Estimating Effective Population Size or Mutation Rate With Microsatellites

Hongyan Xu and Yun-Xin Fu[1]

Human Genetics Center, University of Texas, Houston, Texas 77030

Manuscript received May 30, 2003
Accepted for publication September 24, 2003

ABSTRACT

Microsatellites are short tandem repeats that are widely dispersed among eukaryotic genomes. Many of them are highly polymorphic; they have been used widely in genetic studies. Statistical properties of all measures of genetic variation at microsatellites critically depend upon the composite parameter $\theta = 4N\mu$, where N is the effective population size and μ is mutation rate per locus per generation. Since mutation leads to expansion or contraction of a repeat number in a stepwise fashion, the stepwise mutation model has been widely used to study the dynamics of these loci. We developed an estimator of θ, $\hat{\theta}_F$, on the basis of sample homozygosity under the single-step stepwise mutation model. The estimator is unbiased and is much more efficient than the variance-based estimator under the single-step stepwise mutation model. It also has smaller bias and mean square error (MSE) than the variance-based estimator when the mutation follows the multistep generalized stepwise mutation model. Compared with the maximum-likelihood estimator $\hat{\theta}_L$ by Nielsen (1997), $\hat{\theta}_F$ has less bias and smaller MSE in general. $\hat{\theta}_L$ has a slight advantage when θ is small, but in such a situation the bias in $\hat{\theta}_L$ may be more of a concern.

MICROSATELLITE loci, also known as short tandem repeats, are tandem repeat loci with repeat motifs of two to six nucleotides in length (Tautz 1993). Microsatellites are highly informative as polymorphic markers. Variations at microsatellite loci have been used to study the history and genetic structure of individual populations, such as DNA fingerprinting, paternity and relatedness testing, reconstruction of evolutionary trees, and genetic distance. In addition, they are useful for inferring migration histories, for identifying individuals of unknown origin, and for detecting the hidden population substructure. Microsatellites are also widely used in linkage mapping.

Statistical properties of all measures of genetic variation critically depend upon the composite parameter $\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate per locus per generation. An accurate estimate of θ will greatly facilitate the inference on the basis of variation at microsatellite loci. While the variation at microsatellites is extremely useful, little has been done to estimate θ using microsatellite data. This is partly due to the unknown mutation mechanism at such loci. Microsatellite loci are hypervariable and the mechanisms that produce new variation at such loci are unusual in comparison with those of classical loci. While the exact mechanism of mutations at such loci is still not well characterized at a molecular level (Jeffreys *et al.* 1994), it is generally believed that the processes and the patterns of mutations at different loci may differ from locus to locus, depending on the motif as well as the size of alleles at each locus. Empirical and theoretical studies indicate that for most microsatellite loci, mutations lead to stepwise changes of the repeat size of alleles although the rate of mutation leading to expansion may not be equal to that of contraction of allele size (Chakraborty *et al.* 1997; Deka *et al.* 1999). The stepwise mutation model, originally proposed for the study of protein charge changes (Ohta and Kimura 1973), in a more generalized form may be more suitable for the study of most microsatellite loci (Kimmel *et al.* 1996).

Although a number of estimators of θ (Wehrhahn 1975; Nielsen 1997; Fu and Chakraborty 1998) use microsatellite data, each has its limitations, in part being either too complicated or too simple. There is need for a relatively simple yet robust estimator and the purpose of this article is to develop one such estimator of θ using microsatellite data. Here we assume the neutral Wright-Fisher model without population substructure. The estimation of θ becomes the estimation of effective population size, N, when the mutation rate, μ, is known or the estimation of mutation rate, μ, when the effective population size, N, is known.

METHODS AND RESULTS

Existing estimators: Assuming the single-step stepwise mutation model, in which each mutation produces either one-step contraction or expansion in allele size, for a population without substructure and a neutral locus, the variance in allele size from a sample, V_s, has a mean equal to $\theta/2$ (Wehrhahn 1975). Then a convenient unbiased moment estimator is given by

[1]*Corresponding author:* Human Genetics Center, School of Public Health, University of Texas, 1200 Herman Pressler, Houston, TX 77030. E-mail: yunxin.fu@uth.tmc.edu

TABLE 1

Large sample variance of estimator $\hat{\theta}_V$

θ	$\mathrm{Var}(\hat{\theta}_V)$	SD[a]
1	1.67	1.29
5	35.0	5.92
10	136.67	11.69
50	3350.0	57.88

[a] Standard deviation of $\hat{\theta}_V$.

$$\hat{\theta}_V = 2V_s. \quad (1)$$

The estimator $\hat{\theta}_V$ is rather simple, but the price of its simplicity is a large variance. The variance of allele size variance, V_s, was given by Zhivotovsky and Feldman (1995) as

$$\mathrm{Var}(V_s) = \frac{1}{12}\theta + \frac{1}{3}\theta^2. \quad (2)$$

Consequently, the variance of $\hat{\theta}_V$ is given by

$$\mathrm{Var}(\hat{\theta}_V) = \frac{1}{3}\theta + \frac{4}{3}\theta^2. \quad (3)$$

Several examples of the value of $\hat{\theta}_V$ are shown in Table 1. In general the standard deviation is $>\theta$.

An even better known quantity is heterozygosity, denoted as H and defined as the probability that two randomly chosen sequences are of different allelic type; it is a measure of genetic variation at a microsatellite locus. The complement of heterozygosity, $F = 1 - H$, is called homozygosity. Since F contains the information of both number of alleles and allele frequency, an estimator based on F may be a possible solution.

Under the single-step stepwise mutation model, for a population without substructure and a neutral locus, the expected homozygosity (Ohta and Kimura 1973) is given by

$$E(F) = \frac{1}{\sqrt{1 + 2\theta}}. \quad (4)$$

Supposing a sample is taken from a population and letting k be the number of alleles in the sample, the homozygosity F can be estimated by

$$\hat{F} = \sum_{i=1}^{k} p_i^2, \quad (5)$$

where p_i is the allele frequency of the ith allele in the sample. Then a moment estimator of θ can be derived from Equation 4, replacing F with $\hat{F}$:

$$\tilde{\theta}_F = \frac{1}{2}\left(\frac{1}{\hat{F}^2} - 1\right). \quad (6)$$

Since the transformation is not linear, the estimator $\tilde{\theta}_F$ is usually biased, particularly when θ is large. Simple correction based on the infinite allele model was proposed before (Zouros 1979; Chakraborty and Weiss 1991), which is based on an analytical relationship between the expected value of the θ estimator and real value of θ. Unfortunately, such an analytical formula is not yet known for genetic loci evolving under the stepwise mutation model.

Besides the two estimators of θ using microsatellite data, Nielsen (1997) proposed an estimator using the maximum-likelihood approach. In addition to being restricted to the single-step stepwise mutation model, the estimator is rather demanding computationally and can handle only modest sample size. Fu and Chakraborty (1998) proposed an approach to simultaneously estimate all the parameters in a generalized stepwise mutation model, including θ. They use a minimum chi-square method to perform a grid search of all the possible values in the multidimensional parameter space, which makes it a challenge to analyze a large amount of data. To date, many population studies using microsatellites involve larger and larger samples and multiple loci. A relatively simple yet efficient estimator is highly desirable. In many ways, such an estimator can serve a role similar to that of Watterson's or Tajima's estimator of θ for DNA sequence data, despite the fact that several sophisticated estimators of θ for DNA sequence data have been available.

New estimator: The approach we take uses a combination of computer simulation and statistical regression, trying to find the relationship between the expectation of $\tilde{\theta}_F$ and the real value of θ. On the basis of the relationship, we try to develop a new unbiased estimator of θ. Computer simulation is an efficient way to study the properties of the homozygosity-based estimator $\tilde{\theta}_F$. For each combination of θ value and sample size, n, a large number of samples are simulated according to coalescent theory. For each sample, the homozygosity is estimated through Equation 5. Then the homozygosity-based estimate is obtained through Equation 6. Some of the results are shown in Figure 1, where each point in the figure is the mean of $\tilde{\theta}_F$ over 50,000 simulated samples. Figure 1 shows that $\tilde{\theta}_F$ on average overestimates θ. The magnitude of overestimation is a function of sample size n and θ, and, in many cases, the biases are severe.

To summarize the relationship among θ, n, and the mean of $\tilde{\theta}_F$, a regression approach can be used. The challenge is to find the simplest equation that is sufficiently accurate for describing the relationship. From Figure 1, it seems that mean of $\hat{\theta}_F$ is reversely related to sample size and positively proportional to θ. We include the terms $1/n$ and θ in the regression formula. We started to consider equations that incorporate $1/n$ and $\sqrt{\theta}$ in various ways. Choosing $\sqrt{\theta}$ as the basic unit was partly inspired by Equation 4. The most complex equation we consider is a polynomial including all combinations of $1/n$, $\sqrt{\theta}$, and $(1/n)^2$.

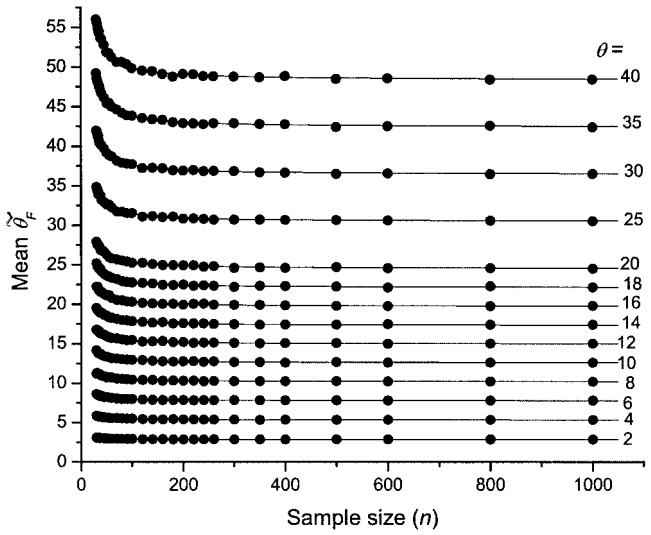

FIGURE 1.—Relationship and regression of θ, sample size n, and mean of $\tilde{\theta}_F$. Each dot is the mean of $\tilde{\theta}_F$ over 50,000 simulated samples and curves are the regression equations. The number on the right side of each curve is the θ value for simulating the samples upon which the mean $\tilde{\theta}_F$ is taken.

The regression analysis shows that two regression equations summarize remarkably well ($R^2 = 99.99\%$) the relationship of θ, n, and mean of $\tilde{\theta}_F$ (see Figure 1). For $\theta \leq 10$,

$$E(\tilde{\theta}_F) = \left(1.1313 + \frac{3.4882}{n} + \frac{28.2878}{n^2}\right)\theta + 0.3998\sqrt{\theta}. \quad (7)$$

For $\theta > 10$,

$$E(\tilde{\theta}_F) = \left(1.1675 + \frac{3.3232}{n} + \frac{63.698}{n^2}\right)\theta + 0.2569\sqrt{\theta}. \quad (8)$$

The regression equations have two nice properties. First, when $\tilde{\theta}_F = 0$, we have $\theta = 0$. Second, when sample size $n \to \infty$, $\tilde{\theta}_F$ has a limit value, which does not depend on n. Actually, when $n > 200$, the effect of sample size is very small.

On the basis of the above regression equations, we propose the following new estimator $\hat{\theta}_F$:

$$\hat{\theta}_F = \begin{cases} \theta \text{ value satisfies Equation 7 with } \tilde{\theta}_F \text{ replacing } E(\tilde{\theta}_F) & \text{if } \tilde{\theta}_F \leq 15 \\ \theta \text{ value satisfies Equation 8 with } \tilde{\theta}_F \text{ replacing } E(\tilde{\theta}_F) & \text{otherwise.} \end{cases}$$

The threshold value 15 is based on the observation that 90% of the value of $\tilde{\theta}_F$ is <15.0 with $\theta = 10$. However, we found that the choice is not critical, because choosing 10 as the threshold value does not make much difference. This is because when θ is ~10, Equations 7 and 8 give very similar results.

The performance of $\hat{\theta}_F$ was investigated through simulation. For a given combination of θ and sample size n, 50,000 samples were simulated and for each sample $\tilde{\theta}_F$ was estimated by Equation 6 and then corrected through Equation 7 or Equation 8. Some of the results are summarized in Table 2. Table 2 shows that the estimator $\hat{\theta}$ is unbiased (or nearly so). The small bias is likely due to fluctuation in simulation and is insignificant compared to the variance.

Next we compare the performance of our estimator $\hat{\theta}_F$ with that of the estimator based on allele size variance, $\hat{\theta}_V$. There are two ways to compute the variance of $\hat{\theta}_V$. The theoretical value of the large sample variance can be computed through Equation 3 and the variance can also be estimated through computer simulation. We computed it in both ways because on the one hand the validity of our simulation program can be checked and on the other hand the results can corroborate each other. The results are summarized in Table 3. Table 3 shows that the theoretical value of the variance of $\hat{\theta}_V$ agrees well with the simulation value, which indicates that our simulation is accurate. More importantly, Table 3 shows that while both estimators are unbiased, our homozygosity-based estimator $\hat{\theta}_F$ is better than the size variance-based estimator $\hat{\theta}_V$ in that the variance of $\hat{\theta}$ is smaller than that of $\hat{\theta}_V$. The relative efficiency of $\hat{\theta}_F$ against $\hat{\theta}_V$, defined as the ratio of the variance of $\hat{\theta}_V$ and variance of $\hat{\theta}_F$, is also given in Table 3. The relative efficiency increases as θ increases, which means that $\hat{\theta}_F$ becomes more and more efficient with increasing θ value. Note that since microsatellite loci have a relatively high mutation rate, the θ value can easily be of the range of 10–100, which makes $\hat{\theta}_F$ superior to $\hat{\theta}_V$ for most microsatellite loci.

Comparison with the maximum-likelihood estimator: The performance of the homozygosity-based estimator $\hat{\theta}_F$ is further compared to that of the maximum-likelihood (ML) estimator $\hat{\theta}_L$ proposed by NIELSEN (1997). Assuming the single-step stepwise mutation model, 10,000 samples are simulated for a number of combinations of θ and sample size. The two estimators, $\hat{\theta}_F$ and $\hat{\theta}_L$, are computed for each simulated sample. The mean value and mean square error (MSE) for the corresponding estimates are then computed and the results are summarized in Table 4. Two conclusions are obvious from Table 4. First, the ML estimator $\hat{\theta}_L$ is, in general, upwardly biased. Although the bias decreases with sample size, it is still appreciable even when the sample size is 300. In comparison, the mean value of the homozygosity-based estimator $\hat{\theta}_F$ exhibits little bias, similar to the case of comparing $\hat{\theta}_F$ and $\hat{\theta}_V$. Second, in general the ML estimator $\hat{\theta}_L$ has a larger MSE than that of $\hat{\theta}_F$, except in the cases where θ is small and sample size is large. It is somehow surprising that as θ increases, the relative performance of $\hat{\theta}_L$, measured by MSE, gets worse compared to $\hat{\theta}_F$. Two possible causes might be that the ML estimator implemented by Nielsen may not be a true ML estimator and it is not efficient. Indeed, in Nielsen's algorithm, a *k*-allele model was used to approximate the stepwise mutation model (NIELSEN 1997) in which the accuracy is not well known. Because of a high mutation rate for microsatellites, the θ value can be quite large

TABLE 2

Properties of $\hat{\theta}_F$

θ	n	$\bar{\theta}_F$	$\bar{\theta}_F e$	Bias	MSE	Variance
2	30	3.094	2.019	0.019	3.66	3.66
	50	3.009	2.053	0.053	3.40	3.39
	100	2.927	2.056	0.056	3.16	3.16
	200	2.884	2.054	0.054	3.11	3.11
	300	2.878	2.058	0.058	3.02	3.02
	400	2.869	2.056	0.056	3.06	3.06
	600	2.865	2.057	0.057	3.05	3.05
	1000	2.878	2.071	0.071	3.07	3.07
5	30	7.225	4.999	−0.001	16.52	16.52
	50	6.939	5.031	0.031	14.25	14.25
	100	6.742	5.043	0.043	12.90	12.90
	200	6.636	5.036	0.036	12.26	12.26
	300	6.604	5.035	0.035	12.10	12.10
	400	6.603	5.046	0.046	12.00	12.00
	600	6.561	5.023	0.023	11.78	11.78
	1000	6.575	5.045	0.045	11.98	11.97
10	30	14.186	10.157	0.157	61.03	61.00
	50	13.354	10.026	0.026	48.81	48.81
	100	12.958	10.051	0.051	42.02	42.01
	200	12.778	10.062	0.062	39.51	39.51
	300	12.691	10.041	0.041	38.43	38.43
	400	12.605	9.994	−0.006	37.79	37.79
	600	12.625	10.035	0.035	36.90	36.90
	1000	12.610	10.043	0.043	36.88	36.88
20	30	27.940	19.909	−0.091	217.52	217.51
	50	26.249	19.973	−0.027	171.77	171.77
	100	25.260	20.012	0.012	142.82	142.82
	200	24.941	20.099	0.099	131.68	131.67
	300	24.600	19.923	−0.077	125.90	125.89
	400	24.604	19.977	−0.023	124.55	124.55
	600	24.579	20.006	0.006	123.72	123.72
	1000	24.493	19.972	−0.028	122.48	122.48
30	30	41.942	30.104	0.104	514.50	514.49
	50	39.140	30.010	0.010	377.85	377.85
	100	37.721	30.126	0.126	308.05	308.03
	200	36.932	30.002	0.002	278.08	278.08
	300	36.850	30.094	0.094	270.54	270.53
	400	36.646	30.000	0.000	262.49	262.49
	600	36.565	30.007	0.007	262.79	262.79
	1000	36.480	29.994	−0.006	256.49	256.49
40	30	55.986	40.358	0.358	927.53	927.40
	50	51.862	39.946	−0.054	643.28	643.28
	100	49.832	39.986	−0.014	527.34	527.34
	200	49.100	40.085	0.085	481.06	481.05
	300	48.776	40.028	0.028	466.07	466.07
	400	48.829	40.174	0.174	451.92	451.89
	600	48.553	40.044	0.044	450.49	450.49
	1000	48.433	40.020	0.020	441.63	441.63

even for a modest population size. For example, many samples from human populations have yielded estimates of $\theta > 10$. This makes $\hat{\theta}_F$ more preferable in general than $\hat{\theta}_L$.

To address the issue of efficiency, we performed a large-scale simulation to see the extent to which performance of the ML estimator is affected by the number of runs through the Markov chain. In the comparison

TABLE 3

Comparison of $\hat{\theta}_F$ and $\hat{\theta}_V$

	$\hat{\theta}_F$			$\hat{\theta}_V$				
θ	Bias	MSE	Var	Bias	MSE	Var	Var(T)[a]	Efficiency[b]
1	0.052	1.16	1.16	−0.002	1.65	1.65	1.67	1.44
2	0.071	3.07	3.07	0.015	6.11	6.11	6.00	1.96
3	0.062	5.36	5.35	−0.010	12.57	12.57	13.00	2.43
4	0.045	8.36	8.36	0.001	23.32	23.32	22.67	2.71
5	0.045	11.98	11.97	0.012	35.71	35.71	35.00	2.92
6	0.049	16.25	16.25	0.006	51.54	51.54	50.00	3.08
8	0.040	25.85	25.85	−0.016	86.66	86.66	88.00	3.40
10	0.043	36.88	36.88	−0.059	129.69	129.69	136.67	3.71
12	0.064	51.23	51.22	−0.023	191.19	191.19	196.00	3.83
14	0.071	65.10	65.09	0.005	262.74	262.74	266.00	4.09
16	0.055	81.94	81.93	−0.109	337.60	337.59	346.67	4.23
18	0.021	101.61	101.61	−0.070	423.15	423.14	438.00	4.31
20	−0.028	122.48	122.48	0.108	549.14	549.13	540.00	4.41
25	−0.003	182.73	182.73	0.019	847.60	847.60	841.67	4.61
30	−0.006	256.49	256.49	0.124	1221.03	1221.01	1210.00	4.72
35	−0.068	340.53	340.53	−0.105	1649.56	1649.54	1645.00	4.83
40	0.020	441.63	441.63	−0.129	2103.78	2103.76	2146.67	4.86

[a] Theoretical value of variance of $\hat{\theta}_V$.
[b] Relative efficiency of $\hat{\theta}_F$ over $\hat{\theta}_V$.

with the ML estimator $\hat{\theta}_L$ shown in Table 4, the $\hat{\theta}_L$ was computed using the default Markov chain steps, 100,000 runs. Table 5 shows the results with three different numbers of runs through the Markov chain, 10,000, 100,000 and 1,000,000, where θ is set to 10.0. It is clear from Table 5 that there is a big improvement in the performance of $\hat{\theta}_L$ in terms of MSE when the number of runs through the Markov chain changes from 10,000 to 100,000, but only a small improvement when the replicate number changes from 100,000 to 1,000,000. More importantly, even when 1,000,000 replicates were used for the $\hat{\theta}_L$, it still has larger bias and MSE than the homozygosity-based estimator $\hat{\theta}_F$ when $\theta = 10.0$. An extreme case was carried out in which the number of runs through the Markov chain for $\hat{\theta}_L$ was set to 10,000,000 when $\theta = 10.0$ and sample size $n = 50$. In this case, the MSE of $\hat{\theta}_L$ was 69.53, which is still >50.62, the MSE of $\hat{\theta}_F$.

Robustness of the estimator: So far, the analysis is based on the single-step stepwise mutation model. While this may be true for some microsatellite loci, statistical analysis suggests that not all of them adhere to this simple version of the stepwise mutation model (Shriver *et al.* 1993; Di Rienzo *et al.* 1994). Furthermore, direct mutation assays at several loci showed that occasionally mutation may lead to jumps of allele sizes beyond one repeat unit (Weber and Wong 1993). On the basis of these lines of evidence, a generalized version of the stepwise mutation model (Kimmel and Chakraborty 1996; Fu and Chakraborty 1998) was proposed in which each mutation is supposed to change the allele size from X to $X + U$. The mutation is symmetric and the absolute value of the offset U is sampled from a geometric distribution with parameter λ; that is,

$$P(|U| = x) = (1 - \lambda)^{x-1}\lambda, \quad 0 < \lambda \leq 1. \qquad (9)$$

The performance of both estimators under this generalized stepwise mutation model was investigated through computer simulation. A total of 50,000 samples were simulated assuming the generalized model with $\lambda = 0.67$. With this λ value,

$$E(|U|) = 1/\lambda = 1.5.$$

That is, on average each mutation causes a jump of allele sizes of ~1.5 repeat units. For each simulated sample, the sample procedure as before was taken to obtain the two estimators, $\overline{\theta}_F$ and $\overline{\theta}_V$. The bias and MSE were also taken for each estimator. The corresponding theoretical values for the bias and MSE of $\hat{\theta}_V$ were also computed. The details are in the appendix. The simulation value agrees well with the theoretical value. The results are shown in Table 6.

Table 6 shows that under the generalized stepwise mutation model, both estimators are upwardly biased. That is, both estimators on average overestimate the real θ value. The bias is an increasing function of θ. When the bias of $\hat{\theta}_F$ is compared to that of $\hat{\theta}_V$, the former always has a smaller bias than the latter, which means that $\hat{\theta}_F$ is less biased than $\hat{\theta}_V$ especially when θ is high. Comparison between the corresponding MSEs also shows that $\hat{\theta}_F$ has a smaller MSE than $\hat{\theta}_V$. These two points make $\hat{\theta}_F$ still more preferable than $\hat{\theta}_V$ even when

TABLE 4

Comparison of $\hat{\theta}_F$ and $\hat{\theta}_L$ under various combinations of θ and sample size (n)

		$\hat{\theta}_F$		$\hat{\theta}_L$	
θ	n	Mean	MSE	Mean	MSE
2	30	2.027	3.635	2.358	3.249
	50	2.028	3.403	2.306	2.812
	100	2.041	3.193	2.238	2.159
	200	2.086	3.189	2.203	1.901
	300	2.039	3.015	2.158	1.725
5	30	4.921	16.328	5.727	19.949
	50	4.956	13.502	5.534	14.356
	100	5.036	12.618	5.382	11.569
	200	5.014	11.694	5.321	10.029
	300	5.075	12.091	5.224	9.579
10	30	9.987	58.355	12.189	106.184
	50	10.002	47.552	11.945	85.583
	100	9.967	42.251	11.296	58.665
	200	9.930	38.276	10.945	47.266
	300	10.058	38.725	10.866	45.992
20	30	20.044	241.480	26.200	635.676
	50	20.008	178.937	25.259	392.604
	100	20.200	151.968	24.227	272.930
	200	20.196	136.492	23.290	217.196
	300	19.945	129.011	22.569	192.065

The default value, 100,000 for the number of runs through the Markov chain, was used to compute $\hat{\theta}_L$.

TABLE 5

Comparison of $\hat{\theta}_F$ and $\hat{\theta}_L$ when θ = 100 for different numbers of runs through the Markov chain

		$\hat{\theta}_F$		$\hat{\theta}_L$	
MC replicates	n	Mean	MSE	Mean	MSE
10,000	30	10.12	60.28	12.59	129.90
	50	9.93	48.83	11.93	96.72
	100	10.07	42.87	11.55	81.39
	200	10.03	39.50	11.08	73.05
	300	10.03	37.65	11.01	70.18
100,000	30	9.99	58.36	12.19	106.18
	50	10.00	47.55	11.95	85.58
	100	9.97	42.25	11.30	58.67
	200	9.93	38.28	10.95	47.27
	300	10.06	38.73	10.87	45.99
1,000,000	30	9.95	61.30	11.90	102.60
	50	10.07	48.44	11.59	76.51
	100	9.98	40.65	11.08	55.96
	200	10.10	40.56	10.86	47.88
	300	9.98	36.89	10.56	39.73

the actual mutation model is the generalized stepwise mutation model.

APPLICATION

To test the performance of the homozygosity-based estimator $\hat{\theta}_F$ with real data, we use the allele frequency data from the ALFRED database at Yale University (Cheung *et al.* 2000). There are altogether 115 dinucleotide repeats with data from 10 worldwide populations. The 10 populations are Biaka, Mbuti, Druze, Danes, Han, Japanese, Melanesian-Nasioi, Yakut, Maya-Yucatan, and Surui. More information about the loci and populations can be found at http://alfred.med.yale.edu/alfred/index.asp.

For each population-locus combination, $\hat{\theta}_F$ and $\hat{\theta}_V$ are computed. To compare the consistency of the estimators, one locus is randomly chosen as the base locus and the ratio of the estimate for other loci in the same population is taken over the estimate for the base locus. Since the effective population size is generally supposed to be the same in the same population for all loci from the same sample, we are estimating the ratio of mutation rates using information from different populations. Assuming the mutation rate for a particular locus is constant across the populations, the estimates of the ratio of mutation rates from different populations are the estimates of the same quantity. Consequently, the dispersion of the results is an indicator of the consistency of the estimator. The coefficient of variance (ratio of standard deviation to mean) is taken as a measure of dispersion. In almost all the cases, the coefficient of variance is smaller with $\hat{\theta}_F$ than with $\hat{\theta}_V$, which indicates that the homozygosity-based estimator $\hat{\theta}_F$ is more stable and more consistent than the variance-based estimator $\hat{\theta}_V$. Examples of the results from four loci are tabulated in Table 7, where the base locus (locus 1) is D11S935, locus 2 is D7S640, locus 3 is D6S441, and locus 4 is D5S408, with the corresponding mutation rates denoted as μ_1–μ_4, respectively.

DISCUSSION

Kimmel and Chakraborty (1996) showed that sample homozygosity at a microsatellite locus depends not only on θ, but also on the pattern of allele size change caused by mutation. Therefore, any attempt to estimate θ on the basis of homozygosity has to be mutation model dependent. Interestingly, the regression formula we found on the basis of the single-step stepwise mutation model is reasonably robust against deviations from the single-step model. This is a useful property since it is very difficult to specify the model with confidence. On the other hand, if one has sufficient confidence in a particular model, a similar approach can be used to derive the regression formula under the model. This can be seen from our simulation study when the mutation

TABLE 6

Comparison of $\hat{\theta}_F$ and $\hat{\theta}_V$ under the generalized model

	$\hat{\theta}_F$		$\hat{\theta}_V$			
θ	Bias	MSE	Bias	Bias(T)[a]	MSE	MSE(T)[b]
1	0.39	2.30	1.99	2	25.72	26
2	0.96	7.42	3.97	4	84.52	84
3	1.63	15.46	5.90	6	170.98	174
4	2.33	26.38	7.75	8	288.51	296
5	3.17	40.12	9.86	10	453.10	450
6	4.05	58.32	11.97	12	641.03	636
8	5.88	105.93	15.75	16	1,086.61	1,104
10	7.88	168.99	19.72	20	1,660.19	1,700
12	9.95	250.23	23.73	24	2,387.10	2,424
14	12.00	341.60	27.22	28	3,047.05	3,276
16	14.41	468.47	31.88	32	4,396.28	4,256
18	16.56	601.07	35.07	36	5,181.32	5,364
20	19.01	770.27	39.00	40	6,440.48	6,600
25	25.16	1,276.12	49.09	50	10,362.25	10,250
30	31.66	1,921.98	59.05	60	13,896.41	14,700
35	38.43	2,756.88	67.99	70	18,798.79	19,950
40	45.13	3,706.23	77.91	80	25,021.12	26,000

[a] Theoretical value of bias of $\hat{\theta}_V$ under the generalized model.
[b] Theoretical value of variance of $\hat{\theta}_V$ under the generalized model.

model deviates from the single-step stepwise mutation model to the generalized stepwise mutation model.

Although the maximum-likelihood estimator, $\hat{\theta}_L$, proposed by Nielsen (1997) is computationally demanding, its performance was compared to that of the homozygosity-based estimator $\hat{\theta}_F$ through a large-scale simulation. The ML estimator $\hat{\theta}_L$ is found to be slightly upwardly biased. This is not too surprising because many maximum-likelihood estimators are known to be biased for small sample sizes. Indeed we found that the $\hat{\theta}_L$ approaches the true value as sample size (n) increases. However, even when $n = 300$, there is still an appreciable amount of bias. The MSE of $\hat{\theta}_L$ decreases with the increase of the sample size. However, in the most likely range of θ for microsatellites, $\hat{\theta}_L$ has in general larger MSE than $\hat{\theta}_F$ unless the sample size is extremely large. $\hat{\theta}_L$ has a slight advantage when θ is small. However, in such a situation, the bias of $\hat{\theta}_L$ may be more of a concern.

TABLE 7

Comparison of estimates of ratio of mutation rates with $\hat{\theta}_F$ and $\hat{\theta}_V$

	$\hat{\theta}_F$			$\hat{\theta}_V$		
Population	μ_2/μ_1	μ_3/μ_1	μ_4/μ_1	μ_2/μ_1	μ_3/μ_1	μ_4/μ_1
Biaka	5.58	3.61	2.28	1.68	1.76	2.42
Mbuti	5.48	8.96	4.40	4.23	5.66	10.55
Druze	3.11	4.18	0.94	0.93	1.69	2.58
Danes	5.61	3.30	1.48	0.61	0.84	0.56
Han	5.19	1.03	0.72	0.66	0.67	0.88
Japanese	10.16	3.86	1.19	0.60	0.24	1.48
Nasioi	1.68	3.61	4.91	0.77	1.77	3.14
Yakut	3.83	2.33	0.60	0.27	0.90	0.05
Yucatan	2.86	3.18	0.23	0.91	1.51	1.43
Surui	5.33	5.20	3.24	1.47	1.99	3.58
Mean	4.88	3.93	2.00	1.21	1.70	2.67
SD of mean	0.73	0.66	0.52	0.36	0.48	0.95
Variance	5.34	4.34	2.74	1.30	2.26	8.96
Coefficient of variance	0.47	0.53	0.83	0.94	0.88	1.12

For example, from Table 4 when $\theta = 2.0$ and $n = 30$, the bias can be nearly 18%. All these factors make $\hat{\theta}_F$ an attractive alternative to $\hat{\theta}_L$.

We have relied on regression to find a way to remove bias as an estimator of θ from $\tilde{\theta}_F$. It should be pointed out that jackknife is a widely used approach to reduce bias in estimation (*e.g.*, Manly 1997). The underlying theory is that a jackknife estimator removes the bias of order $1/n$; that is, if the original biased estimate $\hat{\theta}$ has the form

$$E(\hat{\theta}) = \theta\left(1 + \frac{A}{n}\right), \tag{10}$$

where A is a constant, then the jackknife estimator can remove the bias. However, the relationship between $E(\tilde{\theta}_F)$ and θ is rather complex. Although the exact relationship is unknown, Equations 7 and 8 indicate that the relationship is certainly not in the form of Equation 10. So the jackknife estimator is unlikely to be able to remove much of the bias in $\tilde{\theta}_F$. Indeed, when the jackknife method was applied in our simulated sample, we found that it was able to remove only ~10% of the bias in many combinations of parameters. Therefore, jackknife is not an appropriate approach to use in this situation.

From Equation 5 of Kimmel and Chakraborty (1996), the estimator based on allele size variance under any arbitrary stepwise mutation model is given by

$$\tilde{\theta}_V = \frac{V}{E(U_0^2)}, \tag{11}$$

where $V = 2E(V_s)$ and U_0 is the symmetrized allele size change in a single generation and is mutation model dependent. Consequently, the variance-based estimator $\tilde{\theta}_V$ is mutation model dependent and is applicable to the particular model itself. In the case of the single-step stepwise mutation model, Equation 11 is reduced to Equation 1 since $E(U_0^2) = 1$. Therefore, $\hat{\theta}_V$ is a special case of $\tilde{\theta}_V$ and is mutation model dependent and applicable to the single-step stepwise mutation model. Hence it is no surprise that $\hat{\theta}_V$ becomes biased under the generalized stepwise mutation model.

Rubinsztein *et al.* (1995) argued that the mutational transitions may be asymmetric. During the analysis in this article we did not differentiate the asymmetric model from the symmetric model. This is because from Kimmel and Chakraborty (1996) homozygosity and allele size variance are independent of mutation direction. Indeed, these are confirmed in our simulation (data not shown). Consequently, our homozygosity-based estimator $\hat{\theta}_F$ is applicable for single-step stepwise mutation, symmetric or not. Computer programs to carry out the analysis and to estimate $\hat{\theta}_F$ are available upon request.

We thank R. Nielsen for sharing his ML program. This work was supported partly by National Institutes of Health grants R01 GM50428 and R01 GM60777 to Y.-X. Fu.

LITERATURE CITED

Chakraborty, R., and K. M. Weiss, 1991 Genetic variation of the mitochondrial DNA genome in American Indians is at mutation-drift equilibrium. Am. J. Anthropol. **86:** 497–506.

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

Cheung, K. H., M. V. Osier, J. R. Kidd, A. J. Pakstis, P. L. Miller *et al.*, 2000 ALFRED: an allele frequency database for diverse populations and DNA polymorphisms. Nucleic Acids Res. **28:** 361–363.

Deka, R., G. Sun, D. Smelser, Y. Zhong, M. Kimmel *et al.*, 1999 Rate and directionality of mutations and effects of allele size constraints at anonymous, gene-associated and disease-causing trinucleotide loci. Mol. Biol. Evol. **16:** 1166–1177.

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

Fu, Y. X., and R. Chakraborty, 1998 Simultaneous estimation of all the parameters of a stepwise mutation model. Genetics **150:** 487–497.

Jeffreys, A. J., K. Tamaki, A. MacLeod, D. G. Monckton, D. L. Neil *et al.*, 1994 Complex gene conversion events in germline mutation at human minisatellites. Nat. Genet. **6:** 136–145.

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

Kimmel, M., R. Chakraborty, D. Stives 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.

Manly, B. F. J., 1997 *Randomization, Bootstrap and Monte Carlo Methods in Biology.* Chapman & Hall, New York.

Nielsen, R., 1997 A likelihood approach to population 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. Nat. Genet. **10:** 337–343.

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

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

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

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

Zhivotovsky, L. A., and M. W. Feldman, 1995 Microsatellite variability and genetic distances. Proc. Natl. Acad. Sci. USA **92:** 11549–11552.

Zouros, E., 1979 Mutation rates, population sizes and amounts of electrophoretic variation of enzyme loci in natural populations. Genetics **92:** 623–646.

Communicating editor: J. B. Walsh

APPENDIX: CALCULATING THE BIAS AND MSE OF $\hat{\theta}_V$ UNDER THE GENERALIZED STEPWISE MUTATION MODEL

Given that

$$P(|U| = x) = (1 - \lambda)^{x-1}\lambda, \quad \text{where } \lambda = 0.67,$$

that is, $|U| \sim$ geometric(0.67), since the mutation is symmetric, $U_0 = U$, we have

$$\begin{aligned} E(U_0^2) &= E(U^2) \\ &= \mathrm{Var}(U) + (E(U))^2 \\ &= \frac{1-\lambda}{\lambda^2} + \frac{1}{\lambda^2} \\ &= \frac{2-\lambda}{\lambda^2}. \end{aligned} \tag{A1}$$

Since $\hat{\theta}_V = 2V_s$ and from Equation 4 of Kimmel and Chakraborty (1996) we have

$$E(V_s) = \frac{V}{2} = \frac{\theta}{2}E(U_0^2), \tag{A2}$$

where V is defined in Kimmel and Chakraborty (1996),

$$E(\hat{\theta}_V) = 2E(V_s) = 2 \times \frac{\theta}{2}E(U_0^2) = \frac{2-\lambda}{\lambda^2}\theta. \tag{A3}$$

Substituting $\lambda = 0.67$ into Equation A3, we have

$$E(\hat{\theta}_V) = 3\theta. \tag{A4}$$

Therefore,

$$\mathrm{Bias}(\hat{\theta}_V) = 3\theta - \theta = 2\theta. \tag{A5}$$

To calculate the MSE of $\hat{\theta}_V$ we need to calculate variance of size variance V first. From Equation 16 of Kimmel and Chakraborty (1996),

$$\mathrm{Var}(V_s) = \frac{1}{3}V^2 + \frac{1}{12}V\frac{E(U_0^4)}{E(U_0^2)}. \tag{A6}$$

Since $U_0 = U$ and $U \sim$ geometric(0.67), the moment-generating function of U_0 is

$$M(t) = \frac{\lambda e^2}{1 - \lambda e^2}. \tag{A7}$$

Taking the fourth derivative of Equation A7 and setting $t = 1$, $\lambda = 0.67$, we have

$$E(U_0^4) = 30. \tag{A8}$$

From Equation 5 of Kimmel and Chakraborty (1996), we have

$$V = \theta E(U_0^2). \tag{A9}$$

Substituting Equations A1, A8, and A9 into Equation A6, we have

$$\mathrm{Var}(V_s) = 3\theta^2 + \frac{5}{2}\theta. \tag{A10}$$

Since $\hat{\theta}_V = 2V_s$, we have

$$\mathrm{Var}(\hat{\theta}_V) = 4\,\mathrm{Var}(V_s) = 12\theta^2 + 10\theta. \tag{A11}$$

Therefore,

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}_V) &= [\mathrm{Bias}(\hat{\theta}_V)]^2 + \mathrm{Var}(\hat{\theta}_V) \\ &= (2\theta)^2 + 12\theta^2 + 10\theta \\ &= 16\theta^2 + 10\theta. \end{aligned} \tag{A12}$$