^2. The accumulation of variance on each branch can be thought of as a random walk process on each branch. If that random walk process could be approximated by a Brownian motion process, the fourth central moment, m_4, would depend only on the value of σ_m^2 and not on u and v separately. Table 1 shows that not quite to be the case, which tells us that the two-step model is not equivalent to a Brownian motion process for these parameter values. Nevertheless, m_4 varies only slightly

TABLE 1

Moments of allele size in the stepwise mutation model

$4\sigma^2$	$4Nu$	$4Nv$	m_1	m_2	m_4
0.5	0.5	0	−0.0083	0.240	0.408
			(0.467)[a]	(0.341)	(1.250)
1.0	1	0	−0.0158	0.476	1.188
			(0.665)	(0.534)	(3.306)
	0	0.25	−0.211	0.524	2.713
			(0.679)	(1.018)	(16.74)
2.5	2.5	0	−0.028	1.2877	8.388
			(1.047)	(1.603)	(25.89)
	0.5	0.5	0.054	1.240	9.794
			(1.057)	(1.656)	(31.86)
	0	0.625	0.021	1.266	10.257
			(1.062)	(1.884)	(42.72)
5.0	5	0	0.0336	2.5610	34.851
			(1.554)	(3.02)	(106.2)
	1	1	0.0613	2.6026	35.410
			(1.617)	(3.18)	(122.6)
	0	1.25	0.0448	2.6040	41.018
			(1.5366)	(3.267)	(186.9)
9.0	9	0	0.0916	4.4755	89.776
			(2.075)	(4.597)	(225.9)
	5	1	0.0064	4.3579	93.113
			(1.945)	(4.714)	(274.4)
	1	2	0.095	4.6111	113.577
			(2.011)	(5.255)	(419.37)
10.0	10.0	0	0.049	5.1378	130.373
			(2.187)	(6.499)	(453.95)

[a] SD values are shown in parentheses. Each value is the result of 1000 replicates (average values and standard deviations did not vary significantly from a set of 1000 replicates to the next). The number of individuals (tips of the tree) in each simulation was 100. $\sigma^2 = N(u + 4v)$, u = one-step mutation rate; v = two step mutation rate.

with u and v for a given value of ${\sigma_m}^2$ which means that our use of m_2 as an estimator of ${\sigma_m}^2$ is a reasonable first approximation for the two-step model.

Table 2 shows the results of estimating the parameter $4Nu$ from 1000 simulated data sets. For each parameter value, we used both the WEIR, BROWN and MARSHALL (1976) estimator and the expected observed variance in i, m_2. In many cases, the WEIR, BROWN and MARSHALL estimator could not be obtained because the value of λ_1 found by solving (3) was not in the interval (0, 1). Estimates of $4Nu$ in such cases were undefined. The proportion of simulated data sets in which estimates of $4Nu$ could not be obtained by this method increased dramatically with $4Nu$. For all parameter values, the estimate of $4Nu$ from the variance existed and was apparently unbiased.

APPLICATIONS

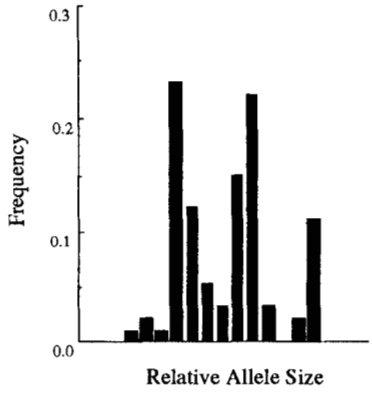
Our simulation results suggest that the observed variance of i in the distribution of allelic types at a locus is a reasonable estimator of ${\sigma_m}^2$ under the assumption of the stepwise mutation model, where ${\sigma_m}^2$ is $2N$ multiplied by the per generation variance of the change in i under mutation. In the one-step model ${\sigma_m}^2 = 2Nu$ and in the two-step model ${\sigma_m}^2 = 2N(u + 4v)$. We do not claim that this is an optimal estimator but it appears to be unbiased and defined for all parameter values we considered. An estimator that used more information about the gene genealogy would perform better but it is not clear how best to extract that information from available data.

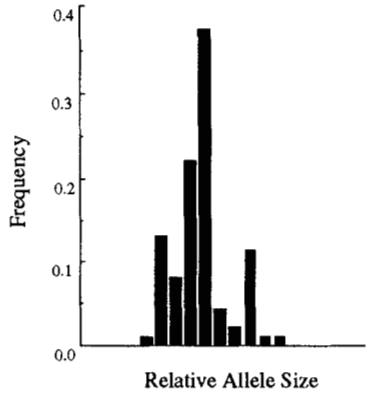
A list of the markers used and the values of m_2 and m_4 is presented in the APPENDIX. The distribution of values of m_2 is presented in Figure 4A. For comparison, we also show the distribution of m_2 from computer simulations using $4Nu = 9.94$ (one-step mutation only) from 1000 replicates (Figure 4B) and from two runs of 100 replicates (Figure 4, C and D). These runs are shown simply to illustrate that the difference between distribution in Figure 4, A and B, can be due simply to the different number of loci used in each. The results from the human markers are remarkably similar to simulation results and are consistent with a one-step mutation model with $4Nu \simeq 10$. This has several implications since the simulations assume that all loci evolve independently under the same mutation parameter in a constant size panmictic population. It is not clear why this would be the case with human markers, although it is interesting that such a simple model can account for the observed patterns.

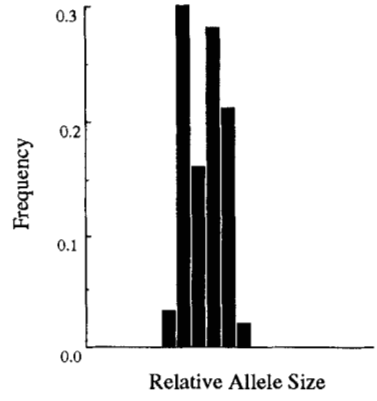
We also considered the relationship between the number of alleles at each locus and the variance in size. Figure 5B shows the observed pattern in the data and Figure 5A shows the results of 100 replicates of the simulation of the one-step model with $4Nu = 9.94$, the same parameter value used in Figure 4.

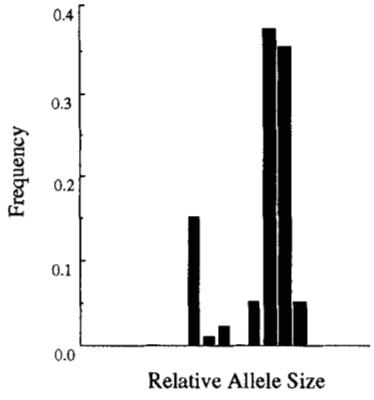
For four of the markers–*mfd66*, *mfd72*, *d9s58* and *d9s63*–empirical estimates of mutation rate are available in the literature (HUANG *et al.* 1992; KWIATKOWSKI *et al.* 1992). The mutation rates per generation reported are 0.00468, 0.00549, 0.02 and 0.02, respectively, although it should be noted that these estimates were based on incidental observations. Interestingly, the two loci with lower estimated mutation rates (*mfd66* and *mfd72*) have lower observed values of m_2 than do the two loci with higher rates (*d9s58* and *d9s63*) (see APPENDIX). If we believe the assumptions behind the model and our variance estimator for $4Nu$ the corresponding estimates of N using the empirical mutation rates are 287.9, 143.8, 503.5 and 429.5. Except for the fact that all of these estimates are of the same order of magnitude there is not much that can be said. The data appear to be consistent with the stepwise mutation model and even with the one-step model, although we do not have a rigorous test of consistency.

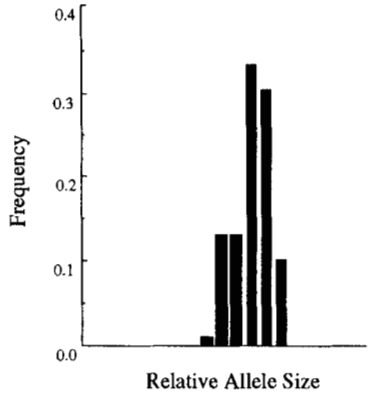
We can use our theoretical results to determine whether there is a correlation between m_2 and average allele size. WEBER (1990) did find evidence of higher

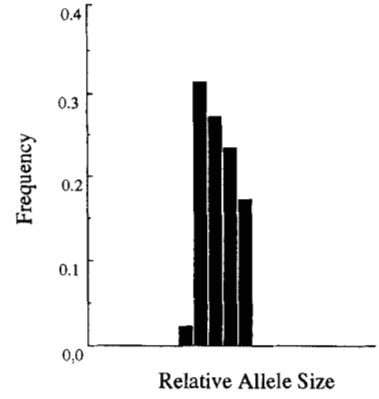

Frequency
0.0
0.1
0.2
0.3
Relative Allele Size

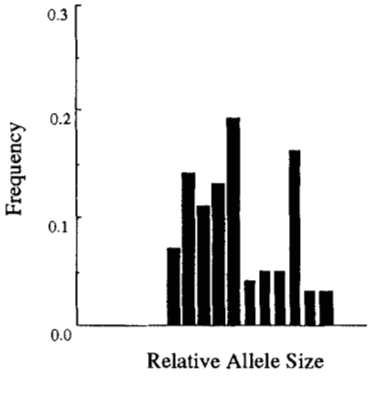

Frequency
0.0
0.1
0.2
0.3
0.4
Relative Allele Size

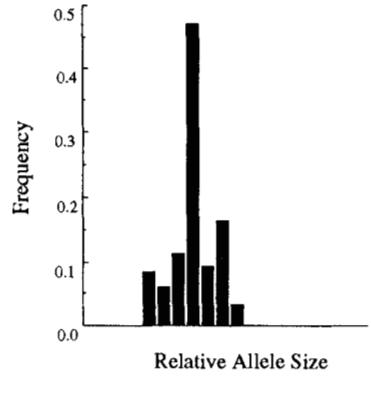

Frequency
0.0
0.1
0.2
0.3
Relative Allele Size

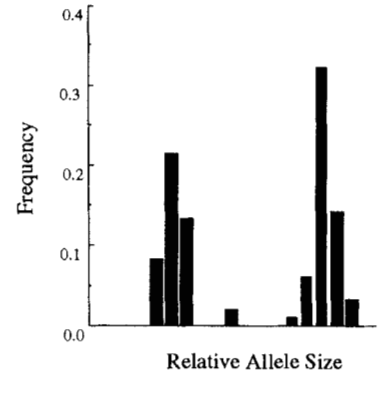

Frequency
0.0
0.1
0.2
0.3
0.4
Relative Allele Size


Frequency
0.0
0.1
0.2
0.3
0.4
Relative Allele Size


Frequency
0,0
0.1
0.2
0.3
0.4
Relative Allele Size


Frequency
0.0
0.1
0.2
0.3
Relative Allele Size


Frequency
0.0
0.1
0.2
0.3
0.4
0.5
Relative Allele Size


Frequency
0.0
0.1
0.2
0.3
0.4
Relative Allele Size

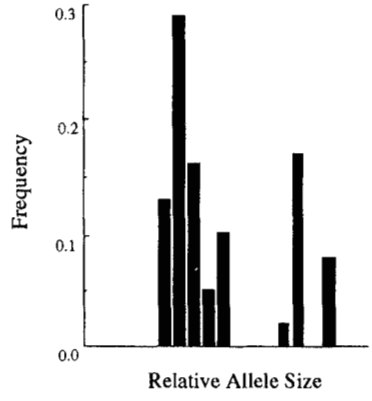

Frequency
0.0
0.1
0.2
0.3
Relative Allele Size

TABLE 2

Comparison of WEIR, BROWN and MARSHALL (1976) $4Nu$ estimator and the variance estimator

$4Nu$	WEIR, BROWN and MARSHALL (1976)	Percent time that $0 \leqq \lambda_1 \leqq 1$	$2m_2$ when $0 \leqq \lambda_1 \leqq 1$	$2m_2$ when $\lambda_1 < 0, \lambda_1 > 1$	$2m_2$ overall[a]
0.50	0.35733 (1.096)[b]	99.00	0.43792	4.35396	0.4811
2.00	1.16197 (5.38)	83.80	1.2171	5.5907	1.96998
5.00	1.05167 (0.7376)	45.60	1.8420	7.5832	4.96312
10.00	1.07464 (0.488)	12.70	2.17792	11.4496	10.2756

[a] The number of individuals sampled was 100, the number of replicates was 1000.

[b] Standard deviation. The standard deviation for the variance estimator is not shown and is twice the SD of $m_2 + \lambda_1$ in WEIR, BROWN and MARSHALL (1976, p. 647).

mutation rates in loci with above average repeat numbers. Figure 6A shows that there is no correlation between average allele size and m_2 in the 108 alleles in our study. Because m_2 estimates σ_m^2, we can conclude that there is no evidence for a correlation between mutation rate and size in these loci. Our data are, however, only for allele sizes and not repeat number directly, so there is an additional source of variation in the data, namely variation in the length of flanking sequences for each loci. Typically, the total length of the flanking sequence at a locus is about half of the total allele size and is probably the same for every allele at a locus. Variation in the lengths of flanking sequences at different loci adds some noise to the horizontal axis in Figure 6A. However, that variation would not invalidate our conclusion unless there were a systematic bias in the lengths of flanking sequences that acted to obscure a trend that was otherwise present in the data. For example, if there were actually a positive correlation between the mean and variance in repeat number, that would be not be detectable if loci with higher than average repeat number also had shorter than average flanking sequences.

We also found that there is no correlation between allele number and average allele size (Figure 6B). KIMURA and OHTA (1975) showed that allele number increases with $4Nu$ in the one-step model, although R. CHAKRABORTY (personal communication) has simulation results that show that the KIMURA and OHTA formula for allele number is not accurate for value of $4Nu$ greater than one. Nevertheless, allele number is

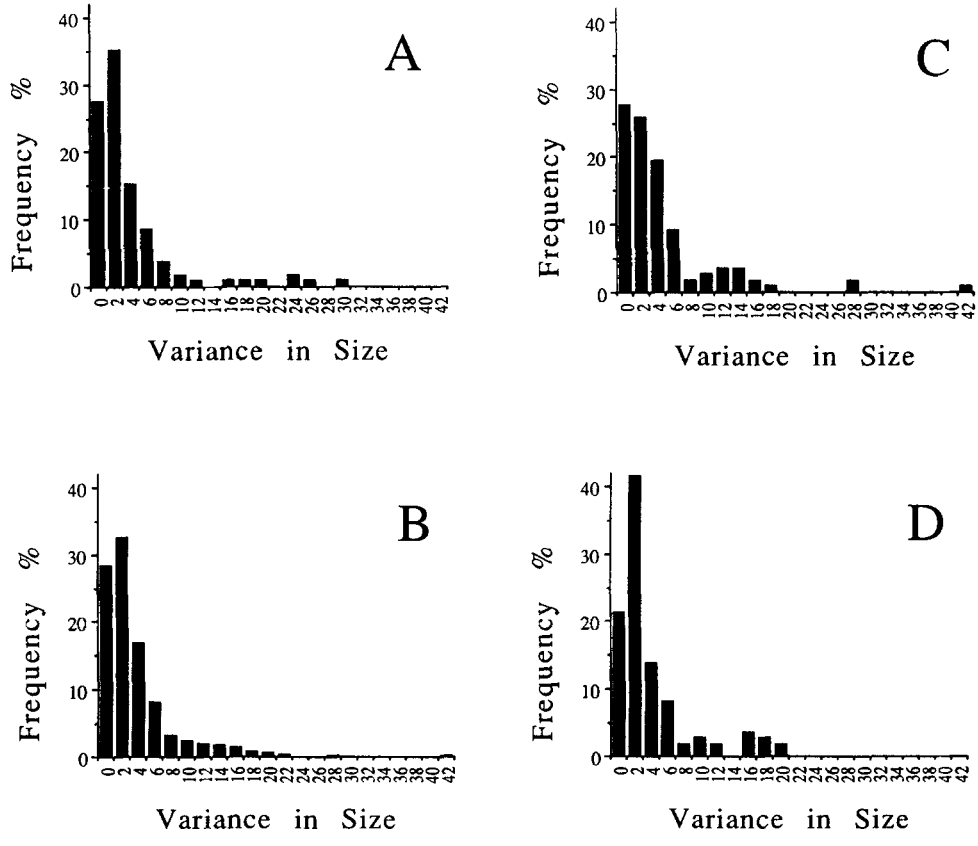


FIGURE 4.—(A) The distribution of m_2 in in the 108 loci listed in the APPENDIX. The average estimate of $4Nu$ for these data is 9.94. (B) The distribution of values of m_2 in 1000 replicate simulations of the one-step mutation model with $4Nu = 9.94$ and a sample size of 100. (C and D) Two distributions of m_2 in samples of 100 copies of the locus in two sets of 100 replicates of the one-step mutation model with $4Nu = 9.94$.

FIGURE 3.—Frequency distributions for 10 independent replicates of the one-step model with $4Nu = 10$ and a sample size of 100. The relative allele sizes are all centered on 0. Adjacent bars on the histograms indicate the frequencies of alleles that differ by one repeat unit.

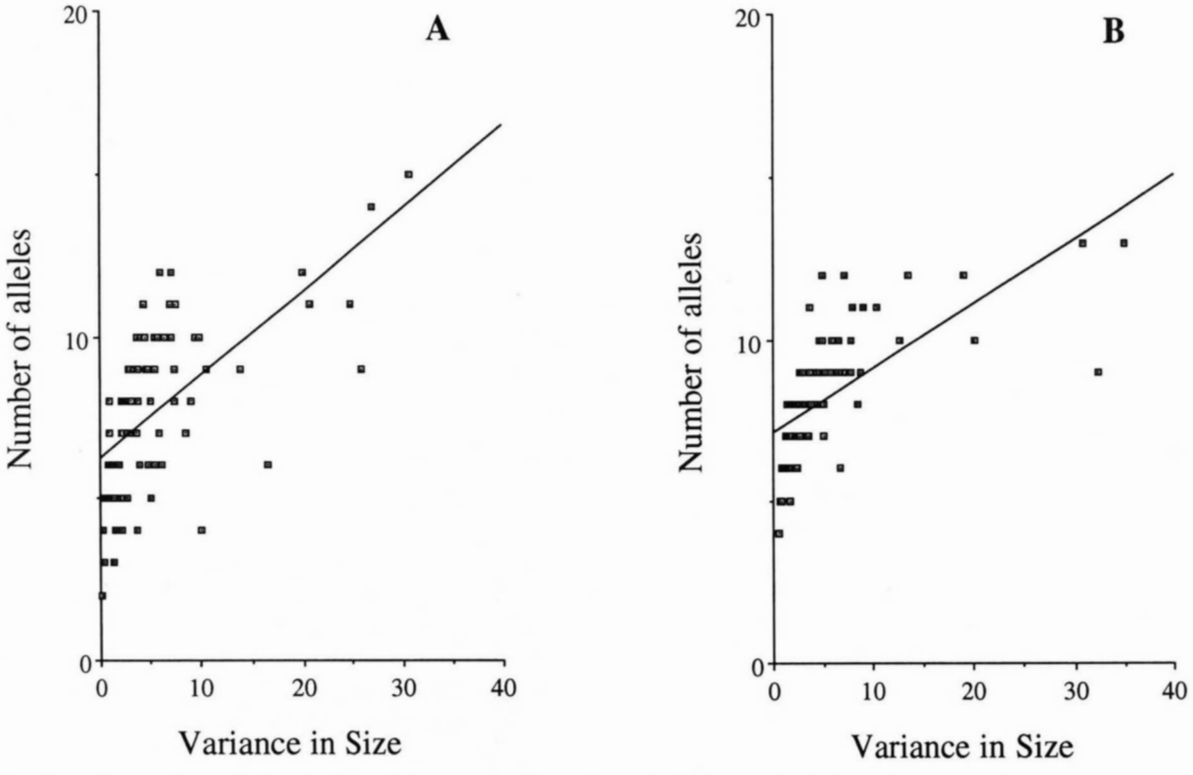


FIGURE 5.—The numbers of distinct alleles plotted against the variance in allele size. (A) Values obtained in 100 replicate simulations with $4Nu = 9.94$. (B) Observed values in the 108 loci listed in the APPENDIX.

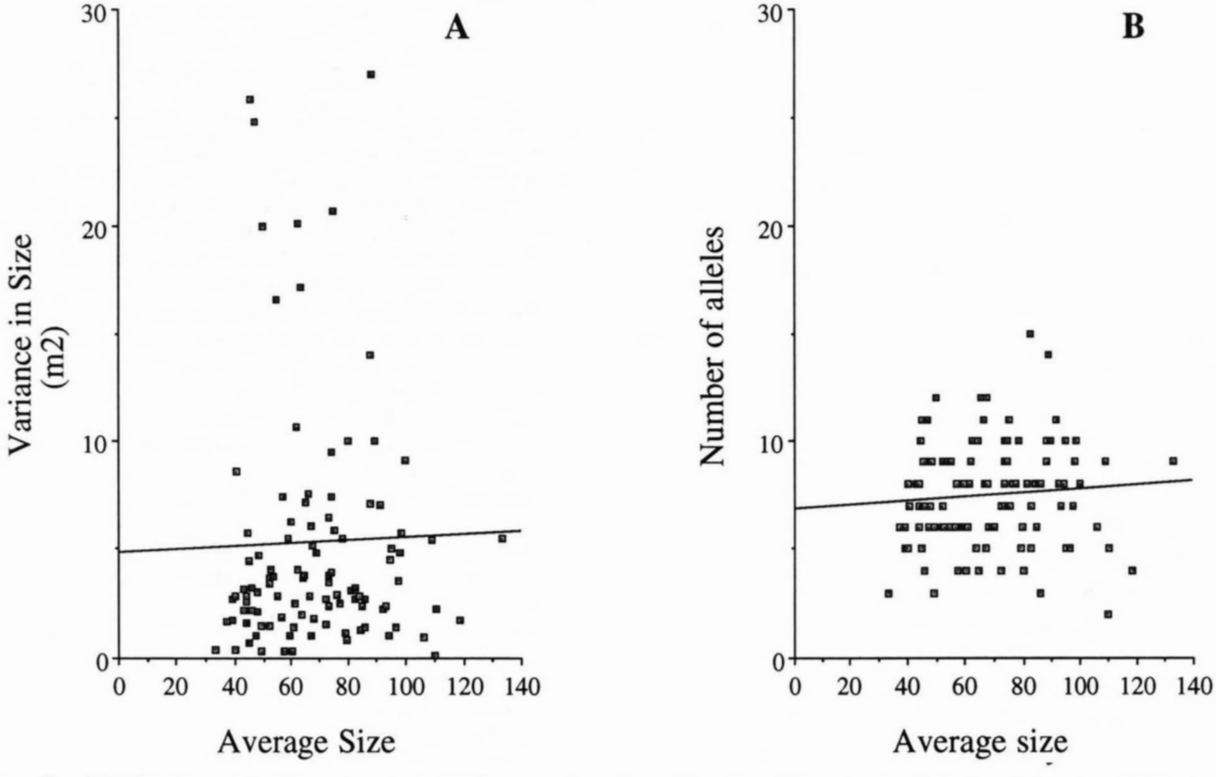


FIGURE 6.—(A) The observed variance in repeat number, m_2, plotted as a function of the average allele size of the same 108 loci plotted in Figure 4A. (B) The observed number of alleles plotted against average allele size in the 108 loci.

TABLE 3

Fragile X gene: moments of size for different populations

Population	m_1	m_2	m_4
Asian (19)	30.775	13.117	1,705.82
Hispanic (65)	28.154	19.484	1,265.32
Black (136)	28.309	13.9756	1,009.65
Caucasian (100)	27.630	34.813	5,168.20
Total ethnic (310)	27.915	21.193	2,433.98
Premutations (57)	88.215	605.582	3,431,906.87
Normals (492)	28.471	23.0378	3,536.53

still an increasing function of $4Nu$. Therefore, Figure 6B also supports the idea that there is no relationship between average allele size and mutation rate, although the same qualifications concerning our estimate of average allele size must be borne in mind.

We also analyzed the distributions of allele sizes in the fragile X gene using the data of FU *et al.* (1991), which was kindly provided by the authors of that paper. We found that the average allele size was approximately the same in all of the groups analyzed (Table 3), except for the premutation stage group which exhibited an average size more than twice that of the other groups. The variance in size showed more variation among all normal groups than did the mean and was 200–300 times larger in the premutation group than in the normals. All of the results are consistent with what has been reported experimentally for this system.

DISCUSSION AND CONCLUSIONS

Our results show that observations of allele frequencies at 108 microsatellite loci are consistent with the stepwise model of mutation at those loci in a population of constant size. This conclusion is similar to that of EDWARDS *et al.* (1992). They compared the observed distribution of allele frequencies at six loci with the expectation under the stepwise mutation model and concluded that the predictions of the stepwise mutation model were somewhat closer to their observations than predictions of the infinite alleles mutation model. Our simulations indicate that the stepwise mutation model is able to account for patterns in allele frequencies at microsatellite loci found in human populations, including the premutation alleles in the fragile X locus.

Our results do not allow us to conclude that the stepwise mutation model is a correct description of the mutation process at these loci. We know relatively little about the individuals represented in the sample but we can be reasonably certain that they do not represent a random sample from a single panmictic population of constant size. Nevertheless, because the samples are from different geographic areas, it is likely that there is some structure to the gene genealogies and it is this structure that could result in the different patterns found in the distributions of allele frequencies shown in Figure 1.

Much more information will be needed before firm conclusions about the mutation process can be drawn from population level data. Data from single relatively isolated populations could be related more easily to the kind of model we have analyzed here and theoretical results for populations with a history of growth and subdivision will be needed to model real populations more accurately.

Our results show that previously developed methods for estimating parameters of the one-step and two-step mutation models are not adequate when σ_m^2, the net increase in variance caused by mutation, exceeds one. Methods, such as that developed by WEIR, BROWN and MARSHALL (1976), that depend on the fit to the expected distribution of pairwise differences, do not take adequate account of correlations in state caused by the gene genealogy. For high mutation rates, these correlations cause the observed distributions to differ substantially from the expected distribution, even when the assumptions of the model are satisfied.

This study has been supported in part by grant from the National Institutes of Health to M.S. and grants from the National Institutes of Mental Health and the Lucille Markey Trust to N.B.F. We thank A. DI RIENZO for helpful comments during the course of this work, D. L. NELSON for sending us data on the fragile X locus, and R. CHAKRABORTY, A. G. CLARK, M. CUMMINGS, P. DONNELLY, J. C. GARZA, S. P. OTTO, J. WAKELEY, and an anonymous reviewer for helpful comments on an earlier version of this paper.

LITERATURE CITED

BECKMANN, J. S., and J. L. WEBER 1992 Survey of human and rat microsatellites. Genomics **12:** 627–631.

BROOK J. D., M. MCCURRACH, H. G. HARLEY, A. J. BUCKLER, D. CHURCH, H. ABURATANI, K. HUNTER, V. P. STANTON, J.-P. THIRION, T. HUDSON, R. SOHN, B. ZEMELMAN, R. G. SNELL, S. A. RUNDLE, S. CROW, J. DAVIES, P. SHELBOURNE, J. BUXTON, C. JONES, V. JUVONEN, K. JOHNSON, P. S. HARPER, D. J. SHAW and D. E. HOUSMAN, 1992 Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. Cell **68:** 799–808.

BROWN, A. H. D., D. R. MARSHALL and L. ALBERCH, 1975 Profiles of electrophoretic alleles in natural populations. Genet. Res. **25:** 137–143.

BULMER, M. G., 1971 Protein polymorphism. Nature **234:** 410–411.

CASKEY C. T., A. PIZZUTI, Y.-H. FU, R. G. FENWICK, JR., and D. L. NELSON, 1992 Triplet repeat mutations in human disease. Science **256:** 784–789.

CHAKRABORTY, R., and M. NEI, 1977 Bottlenect effect on average heterozygosity and genetic distance with the stepwise mutation model. Evolution **31:** 347–356.

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

DAUSSET, J., H. CANN, D. COHEN, M. LATHROP, J. M. LALOUEL and R. WHITE, 1990 Centre d'etude du polymorphisme humain

(CEPH): collaborative genetic mapping of the human genome. Genomics **6:** 575–577.

DEKA, R., R. CHAKRABORTY and R. E. FERRELL, 1991 A population genetic study of six VNTR loci in three ethnically defined populations. Genomics **11:** 83–92.

DIETRICH, W., H. KATZ, S. E. LINCOLN, H.-S. SHIN, J. FRIEDMAN, N. C. DRACOPOLI and E. S. LANDER, 1992 A genetic map of the mouse suitable for intraspecific crosses. Genetics **131:** 423–447.

EDWARDS, A., H. A. HAMMOND, L. JIN, C. T. CASKEY and R. CHAKRABORTY, 1992 Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics **12:** 241–253.

FU, Y.-H., D. P. A. KUHL, A. PIZZUTI, M. PIERETTI, J. S. SUTCLIFFE, S. RICHARDS, A. VERKERK, J. HOLDEN, R. FENWICK, JR., S. T. WARREN, B. OOSTRA, D. L. NELSON and C. T. CASKEY, 1991 Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. Cell **67:** 1047–1058.

FUERST, P. A., and R. E. FERRELL, 1980 The stepwise mutation model: an experimental evaluation utilizing hemoglobin variants. Genetics **94:** 185–201.

HUANG T., R. COTTINGHAM, JR., D. LEDBETTER and H. ZOGHBI, 1992 Genetic mapping of four dinucleotide repeat loci, DXS453, DXS458, DXS454, and DXS424 on the X chromosome using multiplex polymerase chain reaction. Genomics **13:** 375–380.

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

HUDSON, T. J., M. ENGELSTEIN, M. K. LEE, E. C. HO, M. J. RUBENFELD, C. P. ADAMS, D. E. HOUSMAN and N. C. DRACOPOLI, 1992 Isolation and chromosomal assignment of 100 highly informative human simple sequence repeat polymorphisms. Genomics **13:** 622–629.

JEFFREYS, A. J., A. MACLEOD, K. TAMAKI, D. L. NEIL and D. G. MONCKTON, 1991 Minisatellite repeat coding as a digital approach to DNA typing. Nature **354:** 204–209.

KIMURA, M., and T. OHTA, 1975 Distribution of allele frequencies in a finite population under stepwise production of neutral alleles. Proc. Natl. Acad. Sci. USA **72:** 2761–2764.

KIMURA, M., and T. OHTA, 1978 Stepwise mutation model and distribution of allelic frequencies in a finite population. Proc. Natl. Acad. Sci. USA **75:** 2868–2872.

KINGMAN, J. F. C., 1977 A note on multi-dimensional models of neutral mutation. Theor. Popul. Biol. **11:** 285–290.

KWIATKOWSKI D., E. HENSKE, K. WEIMER, L. OZELIUS, J. GUSELLA and J. HAINES, 1992 Construction of a GT polymorphism map of human 9q. Genomics **12:** 229–240

LEVINSON G., and G. A. GUTMAN, 1987a Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4:** 203–221.

LEVINSON G., and G. A. GUTMAN, 1987b High frequency of short frameshifts in poly-CA/GT tendem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. Nucleic Acid Res. **15:** 5322–5338.

LITT M., and J. A. LUTY, 1989 A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am. J. Hum. Genet. **44:** 397–401.

MORAN, P. A. P., 1975 Wandering distributions and the electrophoretic profile. Theor. Popul. Biol. **8:** 318–330.

OHTA, T., and M. KIMURA, 1973 The model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population. Genet. Res. **22:** 201–204.

RAMSHAW, J. A. M., J. A. COYNE and R. C. LEWONTIN, 1979 The sensitivity of gel electrophoresis as a detector of genetic variation. Genetics **93:** 1019–1037.

RICHARDS R. I., and G. R. SUTHERLAND, 1992 Dynamic mutations: a new class of mutations causing human disease. Cell **70:** 709–712.

RICHARDS, R. I., K. HOLMAN, K. FRIEND, E. KREMER, D. HILLEN, A. STAPLES, W. T. BROWN, P. GOONEWARDENA, J. TARLETON, C. SCHWARTZ and G. R. SUTHERLAND, 1992 Evidence of founder chromosome in fragile X. Nature Genet. **1:** 257–260.

ROE, A., 1992 Correlations and interactions in random walks and population genetics. Ph.D. Thesis, University of London, London, U.K.

ROYLE, N. J., R. E. CLARKSON, Z. WONG and A. J. JEFFREYS, 1988 Clustering of hypervariable minisatellites in the proterminal regions of human autosomes. Genomics **3:** 352–360.

SCHLOTTERER C., and D. TAUTZ, 1992 Slippage synthesis of simple sequence DNA. Nucleic Acids Res. **20:** 211–215.

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

STALLINGS R. L., A. F. FORD, D. NELSON, D. C. TORNEY, C. E. HILDEBRAND and R. K. MOYZIS, 1991 Evolution and distribution of $(GT)_n$ repetitive sequences in human genomes. Genomics **10:** 807–815.

WEBER J. L., 1990 Informativeness of human $(dC\text{-}dA)_n.(dG\text{-}dT)_n$ polymorphisms. Genomics **7:** 517–524.

WEBER J. L., and P. E. MAY, 1989 Abundant classes of human DNA polymorphism which can be typed by the polymerase chain reaction. Am. J. Hum. Genet. **44:** 388–397

WEHRHAHN, C., 1975 The evolution of selectively similar electrophoretically detectable alleles in finite natural populations. Genetics **80:** 375–394.

WEIR, B. S., A. H. D. BROWN and D. R. MARSHALL, 1976 Testing for selective neutrality of electrophoretically detectable protein polymorphisms. Genetics **84:** 639–659.

WOLFF R. K., R. PLAETKE, A. J. JEFFREYS and R. WHITE, 1989 Unequal crossing over between homologous chromosomes is not the major mechanism involved in the generation of new alleles at VNTR loci. Genomics **5:** 382–384.

Communicating editor: A. G. CLARK

APPENDIX

The Marshfield markers size moments are given in Table 4.

TABLE 4

Marshfield markers size moments

Marker	m_1	m_2	m_4	Marker	m_1	m_2	m_4
mdf1	96.390	1.4179	50.9826	*mfd48*	77.115	2.5170	22.2383
mfd2	60.210	0.3259	0.7033	*mfd49*	46.810	24.8197	987.5626
mfd3	69.100	4.8600	42.8424	*mfd50*	89.475	9.9941	179.9644
mfd4	84.560	1.2764	9.4724	*mfd51*	64.470	3.7891	15.0869
mfd5	74.825	20.7197	1605.4314	*mfd52*	95.230	5.0571	112.4599
mfd6	97.450	3.5525	18.7902	*mfd57*	67.330	6.0611	53.6383
mfd7	106.290	0.9859	6.4219	*mfd55*	45.445	25.8970	1115.3253
mfd8	92.205	2.2479	27.7334	*mfd58*	57.460	0.3284	0.2641
mfd9	48.160	4.7405	39.9939	*mfd59*	91.235	7.0440	112.5906
mfd10	67.880	1.7887	6.9122	*mfd61*	64.260	3.6924	61.4489
mfd11	56.850	7.3875	71.0068	*mfd62*	49.330	1.5156	3.6460
mfd12	118.880	1.7701	4.9856	*mfd63*	42.790	2.2059	23.1807
mfd13	73.400	2.3800	44.8936	*mfd64*	44.670	4.4670	73.7799
mfd14	47.820	3.0551	87.1738	*mfd65*	79.950	10.0075	107.7029
mfd15	78.200	5.4900	88.6569	*mfd66*	82.730	2.6946	12.9385
mfd17	45.980	3.1996	15.3021	*mfd67*	80.930	3.1217	27.6805
mfd18	39.280	2.7416	10.2849	*mfd69*	45.200	2.1900	10.4229
mfd19	74.190	7.3939	135.9422	*mfd71*	49.135	0.3043	1.5701
mfd20	63.910	2.0019	8.6055	*mfd72*	72.370	1.5789	13.3659
mfd22	74.350	3.9041	37.6524	*mfd73*	40.370	8.5731	104.0462
mfd23	49.875	19.9720	603.6809	*mfd74*	87.640	7.1150	272.6326
mfd24	47.110	1.0379	3.9982	*mfd75*	87.820	13.9982	288.3314
mfd25	66.990	1.0299	3.0870	*mfd77*	45.800	2.1900	5.8629
mfd26	54.805	2.8636	20.9802	*mfd78*	79.110	1.1779	16.3308
mfd27	73.350	6.4275	187.0012	*mfd79*	93.020	2.3796	16.0161
mfd28	73.180	3.8145	32.9856	*mfd83*	110.13	0.1131	0.0747
mfd29	60.795	1.4408	5.7802	*mfd84*	52.490	4.0499	34.6904
mfd30	44.000	2.6100	45.5625	*mfd85*	47.790	2.1459	100.2912
mfd31	39.900	0.3600	0.5424	*mfd86*	38.895	1.7397	13.4264
mfd32	54.670	16.5530	563.8354	*mfd88*	75.220	5.8716	90.3345
mfd33	52.070	1.00500	4.28154	*mfd92*	61.405	2.5442	16.8260
mfd34	44.720	0.7116	4.5721	*mfd94*	74.485	9.4721	239.8718
mfd36	83.825	2.8482	16.2376	*mfd95*	62.360	4.0808	63.7481
mfd37	33.340	0.3888	0.3482	*mfd100*	65.220	7.1436	237.2714
mfd38	61.880	10.6743	246.8319	*mfd101*	133.140	5.4704	52.4802
mfd39	74.865	2.2515	7.9425	*mfd104*	88.170	27.0211	1591.0728
mfd40	85.715	1.4019	5.7327	*mfd106*	76.280	2.9352	45.4241
mfd41	79.640	0.8204	13.2144	*mfd109*	39.900	2.8200	32.9400
mfd42	56.665	1.8961	9.4077	*mfd110*	110.700	2.2800	17.6400
mfd43	70.231	1.4462	3.8384	*mfd114*	53.630	3.7131	42.2502
mfd44	37.350	1.6475	14.8239	*mfd116*	99.870	9.0929	258.9605
mfd45	44.230	5.7571	115.7893	*mfd117*	72.590	2.7350	15.7948
mfd47	73.205	3.4702	32.3059	*mfd120*	85.000	2.3700	21.2025