Simultaneous Estimation of All the Parameters of a Stepwise Mutation Model

Yun-Xin Fu and Ranajit Chakraborty

Human Genetics Center, University of Texas, Houston, Texas 77225

Manuscript received March 13, 1998
Accepted for publication May 26, 1998

ABSTRACT

Minisatellite and microsatellite are short tandemly repetitive sequences dispersed in eukaryotic genomes, many of which are highly polymorphic due to copy number variation of the repeats. Because mutation changes copy numbers of the repeat sequences in a generalized stepwise fashion, stepwise mutation models are widely used for studying the dynamics of these loci. We propose a minimum chi-square (MCS) method for simultaneous estimation of all the parameters in a stepwise mutation model and the ancestral allelic type of a sample. The MCS estimator requires knowing the mean number of alleles of a certain size in a sample, which can be estimated using Monte Carlo samples generated by a coalescent algorithm. The method is applied to samples of seven $(CA)_n$ repeat loci from eight human populations and one chimpanzee population. The estimated values of parameters suggest that there is a general tendency for microsatellite alleles to expand in size, because (1) each mutation has a slight tendency to cause size increase and (2) the mean size increase is larger than the mean size decrease for a mutation. Our estimates also suggest that most of these CA-repeat loci evolve according to multistep mutation models rather than single-step mutation models. We also introduced several quantities for measuring the quality of the estimation of ancestral allelic type, and it appears that the majority of the estimated ancestral allelic types are reasonably accurate. Implications of our analysis and potential extensions of the method are discussed.

SINCE the discovery that a large number of loci with tandemly repeated sequences in human and many eukaryote species are highly polymorphic because of copy number variation of the repeats in different individuals (Jeffreys 1985; Litt and Luty 1989; Weber and May 1989), allele size data from such loci are rapidly becoming the dominant source of genetic markers for genome mapping, forensic testing, and population studies. Loci with repeat sequences longer than 5 bp are generally referred to as minisatellite or variable number tandem repeat loci, and those with repeat sequences between 2 to 5 bp are referred to as microsatellite or short tandem repeat loci (Tautz 1993). Because mutations change the copy number of such loci in a stepwise fashion, rapid accumulation of population samples from minisatellite and microsatellite loci has resurrected the interest of the stepwise mutation model (SMM), which was popular in the 1970s.

To avoid misinterpretation when using the information from these loci, understanding the dynamics of polymorphism at minisatellite and microsatellite loci is important. It is also vital for population and evolutionary study. Important for better understanding of the evolution of such loci is the estimation of relevant population parameters. There are several parameters of a population that can affect the pattern of polymorphism in a sample. The most important is $\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate per locus per generation. θ is primarily responsible for the amount of polymorphism in a sample. Under the infinite-allele model, that is, each mutation at a locus creates a new allele in the population, θ is the only parameter for the distribution of polymorphism in a sample from a steady population. Ewens' (1972) sampling formula provides the basis for estimating θ from a single quantity—the number of alleles in a sample. However, under SMM, the pattern of polymorphism in a sample becomes more complex and depends on additional parameters, the number of which depends on the complexity of the mechanism of evolution for such a locus. Unfortunately, a sampling distribution for a SMM parallel to Ewens' (1972) has not been found, resulting in difficulty in making inference under the SMM. Since the values of parameters are necessary in interpreting the evolution of a locus under the SMM, proper estimation of parameters is critical in studying the mechanism of evolution of loci under the SMM. To date, no method is available for simultaneously estimating all the parameters of the SMM, which limits the usefulness of such loci for studying the history of a population, particularly the human population.

We develop in this article a method for simultaneous estimation of all the parameters of the SMM and the ancestral allelic type of the alleles in a sample. The estimator is a combination of the minimum chi-square estimator (MCS) and Monte Carlo simulation, taking advantage of fast coalescent algorithms. We apply the

Corresponding author: Yun-Xin Fu, Human Genetics Center, University of Texas at Houston, 6901 Bertner Ave., Houston, TX 77030. E-mail: fu@hgc.sph.uth.tmc.edu

method to the samples of dinucleotide repeats of Deka *et al.* (1995) and discuss the implications of our analyses.

STEPWISE MUTATION MODELS

Let π_{ij} be the probability that a mutation causes an allele size change from i to j. For a stable population, which is assumed throughout this article, a SMM is completely specified by the distribution π_{ij} and the mutation parameter $\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate per allele per generation. Following the introduction of the SMM by Ohta and Kimura (1973), most of the subsequent studies in the 1970s were based on single- or two-step SMMs (*e.g.*, Moran 1975; Li 1976; Weir *et al.* 1976; Chakraborty and Nei 1977). In particular, Moran (1975) showed that under a single-step mutation model, allelic frequencies do not reach a steady distribution. Consequently, later studies of SMMs have focused on various moments of allele frequencies, *e.g.*, the variance of allele sizes that have steady-state distributions (Weir *et al.* 1976; Chakraborty and Nei 1982). This tradition appears to continue in the recent surge of interest in the SMM (Shriver *et al.* 1993; Valdes *et al.* 1993; Di Rienzo *et al.* 1994; Kimmel *et al.* 1996).

Although most studies on SMMs assume either a single- or two-step SMM, there are as many SMMs as different distributions for π_{ij}. One problem is that many SMMs can result in patterns of polymorphism that are practically indistinguishable. As a result, the choice of the distribution π_{ij} is not trivial. Distributions that are sufficiently flexible and depend on few parameters, each having clear biological meaning, should be preferred. Although the method developed in this article for estimating parameters of a SMM applies to any distribution for π_{ij}, we shall consider a distribution π_{ij} that is homogeneous for a different value of i, partly because such a model has fewer parameters and partly because only the relative sizes of alleles in the samples analyzed later are known. We shall consider the following distribution:

$$\pi_{ij} = \begin{cases} \alpha(1-P)P^{j-i-1}, & j > i \\ (1-\alpha)(1-P)P^{i-j-1}, & j < i, \end{cases} \qquad (1)$$

where $0 \le \alpha \le 1$ and $0 < P < 1$. We note that because $\Sigma_{j>i}\pi_{ij} = \alpha$, the probability of a mutation increasing allele size is α, and the probability of a mutation decreasing allele size is $1 - \alpha$. In particular, $\alpha = 1$ implies that mutations always increase allele size, $\alpha = 0$ implies that mutations always decrease allele size, and $\alpha = 0.5$ implies that there is equal chance of increasing and decreasing allele size.

It follows from the distribution (1) that given the direction of size change (increasing or decreasing) the size of a change is given by geometric distribution $(1-P)P^{|i-j|-1}$. A small value of P implies that the size of a change is likely small and a large value of P means that the size of a change can be large. An example of π_{ij} is given in Figure 1. Note that the geometric distribution $(1-P)P^{|i-j|-1}$ does not have maximum size (step) for a size change. Any size of change is possible at least theoretically. Although we can consider a truncated geometric distribution that imposes a maximum size of change, doing so will introduce another parameter. We note that, although any size change is possible under a geometric distribution, the probabilities for most changes of large size are small and therefore negligible in practice. For this reason, it is simpler to consider an effective number of steps rather than to impose an absolute maximum step. We define the effective number of steps of a SMM to be the smallest integer s such that when $\pi_{ij} < \varepsilon$, $|i-j| > s$ for some threshold value ε, we shall use $\varepsilon = 10^{-3}$. For our model, s increases with P and is the largest integer that is not larger than

$$1 - \frac{3 + \log_{10}(\max\{\alpha, 1-\alpha\}) + \log_{10}(1-P)}{\log_{10}(P)}.$$

Figure 2 plots the effective number of steps against the value of P.

Given that a locus evolves according to the SMM described above, the values of three parameters θ, α, and P then determine the values of various moments of allele frequencies at equilibrium. Therefore, in general, moments computed from a sample can be used to estimate these parameters. Because moments are computed from allele frequencies of a sample, allele frequencies thus contain more information about the parameters than a set of moments do, and consequently the accuracy of estimation directly based on allele frequencies should be higher than that based on a small set of moments. However, Moran (1975) showed that for any set of initial allele frequencies, the frequency of any allele does not have a steady distribution after a sufficient number of generations. Also, because the allele frequencies of a population many generations ago are generally unknown, Moran's analysis appears to suggest that it would be futile to make an inference about the parameters of a SMM based directly on allele frequencies of a sample. A close examination of what determines the allele frequencies in a sample will change this perception.

The sequences (chromosomes) of a sample of size n can be traced back to their most recent common ancestor (MRCA), who lived on average $4N(1-1/n)$ generations ago. It is obvious from a coalescent point of view that the distribution of allele frequencies in a sample is entirely determined by the allele A possessed by the MRCA and the three parameters θ, α, and P of SMM. The allele frequencies of a population in the more distant past are irrelevant to the allele frequencies in a sample once the ancestral allele A is known or specified. This observation implies that inference on parameters θ, α, and P based on allele frequencies in a sample can be made once A is specified.

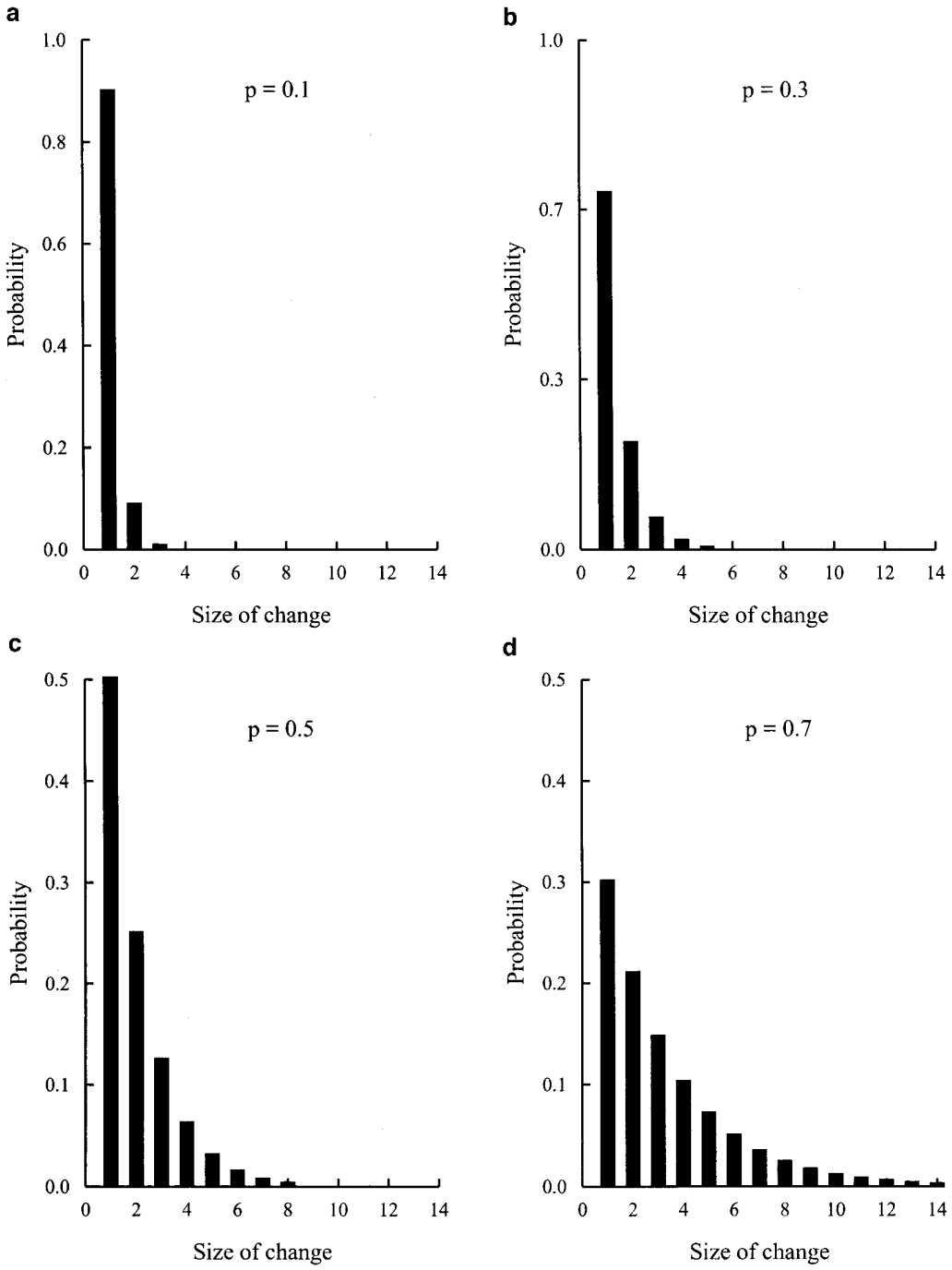

Figure 1.—Numerical examples of geometric distribution $(1 - P)P^{i-1}$.

In addition to the importance of the ancestral allele A in determining the distribution of allele frequencies in a sample, the value of A for a sample is itself of great interest in studying the evolution of the locus from which alleles were sampled. It is thus desirable to be able to infer the value of A as well as the values of θ, α, and P from a sample. Such an inference would be difficult if it were based on a set of moments only. In this article, we treat A as a parameter whose value is to be estimated from a sample. For convenience, we collect these four parameters in a vector

$$\Gamma = (\theta, P, \alpha, A). \quad (2)$$

In the next section, we describe how Γ can be estimated from a sample.

MINIMUM CHI-SQUARE ESTIMATOR OF Γ

Let f_i, $i = 1, \ldots$ be the number of alleles of size i in a sample of n chromosomes. Our aim is to derive an estimate of Γ from $\{f_i\}$.

There are two widely used approaches for estimating

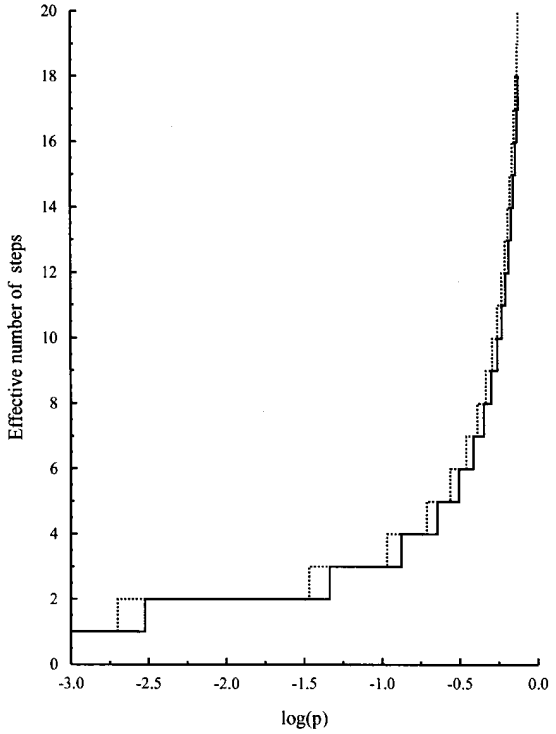

Figure 2.—The effective number of steps of a SMM. Solid and dashed lines correspond to $\alpha = 0.5$ and 0.9, respectively.

parameters: the maximum likelihood method and the least-squares method. Often both methods result in similar estimates and share many properties that are desirable. The maximum likelihood method computes the probabilities of allele frequencies $\{f_i\}$ given the value of Γ. The computation is difficult and time-consuming, but the likelihood approach has the advantage of being able to test hypotheses. On the other hand, least-squares-based methods compute the means and perhaps the variances and covariances of allele frequencies, which are much easier to compute than probabilities. Therefore, least-squares-based methods are generally easier to use in practice and are particularly appealing when a large number of samples need to be analyzed. The major disadvantage is that they are difficult to extend for hypothesis testing.

Let $e_i(\Gamma)$, $i = 1, \ldots$ be the expected number of alleles of size i conditional on the value of Γ, *i.e.*, $e_i(\Gamma) = E(f_i|\Gamma)$. Our strategy is to find the value of Γ that minimizes the quantity

$$L = \sum_{i=1}^{\infty} w_i(\Gamma)|f_i - e_i(\Gamma)|^g, \tag{3}$$

where $w_i(\Gamma)$ is a function of Γ, and $g > 0$. When $w_i = 1$ and $g = 2$, the value of Γ that minimizes L is the well-known least-squares estimator. However, since $e_i(\Gamma)$ is not linear with the parameters in Γ, the least-squares estimator is not necessarily a good choice. For a discrete distribution, a generalized least-squares method, often referred to as the MCS estimator, is usually a better one. The MCS estimator corresponds to $g = 2$ and $w_i = e_i^{-1}(\Gamma)$. Therefore, the MCS estimate of Γ is the value of Γ that minimizes the function

$$\chi^2 = \sum_{i=1}^{\infty} \frac{(f_i - e_i(\Gamma))^2}{e_i(\Gamma)}. \tag{4}$$

It is well known that the MCS estimator has similar asymptotic properties to the maximum likelihood estimator when samples are from a multinominal distribution, with each cell's probability being a function of the parameters to be estimated (*e.g.*, Stuart and Ord 1991). At the time when alleles were sampled from a population, there was a specific number of alleles of each size, and therefore the sample was indeed from a multinomial distribution. However, the probability of the number of alleles of size i is not a deterministic function of Γ, but a random variable whose distribution depends on Γ. Nevertheless, we expect the MCS estimator to be a reasonably good estimator of Γ. Note that the MCS estimator was used in Weir *et al.* (1976) for estimating parameters in a two-step SMM from moments of allele frequencies.

Although some optimization procedures with constraints can be adapted to search for the MCS estimate, a simple grid search is sufficient here because there are only four parameters. What complicates this seemingly straightforward procedure is that a formula or an even numerical solution for $e_i(\Gamma)$ is not available at present. Therefore, their values have to be estimated from simulated samples. Estimating $e_i(\Gamma)$ for a large number of combinations of parameter values can be very time consuming even when samples are generated by a fast algorithm from coalescent theory (Kingman 1982a,b; see Hudson 1991 for a recent review). When there are only a few samples to be analyzed, a two-steps grid search approach can be used. The first step is to carry out a fast full-grid search to identify the vicinity of the MCS estimate. To achieve a fast full-grid search, only a modest number of Monte Carlo samples is used to estimate $e_i(\Gamma)$ for each Γ. The second step is to carry out a fine-scale grid search in the small area identified by the first step. In the fine-scale search, a relatively large number of Monte Carlo samples is used to obtain more accurate estimates of $e_i(\Gamma)$, and thus more accurate MCS estimates. When there are many samples to be analyzed, an alternative approach is to create a database of $\{e_i\}$ for all reasonable combinations of parameter values, and estimation for each sample will retrieve the values of $\{e_i\}$ from the database. This latter approach is the one we used in our analyses of 63 samples of dinucleotide repeats.

Let us consider how $e_i(\Gamma)$ can be estimated. Suppose we have M simulated samples of size n given the value of Γ, and suppose the number of alleles of size i in the jth sample is n_{ij}. Then we can use the mean $\hat{e}_i$ of n_{ij} as an estimate of $e_i(\Gamma)$. That is,

TABLE 1

Standard errors in estimating e_i in a sample of 100 chromosomes

Allelic size	$E(\hat{e}_i)^a$	$V(e_i)^b$	SE(500)[c]	SE(10^3)[c]	SE(10^4)[c]	SE(10^5)[c]
44	0.01	0.07	0.012	0.008	0.003	0.001
45	0.07	2.79	0.075	0.053	0.017	0.005
46	0.25	8.90	0.133	0.094	0.030	0.009
47	1.03	35.55	0.267	0.189	0.060	0.019
48	3.96	136.70	0.523	0.370	0.117	0.037
49	14.47	456.41	0.955	0.676	0.214	0.068
50	48.37	1019.38	1.428	1.010	0.319	0.101
51	20.70	588.45	1.085	0.767	0.243	0.077
52	7.61	251.57	0.709	0.502	0.159	0.050
53	2.47	86.92	0.417	0.295	0.093	0.029
54	0.78	27.03	0.233	0.164	0.052	0.016
55	0.23	7.65	0.124	0.087	0.028	0.009
56	0.05	1.52	0.055	0.039	0.012	0.004
57	0.02	0.42	0.029	0.021	0.006	0.002

The parameters used are $A = 50$, $P = 0.1$, $\alpha = 0.6$, and $\theta = 1.0$.
[a] e_i estimated from 200,000 independent samples.
[b] The estimated variance of e_i.
[c] The standard error of an estimate with the number of samples given in parentheses.

$$\hat{e}_i = \frac{1}{M}\sum_j n_{ij}. \quad (5)$$

The estimator is unbiased with variance $\mathrm{var}(\hat{e}_i) = \mathrm{var}(n_i)/M$, where n_i is the number of alleles of size i in a single random sample of n chromosomes. It follows that estimation accuracy increases with the number M of simulated samples. Therefore, the ability to simulate a large number of samples in a reasonable amount of time is critical for the MCS estimator to be practical. Coalescent algorithm is ideal for this purpose, and a sample of n alleles from a locus from a population evolving according to a SMM with value $\Gamma = (\theta, P, \alpha, A)$ can be simulated as follows:

First, a genealogy of n sequences (alleles) is generated using a coalescent algorithm (*e.g.*, Hudson 1991). For the simulated genealogy, we have not only the topological relationships of these sequences, but also the number of mutations that occurred on each branch of the genealogy. Simulation of such a genealogy requires only the value of θ. Second, assign a value to A, which is by definition the allele at the root of the genealogy; then determine the resulting allele of each mutation that occurred on the genealogy. Obviously the exercise requires knowing the allele type before a mutation and can be accomplished by starting from the root and progressing toward the tips of the genealogy. In this process, the type of a new mutant allele is simulated according to the distribution $\{\pi_i\}$, which is completely specified by the values of P and α.

Table 1 shows examples of the values of $\hat{e}_i$ and their standard errors for different numbers of Monte Carlo samples. Note that the majority of estimates are reasonably accurate even when $M = 500$. These results as well as many other simulation experiments we performed suggest that for the purpose of identifying the vicinity of the MCS estimate, 500–1000 Monte Carlo samples is usually sufficient, and 10,000 Monte Carlo samples is adequate for a fine-scale grid search to obtain the final MCS estimate.

We define $t_i(n) = e_i/n$ as the proportion of alleles being size i and measure the closeness between $\{t_i(n)\}$ for two different sample sizes by Euclidean distance, *i.e.*,

$$d = \sqrt{\sum_i (e_i/n - e_i'/n')^2},$$

where e_i' is the expected number of alleles i in a sample of size n'. An interesting observation that we made in our simulations is that values of $\{t_i(n)\}$ become steady rapidly even for a modest sample size. Table 2 shows several examples of Euclidean distances between $\{t_i(n)\}$ of different sample sizes. Note that the difference between them when sample size is larger than 100 is ex-

TABLE 2

Estimated Euclidean distance between $\{e_i/n\}$ of different sample sizes

	Sample size			
Sample size	25	50	100	200
50	0.0077			
100	0.0101	0.0030		
200	0.0129	0.0055	0.0031	
400	0.0135	0.0062	0.0036	0.0013

The e_i for each sample size was estimated from 200,000 independently simulated samples. $P = 0.1$; $\alpha = 0.5$; $\theta = 1.0$.

tremely small. This observation, although not too surprising from the viewpoint of coalescent process, is a reward for our MCS method, because it means that one can convert with excellent accuracy the estimate $\hat{e}_i$ for a sample of size n to the corresponding estimate for a sample of size m by $m\hat{e}_i/n$, saving considerable computer cpu time when many samples are to be analyzed.

APPLICATIONS

Since the discovery of highly polymorphic CA-repeat loci (Litt and Luty 1989; Weber and May 1989), many samples of such loci from human populations have been reported (*e.g.*, Kamino *et al.* 1993; Bowcock *et al.* 1994; Di Rienzo *et al.* 1994; Deka *et al.* 1995). The samples of Deka *et al.* (1995) are particularly useful for population study because their sample sizes are relatively large and the populations sampled are anthropologically well defined. We thus use their data to illustrate our method and to examine several issues about the evolution of microsatellite loci. Eight CA-repeat loci in nine different populations, the Samoan (SA), Dogrib Indian (DG), Pehuenche Indians (PH), New Guineans (NG), Kachari (KA), German (GR), CEPH (CP), Sokoto (SO), and Chimpanzee (CH), were reported in Deka *et al.* (1995), but we shall exclude the locus D13S137 from our analysis because there are many single nucleotide insertions/deletions within the repeat motifs at the locus. For the remaining seven loci, we also exclude all the alleles that are results of single nucleotide insertion/deletion because the mechanisms for changing allele size and for insertion/deletion are likely different. The allele sizes and their frequencies at these loci are given in the appendix of Deka *et al.* (1995). Figure 3 shows the allele frequencies at the loci FLT1 and D13S122. Because polymerase chain reaction (PCR) was used for amplification and the distances between the upstream primer sequences and the first CA-repeat were unknown for these loci, the resulting alleles were given in terms of sequence length from the primer sequences, instead of copy numbers. This does not pose difficulty in our analysis because the mutation model (1) is only dependent on the relative number $|i - j|$ of copy numbers, which are available from the samples. However, the estimated ancestral allele A at each locus will have to be given in terms of sequence length from the primer sequence.

Our task of estimation appears to be challenging at first glance because there are 63 samples to be analyzed, each requiring a considerable amount of computer cpu time. The observation we made earlier that the proportion of alleles of a given size is rather insensitive to sample size is of great help. Instead of generating a large number of samples for estimating e_i for each of the 63 samples, we choose to obtain good estimates of e_i/n for a sample of 100 chromosomes only and rescale them to obtain e_i for samples of different sizes. We selected $\theta = 0.2(0.2)10$ (*i.e.*, 0.2, 0.4, . . ., 9.8, 10), $P = 0.01(0.01)0.10(0.05)0.9$, and $\alpha = 0.5(0.5)1.0$, a total of 14,500 combinations of values of the three parameters. For each set of parameters, we generated 20,000 independent samples and obtained estimates of e_i/n from (5). Note that we do not need to estimate e_i/n for $\alpha < 0.5$ separately, because they can be obtained from those for $\alpha > 0.5$ because of symmetry. Therefore, we effectively obtained estimates of e_i/n for 29,000 sets of parameters. These estimates were stored in a database and can be retrieved easily.

For each of the 63 samples, we performed a grid search over all the parameter sets to obtain a MCS estimate of Γ. In other words, for each of 29,000 parameter sets, we computed the χ^2 value, updated the minimum χ^2 value and corresponding parameter values, and obtained the MCS estimate after all the parameter sets had been examined. This fine-scale grid search still requires nontrivial computer cpu time but is quite manageable. The estimation results are given in Table 3.

Note that 26 of the 63 samples show contraction ($\alpha < 0.5$) in allele sizes and 37 show expansion ($\alpha > 0.5$) in allele sizes (Table 3). This suggests that there is a slight bias of mutations toward expansion. Table 3 also suggests that most of the loci evolve in a multistepwise fashion. We divided samples into two groups, one showing allele size contraction and the other expansion; we then found that P values of the latter group are considerably larger than those of the former group, which means that size change during expansion is likely greater than that during contraction. This result implies that if the probability α of expanding allele size by a mutation is the same as or even slightly smaller than the probability of contracting allele size, the alleles in a population will still tend to increase in size. Because Table 3 shows that α in the majority of samples is larger than $\frac{1}{2}$, there is a tendency, stronger than that suggested by α alone, that most loci are expanding in allele size.

Rubinsztein *et al.* (1995) compared allele sizes of 42 microsatellite loci in several primate species and found that alleles in humans are generally longer than in other primates. They argued that microsatellite loci can evolve directionally and at different rates in closely related species. Our estimates of the ancestral allele sizes in human populations and in chimpanzees also show a slight bias toward longer alleles in humans than in chimpanzees, because the ancestral alleles of humans in five (FLT1, D13S118, D13S71, D13S122, and D13S124) of the seven microsatellite loci are longer than those of chimpanzees. However, some of these differences may be due to ascertainment bias, and analyses of more loci are needed to resolve this issue.

Weber and Wong (1993) studied 28 microsatellite loci in human chromosome *19* and a total of 20,000 parent-offspring allele pairs. They found that 78% of the 24 size changes *in vivo* were either gain or loss of single repeat unit, and gain or loss of more than three repeat units was not observed. When all the mutations

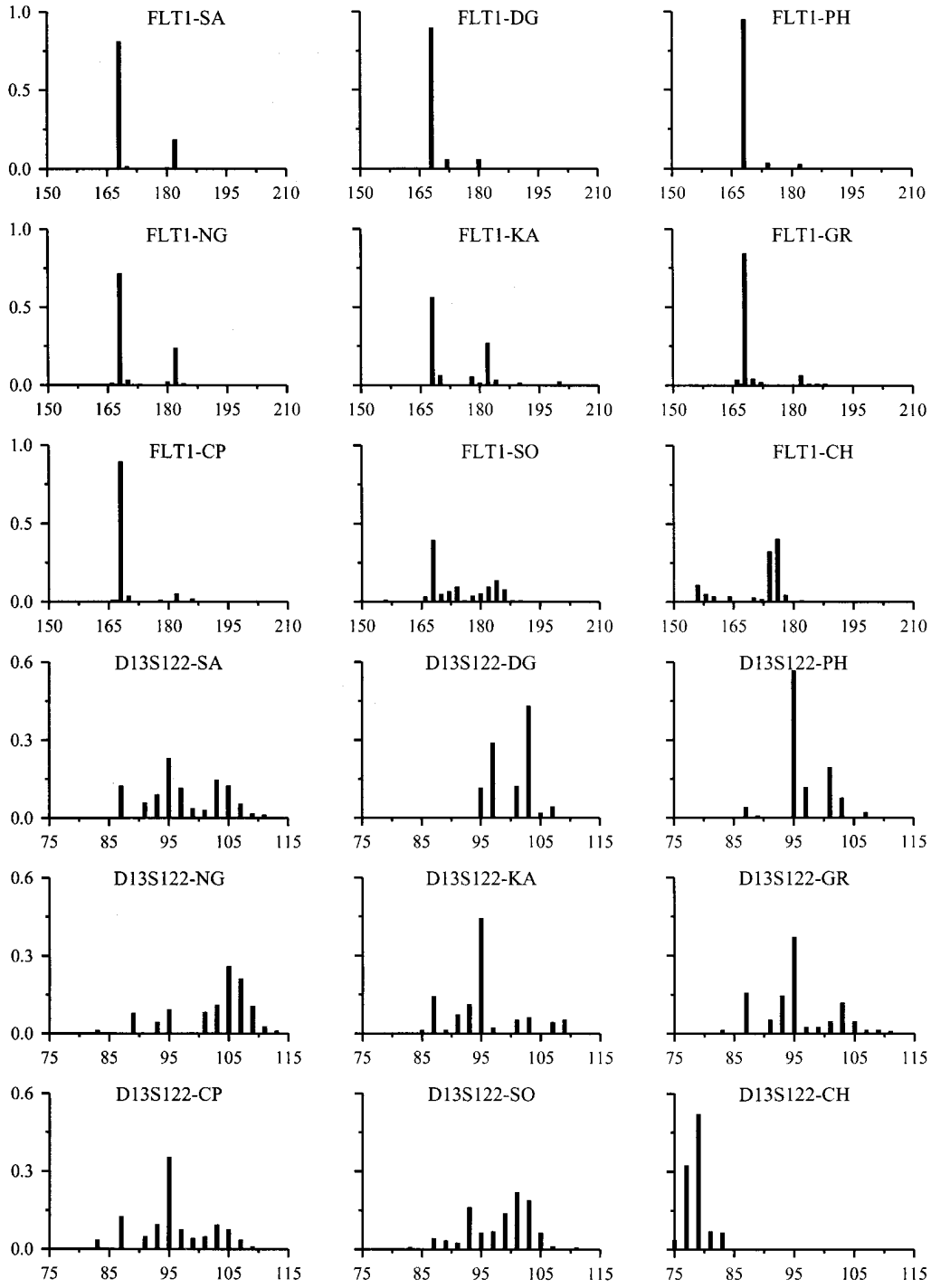

Figure 3.—Allele frequencies at loci FLT1 and D13S122.

in vivo and *in vitro* are considered, there is a strong tendency toward gains over losses in repeat units. Our analysis in general agrees with their observations. *In vivo* mutations as observed by Weber and Wong (1993) do not appear to suggest large *P* values. However, they do not necessarily contradict our estimates: first, because the number of mutations found in each locus examined in their study is simply too small to yield a reliable estimate of *P* for that locus; and second, because it is not unreasonable to suggest that a similar mutational bias may be occurring both *in vitro* and *in vivo*, and when both *in vitro* and *in vivo* mutations found in their study are considered, the mutation pattern would agree well with larger *P* values.

Let θ_{ij} be the value of θ for the *i*th population at the *j*th locus. If these loci are selectively neutral, then it would be reasonable to assume that the mutation rate at each locus is the same for different populations. Therefore, under the neutrality assumption, $\theta_{ij}/\theta_{ij'} = N_j/N_{j'}$, where N_j and $N_{j'}$ are the effective population sizes

TABLE 3

Estimated values (×100) of parameters θ, *P*, α, and *A*

Locus	Parameter	Populations								
		SA	DG	PH	NG	KA	GR	CP	SO	CH
FLT1	θ	60	20	20	140	180	40	40	180	180
	P	70	65	65	60	50	70	70	60	60
	α	95	100	100	90	100	80	80	90	10
	A	168	168	168	168	168	168	168	168	176
D13S118	θ	60	320	120	60	220	260	180	140	200
	P	55	1	50	55	9	10	25	45	2
	α	100	80	95	60	100	45	35	65	35
	A	190	190	190	190	188	194	194	190	186
D13S121	θ	60	60	40	60	140	120	120	240	300
	P	45	55	60	70	30	50	50	40	35
	α	100	90	90	95	75	75	75	60	30
	A	166	166	166	166	166	166	166	166	170
D13S71	θ	220	40	40	80	140	240	260	140	20
	P	25	50	2	50	25	6	2	20	2
	α	45	65	20	55	10	35	15	35	90
	A	73	75	75	73	77	75	77	75	71
D13S122	θ	880	160	100	300	140	240	240	300	100
	P	3	30	60	30	65	50	50	25	2
	α	25	10	80	5	40	45	55	15	40
	A	105	103	95	109	95	95	95	103	79
D13S193	θ	240	240	260	140	260	220	220	220	140
	P	55	40	55	65	40	55	55	45	40
	α	100	100	10	75	95	15	15	80	80
	A	131	131	147	133	131	147	147	131	133
D13S124	θ	120	20	20	60	80	100	100	120	300
	P	10	2	5	1	20	1	15	15	4
	α	25	25	15	0	30	95	95	60	100
	A	185	187	187	187	187	185	185	185	179

of the *j*th and *j*′th populations, respectively. That is, the ratio of θs of two different populations is independent of the locus studied. If estimates of θ are accurate, then we should expect to see a consistent value of ratio of θs over different loci. The estimated θs in Table 3 show that this ratio for most pairs of populations is not very consistent over loci. This suggests that the variance in θ estimate is likely to be large. The estimated θs also vary considerably among different populations for each locus, but this is expected because of different effective population sizes. A large variance in the estimation of θ is not unexpected because the estimate of θ, [for example, by Watterson's (1975) estimator], from segregating sites of DNA sequences that contain more information about θ than microsatellite data, is also accompanied by a relatively large variance. Furthermore, Kimmel and Chakraborty (1996) showed that the variance of a variance estimator of θ does not diminish with sample size.

When we draw conclusions based on estimated values of parameters, which are associated with variances, it is important to have some measures of accuracy in the estimates. Often the variance of an estimate of a single parameter is hard enough to compute, and measuring the accuracy of simultaneous estimation of all four parameters in a SMM model is undoubtedly more difficult, but nevertheless of great importance for recommending a method. This is an area in which answers demand substantially more computer resource than estimation, and it deserves much effort in the future. Therefore, we do not intend to provide all the answers here. Instead, we will focus on discussing the accuracy in the estimation of ancestral allele *A*. Although estimates of the four parameters are interrelated, we observed that values of the three parameters *p*, α, and θ that result in χ^2 that is close to the minimum are also close to the set that results in minimum χ^2, while MCS estimates conditional on different ancestral alleles can differ substantially. This is not difficult to understand. For example, specifying a small ancestral allele size (relative to the sizes of alleles in a sample) will require large α and *P* (or θ) to explain the observed allele frequencies, while, on the other hand, specifying a large ancestral allele size will result in a small estimated value for α. Our experience suggests that accuracy in the estimation of an ancestral allele is a good indicator of the overall accuracy of estimation.

To measure the accuracy of the estimate of *A*, we

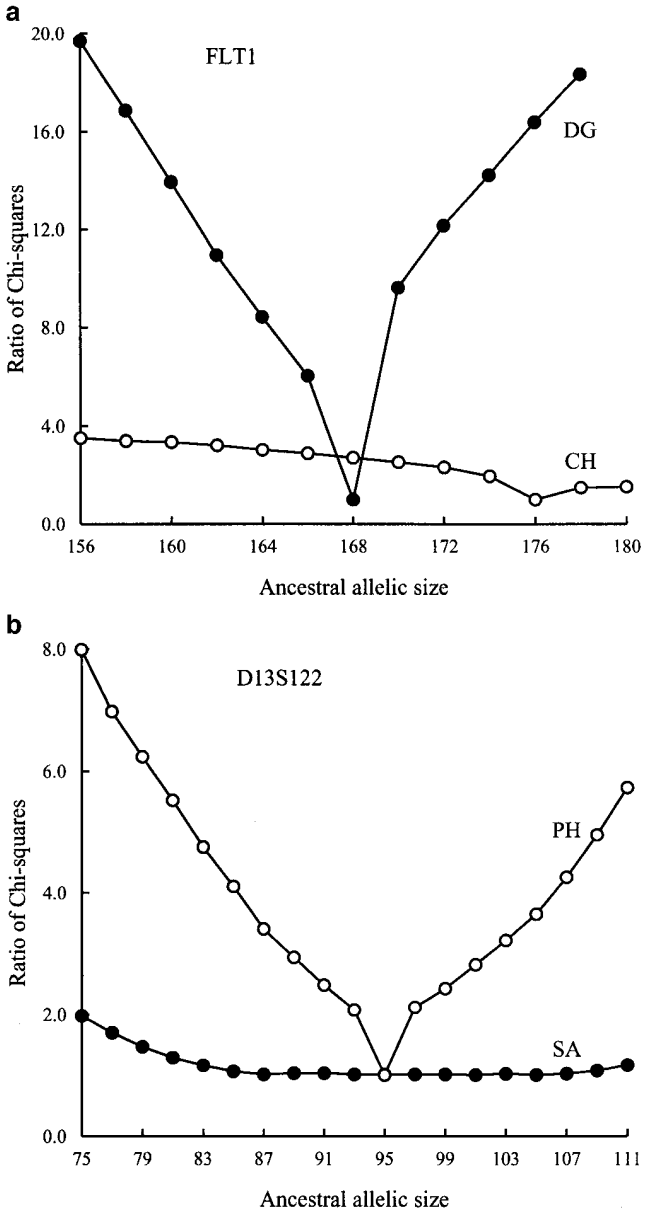

Figure 4.—Examples of the ratio of χ^2 over the minimum χ^2.

compare the minimum χ^2 values conditional on different ancestral allele sizes. One way to facilitate the comparison is a plot of χ^2 values *vs.* various ancestral allele sizes. A sharp decrease in the overall minimum value of χ^2 at allele A should suggest high accuracy of estimation. To allow comparison of the estimates for different samples, we use the ratio of conditional minimum χ^2 to the overall minimum χ^2. Figure 3 shows two examples. Part a suggests that the estimate of A for the FLT1 locus from population DG is accurate but the estimate from population CH is uncertain. The allele frequencies in Figure 3 concur with this analysis, because the frequency of estimated $A = 168$ in the DG sample is extremely high, while the frequency of estimated $A = 176$ in the CH sample is only intermediate, although it is the highest. Figure 4b also shows a similar pattern for locus D13S122. It is interesting to note that estimated $A = 105$ for the SA sample is not the most frequent allele in that sample. Table 3 shows that the θ estimate from this sample is suspiciously large. Because Figure 3b shows that the estimate of A for this sample is rather uncertain, we expect that quite different sets of P, α, and θ can result in χ^2 values close to the minimum. Indeed, when ancestral allele A is set to 85 the conditional minimum χ^2 value is equal to 154.59, which is less than 1% larger than the minimum χ^2 value, and the corresponding estimates of P, α, and θ are 0.3, 1.0, and 4.2, respectively.

When the ancestral allele is reasonably certain, it makes sense to examine closely the estimation of other parameters. Take the locus D13S122, for example; allele size 95 appears to be the ancestral allele for the PH and GR samples (see Figure 3 and Table 4). Under the condition $A = 95$, we examined the minimum χ^2 values for different values of α, and results are given in Figure 5. For the PH sample, Figure 5 shows that it is very unlikely that $\alpha < 0.5$ while for the GR sample, α should not be substantially different from 0.5. These conclusions are reinforced by the allele frequencies in the two samples that are shown in Figure 3.

Another measure R_2 of accuracy is the ratio of the χ^2 value of the second best estimate of A to that of the best. The larger the ratio, the worse the fit for the second best A, and thus the better the estimate of A. Another useful measure R_m is the ratio of the mean χ^2 values of the two neighboring sizes of A to that of A, which measures the goodness of the A compared to its neighbors. Obviously we always have $R_2 \leq R_m$. These two measures are particularly convenient when there are many samples to analyze, as in our situation. The values of R_2 and R_m, as well as the minimum χ^2 value for the 63 samples, are given in Table 4.

Table 4 shows that 43 of the 63 samples result in $R_2 > 1.20$, and 30 result in $R_2 > 1.5$. Although further study is required for proper interpretation of these R_2 values, it appears that other than the MCS estimate an increase of 20% or more in the χ^2 value for an ancestral allele should be a reasonable indication that the estimation is not totally out of line.

DISCUSSION

A strength of the MCS estimator developed in this article is its ability to simultaneously estimate all the parameters of the SMM, including the ancestral allele, making better use of available information in a sample. To date, mutation mechanisms for minisatellite and microsatellite loci are not yet fully understood, and even less is known about the mode of mutations, *i.e.*, whether it is symmetric or nonsymmetric, single-step or multistep. Although Kimmel *et al.* (1996) and Kimmel and Chakraborty (1996) emphasized that allelic size variance-based estimates of intra- and interpopulation varia-

TABLE 4

Some measures of quality in the estimates of ancestral allelic type

	Populations								
Locus	SA	DG	PH	NG	KA	GR	CP	SO	CH
FLT1	380.57[a]	46.01	45.30	627.88	221.04	81.14	60.10	187.20	202.82
	2.22[b]	6.03	10.36	1.73	1.48	5.11	6.10	1.70	1.49
	166[c]	166	166	166	166	166	166	166	178
	2.70[d]	7.83	13.42	2.04	1.73	6.82	8.17	1.95	1.72
D13S118	56.91	73.30	156.57	126.41	72.71	136.72	91.95	128.20	57.63
	3.84	1.02	1.47	3.31	1.05	1.09	1.22	1.41	1.05
	188	198	188	188	190	196	196	188	188
	4.85	1.08	1.69	3.87	1.19	1.09	1.22	1.60	1.20
D13S121	33.86	83.94	40.65	96.97	39.50	67.79	69.79	95.32	37.07
	4.74	2.14	7.80	4.23	1.44	2.03	1.51	1.16	1.04
	164	164	164	164	164	164	164	160	172
	5.94	2.67	8.61	5.46	1.47	2.39	1.71	1.24	1.17
D13S71	179.02	38.61	10.93	139.76	55.84	81.61	57.10	23.36	7.37
	1.05	4.39	11.15	2.24	1.42	1.01	1.03	2.05	38.28
	75	73	77	75	79	77	75	77	73
	1.10	4.99	11.26	2.35	1.54	1.14	1.17	2.47	38.65
D13S122	153.24	97.45	139.09	160.43	78.33	145.01	88.48	59.19	28.95
	1.00	1.26	2.07	1.02	1.61	1.23	1.39	1.09	1.47
	101	105	93	111	91	93	93	105	77
	1.02	1.30	2.09	1.03	1.70	1.27	1.40	1.17	2.49
D13S193	323.41	193.52	460.81	493.91	162.94	306.61	240.44	179.17	116.02
	1.27	1.19	1.16	1.14	1.06	1.13	1.19	1.03	1.34
	129	129	149	131	133	149	149	129	135
	1.43	1.28	1.28	1.37	1.07	1.31	1.43	1.09	1.47
D13S124	73.40	3.48	9.96	22.64	28.30	37.31	31.40	26.88	85.82
	1.44	56.04	20.86	5.81	1.85	3.00	2.82	2.27	1.07
	187	185	185	185	185	187	183	183	187
	163	56.46	21.72	8.02	2.45	3.27	2.88	3.19	1.18

[a] Minimum χ^2 value.
[b] Value of R_2.
[c] Second best ancestral allelic type.
[d] Value of R_m.

tion at repeat loci are not affected by asymmetry of allele size changes by mutations, their analyses reflect as well the concept that knowledge of the distribution of size change by mutation is critical. Furthermore, the modes of mutations are likely to differ from loci to loci. Therefore, being able to estimate simultaneously all the parameters of a SMM has considerable advantages over methods for estimating a single parameter that assume a mode of mutations, such as a symmetric single-step mutation model, which may be grossly incorrect.

Another strength of our method is its flexibility. Although we only considered a mutation model with three parameters and assumed a constant population size, it is not difficult to see that any combination of mutation model and population genetics model can be analyzed in a similar manner as long as alleles under these models can be simulated. In particular, one can consider mutation models with constraints on allele size or mutation rates dependent on allele size and more complex population genetic models, such as growth populations or subdivided populations. This strength should not be overlooked for two reasons. First, rapid accumulation of population samples from microsatellite and minisatellite loci provides excellent opportunities to examine various mutation models, and, second, many natural populations, particularly human populations, are not panmictic. Proper statistical inferences should be based on more realistic population models, allowing for population growth and subdivision. The expectation of alleles of a given size as estimated by Monte Carlo simulation provides great flexibility of the method, although it has the drawback of requiring more computer cpu time. Note that methods for parameter estimation that rely on Monte Carlo samples to obtain some necessary quantities were used in Fu (1994).

The MCS estimator we developed is a generalized least squares estimator and is often used in statistics for discrete distributions. Therefore, we expect our estimator to have many desirable properties. Even though the procedure of estimation takes advantage of a fast

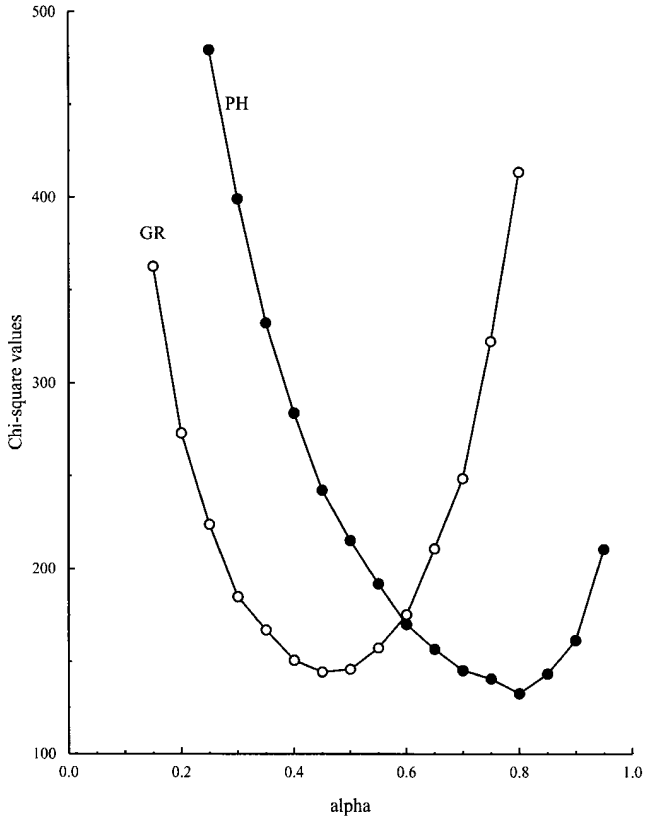

Figure 5.—Minimum χ^2 for different values of α for PH and GR samples at locus D13S122, conditional on ancestral allele size being 95.

coalescent algorithm, it is still a time-consuming method, which makes it hard to investigate the statistical properties of the estimator. Nevertheless, the statistical properties of the estimator will be worth studying in the future.

There are a number of potential extensions to the method we proposed. We chose to obtain parameter estimates from allele frequencies, but the same approach can be applied to a set of summary statistics, including various moments, number of alleles, heterozygosity, etc. It is also possible to incorporate variances and covariances of allele frequencies into an estimator, although doing so will demand even more computer resource. Another potential extension is to use the χ^2 statistics for testing hypotheses, such as the hypothesis that mutations follow a single-step SMM, but hypothesis testing is likely to be more effective using likelihood-based approaches such as the method by Nielson (1997).

This work was supported in part by National Institutes of Health grant R29 GM-50428 (Y.-X.F.) and GM58545 (R.C.).

LITERATURE CITED

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

Chakraborty, R., and M. Nei, 1982 Genetic differentiation of quantitative characters between populations or species. Genet. Res. **39:** 303–314.

Chakraborty, R., and M. Nei, 1997 Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model. Evolution **31:** 347–356.

Deka, R., L. Jin, M. D. Shriver, L. M. Yu, S. DeCroo, J. Hundrieser *et al.*, 1995 Population genetics of dinucleotide $(dC\text{-}dA)_n\cdot(dG\text{-}dT)_n$ polymorphism in world populations. Am. J. Hum. Genet. **56:** 461–474.

Di Rienzo, A., A. C. Peterson, J. C. Garza, A. M. Valdes, M. Statkin *et al.*, 1994 Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA **91:** 3166–3170.

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

Fu, Y. X., 1994 Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. Genetics **138:** 1375–1386.

Hudson, R. R., 1991 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. Futuyma and J. Antonovics. Oxford University Press, New York.

Jeffreys, A. J., V. Wilson and S. L. Thein, 1985 Hypervariable "minisatellite" regions in human DNA. Nature **314:** 67–73.

Kamino, K., J. Nakura, K. Kihara, L. Ye, K. Nagano *et al.*, 1993 Population variation in dinucleotide repeat polymorphism at the D8S360 locus. Hum. Mol. Genet. **2:** 1751.

Kimmell, M., and R. Chakraborty, 1996 Measure of variation at DNA repeat loci under a general stepwise mutation model. Theor. Popul. Biol. **50:** 345–367.

Kimmel, M., R. Chakraborty, D. N. Stivers and R. Deka, 1996 Dynamics of repeat polymorphisms under a forward-backward mutation model: within- and between-population variability at microsatellite loci. Genetics **143:** 549–555.

Kingman, J. F. C., 1982a The coalescent. Stochastic Process. Appl. **13:** 235–248.

Kingman, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

Li, W. H., 1976 Distribution of nucleotide differences between two randomly chosen cistrons in a subdivided population: the finite island model. Theor. Popul. Biol. **10:** 303–308.

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.

Nielsen, R., 1997. A likelihood approach to populations samples of microsatellite alleles. Genetics **146:** 711–716.

Ohta, T., and M. Kimura, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22:** 201–204.

Rubinsztein, D. C., W. Amos, J. Leggo, S. Goodburn, S. Jain *et al.*, 1995 Microsatellite evolution—evidence for directionality and variation in rate between species. Nature Genet. **10:** 337–343.

Shriver, M. D., L. Jin, R. Chakraborty and E. Boerwinkle, 1993 VNTR allele frequency distribution under the stepwise mutation model: a computer simulation approach. Genetics **134:** 983–993.

Stuart, A., and J. K. Ord, 1991 *Kendall's Advanced Theory of Statistics*, Vol. II, Ed. 6. Oxford University Press, New York.

Tautz, D., 1993 Notes on the definition and nomenclature of tandemly repetitive DNA sequence, pp. 21–28 in *DNA Fingerprinting: State of the Science*, edited by S. D. J. Pena, R. Chakraborty, J. T. Eppen and A. J. Jeffreys.

Valdes, A. M., M. Slatkin and N. B. Freimer, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133:** 737–749.

Watterson, G. A., 1975 On the number of segregation sites. Theor. Popul. Biol. **7:** 256–276.

Weber, J. L., and P. E. May, 1989 Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet. **44:** 388–396.

Weber, J. L., and C. Wong, 1993 Mutation of human short tandem repeats. Hum. Mol. Genet. **2:** 1123–1128.

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.

Communicating editor: G. B. Golding