Constraints on Allele Size at Microsatellite Loci: Implications for Genetic Differentiation

Maarten J. Nauta and Franz J. Weissing

Department of Genetics, University of Groningen, 9750 AA Haren, The Netherlands

Manuscript received August 22, 1995
Accepted for publication February 28, 1996

ABSTRACT

Microsatellites are promising genetic markers for studying the demographic structure and phylogenetic history of populations. We present theoretical arguments indicating that the usefulness of microsatellite data for these purposes may be limited to a short time perspective and to relatively small populations. The evolution of selectively neutral markers is governed by the interaction of mutation and random genetic drift. Mutation pressure has the inherent tendency to shift different populations to the same distribution of alleles. Hence, mutation pressure is a homogenizing force, and population divergence is caused by random genetic drift. In case of allozymes or sequence data, the diversifying effect of drift is typically orders of magnitude larger than the homogenizing effect of mutation pressure. By a simple model, we demonstrate that the situation may be different for microsatellites where mutation rates are high and the range of alleles is limited. With the help of computer simulations, we investigate to what extent genetic distance measures applied to microsatellite data can nevertheless yield useful estimators for phylogenetic relationships or demographic parameters. We show that predictions based on microsatellite data are quite reliable in small populations, but that already in moderately sized populations the danger of misinterpretation is substantial.

MICROSATELLITES are a class of tandem repeat loci, where alleles can be distinguished by their size (TAUTZ 1993). At these loci, a mutation may alter the size of an allele by adding or deleting one or more repeats. The mutation rate is exceptionally high (*e.g.*, LEVINSON and GUTMAN 1987; JEFFREYS *et al.* 1988; KELLY *et al.* 1991), implying a high degree of polymorphism. As a consequence, microsatellites seem very promising for studying the genetic structure of natural populations. However, it is not yet clear which genetic distance measures are adequate for studying microsatellite data. The discussion centers around the question whether distance should be based on the variance in allele frequencies or on the variance in repeat number (GOLDSTEIN *et al.* 1995a; SLATKIN 1995).

In this paper, we discuss a general problem that applies to *all* distance measures. The range of allele sizes found at microsatellite loci is typically limited (*e.g.*, GARZA *et al.* 1995). Combined with the high mutation rate, this has important implications since genetic information accumulated by a population may rapidly decay. Generally, mutation is regarded as a factor that enhances the differentiation between populations. However, when the range of target alleles is limited, mutation will lead to the reappearance of alleles lost in the past. As a consequence, mutation may be viewed as a homogenizing factor that counteracts the diversifying effects of random genetic drift.

To formalize this idea, consider a finite number of M alleles $A_1, \ldots, A_M$, segregating in an infinite population with discrete, nonoverlapping generations. Let f_i denote the relative frequency of allele A_i. Mutation pressure can be characterized by a probability matrix $U = (u_{ij})$, where u_{ij} denotes the probability that allele A_j mutates to allele A_i. Evolution under mutation pressure is then governed by the recurrence equations:

$$f_i' = \sum_j u_{ij} f_j = (U \cdot f)_i. \tag{1}$$

Under mild regularity assumptions on the mutation matrix U, Markov theory (*e.g.*, KEMENY and SNELL 1976) predicts that, irrespective of the starting conditions, mutation pressure will shift the population to a fixed limit distribution of allele frequencies that is implicitly given by

$$f^* = U \cdot f^*. \tag{2}$$

Hence, two separated populations have the inherent tendency to converge to the *same* allele frequency distribution. In other words, mutation pressure *per se* leads to genetic convergence rather than to genetic divergence of subpopulations.

Strictly speaking, this result only applies to infinite populations. In finite populations, genetic divergence is possible due to the diversifying action of genetic drift. Whether the combined action of drift and mutation will lead to genetic divergence rather than to genetic convergence will depend on the relative importance of

Corresponding author: Franz J. Weissing, Department of Genetics, University of Groningen, P.O. Box 14, 9750 AA Haren, The Netherlands. E-mail: weissing@biol.rug.nl

these factors. Obviously, the sampling effects leading to genetic drift will be strongest when population size N is small. On the other hand, the inherent tendency to converge to a limit distribution under mutation pressure is strongest when the number M of alleles is small and when the mutation rates are high.

In the case of allozyme and sequence data, mutation rates are very low (10^{-6}, say) and/or the number of possible target alleles is large. As a consequence, the homogenizing effect of mutation will typically be dominated by the diversifying action of drift, even in relatively large populations. In case of microsatellites, the situation is different. In fact, mutation rates are orders of magnitude higher (10^{-3}, say) and the number of alleles is typically smaller than 20 (*e.g.*, BOWCOCK *et al.* 1994; GARZA *et al.* 1995; GOLDSTEIN *et al.* 1995a,b; ZHIVOTOVSKY and SLATKIN 1995). We shall demonstrate that this may have important implications even for relatively small populations. In fact, the interaction of drift and mutation does not necessarily lead to genetic divergence and genetic distance measures cannot be expected to have the desirable property that the distance of two populations increases with the time of their genetic isolation.

To address the question for what population sizes the genetic structure of populations can be correctly inferred from microsatellite data, we consider the stepwise mutation process (OHTA and KIMURA 1973), one of the standard models for mutation at microsatellite loci. Previous work has focused on the infinite alleles version of this model (*e.g.*, OHTA and KIMURA 1973; KIMURA and OHTA 1975; MORAN 1975). For this version, it could be shown that, even for the high mutation rates typical for microsatellites, genetic isolation will lead to genetic differentiation with respect to genetic distance measures such as D_1 (GOLDSTEIN *et al.* 1995a) or R_{ST} (SLATKIN 1995). The empirical relevance of these results may, however, be limited since the homogenizing force of mutation pressure is extremely weak or even absent in infinite alleles models. We shall therefore confront the infinite alleles version of the stepwise mutation model with a more realistic version, which incorporates the assumption that the range of possible allele sizes at a microsatellite locus is restricted.

First, we compare the two versions of the stepwise mutation model with respect to their general characteristics. Taking D_1 as an example, we then study the consequences of a constrained range of allele sizes for genetic distance measures. By means of computer simulations, we subsequently investigate some practical implications such as the use of genetic distance measures for the estimation of divergence time and the reconstruction of phylogenetic relationships. Finally, we discuss the robustness of our results by considering different model parameters and by showing that our conclusions do not depend on the specific model chosen, but that they can be drawn from other models as well. Let us stress from the beginning that it is not our purpose to present the finite alleles version of the stepwise mutation process as the most adequate model for mutation pressure at microsatellite loci. We rather use this simple model to illustrate the general principle that constraints on the number of alleles may be relevant already in the context of moderately sized populations.

THE FINITE ALLELES VERSION OF THE STEPWISE MUTATION MODEL

In its simplest version, the stepwise mutation model assumes that each mutation event at a microsatellite locus leads to the addition or the deletion of a single repeat. If we assume that addition and deletion of repeats have the same probability, an allele with i repeats mutates to either $i - 1$ or to $i + 1$ repeats, each with probability $\mu/2$, where μ is the mutation probability per gamete and generation. In other words, the mutation matrix U is given by $u_{i+1,i} = u_{i-1,i} = \mu/2$, $u_{i,i} = 1 - \mu$, and $u_{i,j} = 0$ otherwise. This one-step version of the stepwise mutation model appears to give a reasonable approximation of the mutation process at microsatellite loci (*e.g.*, SHRIVER *et al.* 1993; VALDES *et al.* 1993; WEBER and WONG 1993) and is often assumed in theoretical studies (*e.g.*, GOLDSTEIN *et al.* 1995a). It can, however, be easily modified to include larger mutation steps (*e.g.*, GARZA *et al.* 1995; SLATKIN 1995). The mutation frequency at microsatellite loci appears to range from 10^{-5} to 10^{-2} (*e.g.*, EDWARDS *et al.* 1992; WEBER and WONG 1993). Typically, we have used a value of $\mu = 10^{-3}$ for our simulations. The effects of the mutation rate on our results will be discussed later on.

Theoretical research has centered around the infinite-alleles version of the stepwise mutation model, which assumes that there are no constraints on allele size (*i.e.*, allele size i ranges from $-\infty$ to $+\infty$). As a consequence of this assumption, the distribution of allele sizes does, even in large populations, not converge to a characteristic limit distribution f^*. Instead, the distribution of allele sizes wanders around indefinitely (MORAN 1975). The mean number of repeats does also not converge, giving two separated populations an infinite potential for divergence. The expected number of alleles actually present in a finite population, however, does converge to an equilibrium value n_a (KIMURA and OHTA 1975). Moreover, the variance in repeat number converges to the value $(2N - 1)\mu$ in a diploid population of size N (MORAN 1975). Following GOLDSTEIN *et al.* (1995a), this variance can be expressed in terms of the average squared difference in allele size,

$$D_0 = \sum_{i,j} f_i f_j (i - j)^2, \tag{3}$$

which is twice the variance in repeat number. Hence, D_0 has an equilibrium expectation that is approximately given by

$$\hat{D}_0 = 2\,\mu\,(2N-1) \approx 4N\mu. \tag{4}$$

(The small difference between the terms $2N$ and $2N - 1$ will henceforth be neglected.)

The infinite alleles version of the stepwise mutation model is mathematically convenient, but its properties are quite different from those of a more realistic finite alleles model (wandering distribution of allele sizes *vs.* convergence to a fixed limit distribution). Therefore we consider a version of the stepwise mutation model where the maximum number of alleles (M) is limited. Alleles are restricted to a range of sizes $\langle a+1, a+2, \ldots, a+M\rangle$ with alleles $A_1 = a+1, A_2 = a+2, \ldots, A_M = a+M$. The shortest allele A_1 will only mutate to A_2 and the longest allele A_M only to A_{M-1}, both with probability $\mu/2$. In our simulations we normally took $M = 10$, a value not uncommon in microsatellite studies (*e.g.*, BOWCOCK *et al.* 1994; GARZA *et al.* 1995; GOLDSTEIN *et al.* 1995a), but the value $M = 20$ will be considered as well.

It is easy to derive from (2) that for the finite alleles version of the stepwise mutation model the limit distribution induced by mutation pressure is the uniform distribution over the M alleles. Hence, in equilibrium the expected frequency of each allele is the same. At the limit distribution, the number of alleles is $n_a = M$ and the average squared difference in allele size is twice the variance of the uniform distribution:

$$\hat{D}_0 = \frac{M^2-1}{6}. \tag{5}$$

As stated above, random genetic drift may prevent convergence to the limit distribution. Whether and how well this distribution is approached, will not only depend on the population size and the mutation rate but also on the range of allele sizes M. In a small population, only a small number of alleles can be maintained. As long as only a fraction of all alleles is present, the allele frequency distribution will wander around and the infinite alleles version of the stepwise mutation model may provide a good approximation. Larger populations, however, can maintain all M alleles, and with increasing population size, the allele distribution will more and more tend toward the uniform limit distribution. In such a case, the infinite alleles version of the stepwise mutation model will not apply.

The question arises for which population sizes the infinite alleles version of the stepwise mutation model is adequate. To give an indication, we performed computer simulations of our finite alleles model for different population sizes. Figure 1 shows the number of alleles n_a and the average squared difference D_0 as a function of population size. The simulation results agree well with the predictions of the infinite alleles model for small population sizes: $N < 5000$ for n_a, and $N < 500$ for D_0. In a larger population, the infinite and the finite model give quite different results. n_a reaches its maximum M when $N = 10{,}000$ and D_0 approximates $(M^2-1)/6$ when $N = 100{,}000$.

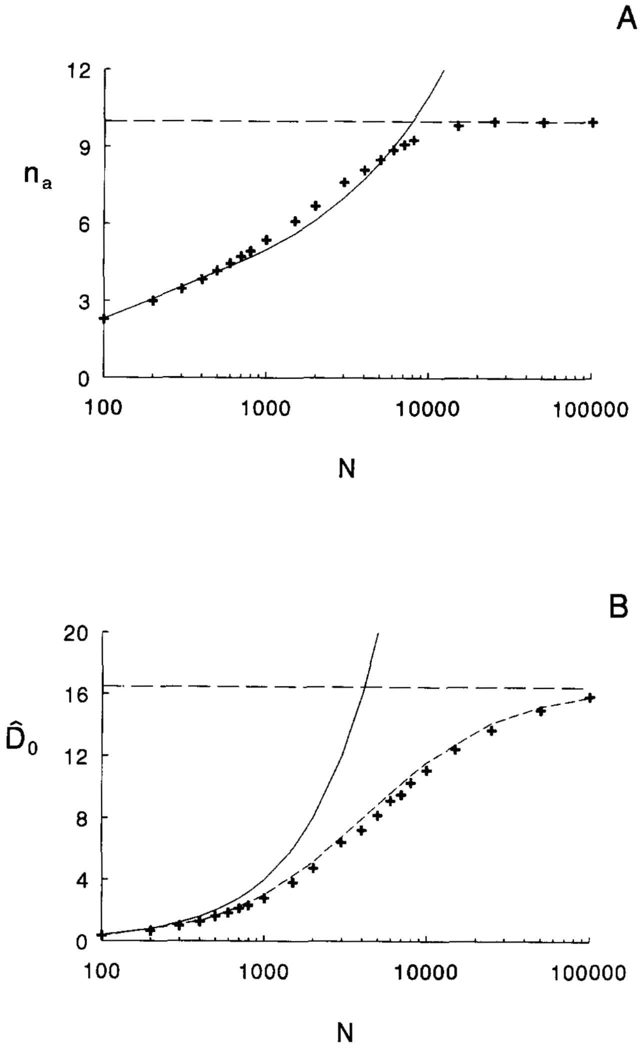

FIGURE 1.—A comparison between the finite and the infinite alleles version of the stepwise mutation model. (A) The mean number of alleles n_a and (B) the mean average squared difference in allele size $\hat{D}_0$ as a function of population size. The symbols (+) indicate the outcome of 1000 computer simulations based on the stepwise mutation model with $M = 10$ alleles and $\mu = 10^{-3}$. In this model, $n_a \leq M$ and $\hat{D}_0$ is limited by $(M^2-1)/6 = 16.5$. Solid lines correspond to the predictions of the infinite alleles version of the model (A, KIMURA and OHTA 1975; B, MORAN 1975). The dashed line in (B) indicates the prediction of the finite alleles model (see APPENDIX).

In the APPENDIX, it is shown how $\hat{D}_0$ can be calculated for the finite alleles model. Even without such a calculation, it is easy to see that the equilibrium prediction (4) of the infinite alleles model will be certainly misleading as soon as

$$N > \frac{M^2-1}{24\mu}. \tag{6}$$

In fact, for larger values of N, the prediction (4) would increase beyond the value (5) for a uniform distribution. Notice that (6) is quite restrictive. For $M = 10$ and $\mu = 10^{-3}$, for example, a large discrepancy with the infinite alleles model is already expected in moderately sized populations (for $N > 4125$).

CONSEQUENCES FOR GENETIC DISTANCE MEASURES

Several genetic distance measures have been developed to quantify differences in the allele frequency distribution between populations. These measures are, for example, used to estimate migration rates, divergence times, or phylogenetic relationships. Here we want to investigate to what extent the reliability of such estimates is affected by constraints on the range of allele sizes at a microsatellite locus. To this end, we focus on the most simple example, the genetic differentiation of two populations that resulted from a common parent population τ generations ago. Both the parent and the daughter populations have a constant size of N diploid (or equivalently $2N$ haploid) individuals. In our simulations, reproduction occurred in discrete nonoverlapping generations. A new generation was generated by sampling (with replacement) $2N$ alleles from the previous one. After reproduction, mutation took place according to the stepwise mutation model as described above. First, a parent population was simulated for $6N$ generations giving D_0 enough time to reach its equilibrium. Then, two daughter populations were formed by sampling twice independently $2N$ alleles from the parent population.

Allele sharing (D_{AS}) is a genetic distance measure that is based on allele frequencies only. It is defined as the probability that two alleles randomly chosen from both populations are not identical (*e.g.*, GOLDSTEIN *et al.* 1995a). Hence, with f_{ik} denoting the frequency of allele i in population k ($k = 1, 2$):

$$D_{AS} = \sum_{i \neq j} f_{i1} f_{j2} = 1 - \sum_{i} f_{i1} f_{i2}. \tag{7}$$

GOLDSTEIN *et al.* (1995a) argue that, for microsatellites, it is better to use a distance measure that takes account of the fact that alleles can be ordered according to size. They suggest to use the average squared difference in allele size between two populations:

$$D_1 = \sum_{i,j} f_{i1} f_{j2} (i - j)^2. \tag{8}$$

Similarly, SLATKIN (1995) proposes to replace the classical F_{ST} statistic (WRIGHT 1951), which is based on the variance in allele frequencies, by a new statistic R_{ST}. For the infinite alleles version of the stepwise mutation model, the R_{ST} statistic shows the same relationship to coalescence times as F_{ST} does in the classical model of arbitrary mutation. In terms of D_1 and D_0 and in case of two populations, R_{ST} can be represented as

$$R_{ST} = \frac{D_1 - \bar{D}_0}{D_1 + \bar{D}_0}, \tag{9}$$

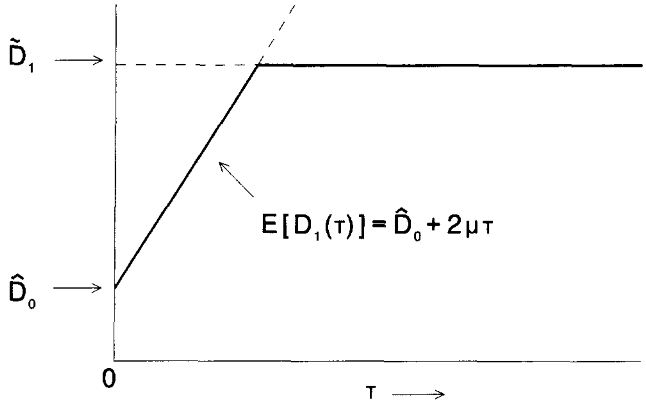

FIGURE 2.—Maximal range of linearity of the genetic distance measure D_1. At time $\tau = 0$, a haploid equilibrium population of size $2N$ is divided into two populations of size $2N$. In the infinite alleles model, D_1 is expected to increase linearly with τ according to $E[D_1(\tau)] = \hat{D}_0 + 2\mu\tau$, where $\hat{D}_0 = 4N\mu$. With a finite number of M alleles, however, D_1 will not increase beyond $\tilde{D}_1 = (M^2 - 1)/6$.

where $\bar{D}_0$ is the mean of the D_0s of the two diverging populations.

GOLDSTEIN *et al.* (1995a) argue that D_1 is a useful genetic distance measure since, in the infinite alleles version of the stepwise mutation model, D_1 is expected to increase linearly with divergence time τ. The slope of the function describing this linear relationship is equal to 2μ, independent of population size. The intercept is given by the value of D_0 at the moment of population subdivision. Hence,

$$E[D_1(\tau)] = D_0(0) + 2\mu\tau. \tag{10}$$

If we assume that the parent population was in equilibrium when separation occurred, $\hat{D}_0 = 4N\mu$ can be inserted for $D_0(0)$:

$$E[D_1(\tau)] = 4N\mu + 2\mu\tau. \tag{11}$$

It is important to realize that this equation strongly reflects the properties of the infinite alleles version of the stepwise mutation model. This is demonstrated by the following two observations.

First, Equation 11 describes an unbounded function. In contrast, a limited number of alleles implies that population divergence must be limited, too. This was also noticed by GOLDSTEIN *et al.* (1995a), who argue that the ultimate population divergence is given by

$$\tilde{D}_1 = \lim_{\tau \to \infty} D_1 = \frac{M^2 - 1}{6}. \tag{12}$$

As a consequence, Equations 10 and 11 can only be correct up to $\tilde{D}_1$. This means that the range of linearity of $E(D_1)$ must lie between $D_0(0)$ and $\tilde{D}_1$ (see Figure 2).

Second, as we have shown above (see Figure 1), the derivation of Equation 4 for D_0 in equilibrium crucially depends on the infinite alleles model. If the number of alleles is finite, $\hat{D}_0$ is not given by $4N\mu$. In fact, $\hat{D}_0$ is always

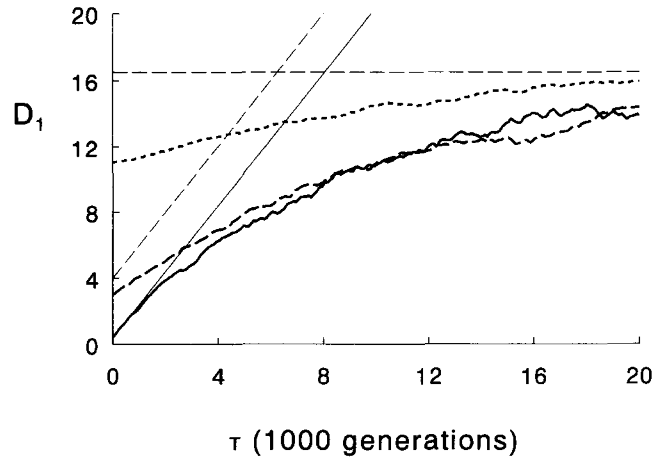

FIGURE 3.—Dynamics of the genetic distance measure D_1 in the finite alleles model. At $\tau = 0$, a haploid equilibrium population of size $2N$ was divided in two populations of size $2N$. The bold lines show the mean values of D_1 over 500 simulations (———, $2N = 200$; – – –, $2N = 2000$, · · ·, $2N = 20{,}000$). The thin straight lines correspond to the time course of D_1 as predicted by the infinite alleles model. In the finite alleles model, D_1 is limited by $\tilde{D}_1 = 16.5$. Notice that $D_1(0) \approx 0.4 = 4N\mu$ for $2N = 200$; $D_1(0) \approx 3 < 4N\mu$ for $2N = 2000$; and $D_1(0) \approx 11 \ll 4N\mu$ for $2N = 20{,}000$.

smaller than twice the variance of the uniform distribution [see (5)], and it approaches $\tilde{D}_1$ in large populations. (The APPENDIX shows how $\hat{D}_0$ can be calculated.)

Taken together, these arguments show that for finitely many alleles Equation 11 cannot be expected to be a good predictor of the behavior of D_1. As the range of linearity is restricted, the slope 2μ will not always offer a good description of the increase of D_1 in time. Moreover, Equation 5 shows that the intercept $4N\mu$ cannot be correct for large populations. Generally, the increase of D_1 will not be properly predicted by Equation 11 if the population size N is large, if the maximal number of alleles M is small, or if a long time has passed since population separation occurred.

Figure 3 illustrates these general observations. It is shown how D_1 changes with divergence time when the number of alleles is limited. Clearly, only for a very small population ($2N = 200$) and a relatively small number of generations ($\tau < 2000$) the time course of $D_1(\tau)$ is reasonably well described by Equation 11. In larger populations, the intercept $D_1(0)$ is smaller than $4N\mu$ and $D_1(\tau)$ increases with a much smaller slope than 2μ. As expected, $\tilde{D}_1$ is approached asymptotically in all cases. Interestingly, the increase in $D_1(\tau)$ is approximately linear even for relatively large population sizes (*e.g.*, $2N = 20{,}000$) and for a longer time than predicted by GOLDSTEIN *et al.* (1995a, Equation 4). This is possible since, with a restricted number of alleles, slope and intercept of $D_1(\tau)$ are both much smaller than in the infinite alleles model.

PRACTICAL IMPLICATIONS

Estimation of divergence time: One of the applications of genetic distance measures is the estimation of the number of generations τ that two populations are genetically isolated (*e.g.*, SLATKIN 1995). According to the infinite alleles version of the stepwise mutation model, this divergence time can be estimated by combining Equations 4 and 11 and eliminating μ:

$$E[D_1(\tau)] = \hat{D}_0 + \frac{\hat{D}_0}{2N}\tau, \qquad (13)$$

which leads to the estimator

$$\tau_{est} = 2N\left(\frac{D_1 - \bar{D}_0}{\bar{D}_0}\right). \qquad (14)$$

In terms of R_{ST}, this can be rewritten as:

$$\tau_{est} = 4N\frac{R_{ST}}{1 - R_{ST}}. \qquad (15)$$

SLATKIN (1995) derived the same estimator of the divergence time by a different method. By means of computer simulations based on an infinite alleles model, SLATKIN studied a pair of populations diverging at 100 microsatellite loci. He found that τ_{est} is a relatively unbiased estimator of the real divergence time τ and that for such a large number of loci the standard error of the estimate is small.

However, Equations 14 and 15 are based on Equations 4 and 11, which, as we have shown above, are not correct if the number of alleles is limited. One might therefore expect that, with constraints on allele size, the estimator τ_{est} is unreliable. To check this, we simulated two populations diverging at 15 microsatellite loci. (We chose a smaller number of loci than SLATKIN since rarely more than 10 or 20 loci are available in empirical studies.)

Our simulation results are illustrated by Figure 4. For 560 replicate pairs of populations, the estimated divergence time τ_{est} is compared with the true divergence time τ (dashed line). Although the mean value of the estimates $\overline{\tau_{est}}$ (solid line) slightly underestimates the real value τ, the estimator appears to be reasonably unbiased. However, there are considerable differences in the estimates between replicates. With a population size $2N = 200$, for example, at $\tau = 2000$, 90% of the estimates range from $\tau_{est} \approx 710$ to $\tau_{est} \approx 3740$ generations and 10% deviate even more. Hence, the standard error of the estimate is large, and it becomes even larger when a smaller number of loci is available (data not shown). Accordingly, τ_{est} appears to be an unbiased but unprecise estimator of divergence time.

In case of a large population (Figure 4B), the agreement between τ and $\overline{\tau_{est}}$ is quite surprising, as we have shown above that in this case Equation 11 is not correct. However, the estimator τ_{est} is not based on Equation 11 but on Equation 13. The fact that τ_{est} appears to be a relatively reliable estimator indicates that, in the context of a limited number of alleles, Equation 13 is applicable to a rather broad (but restricted) time horizon,

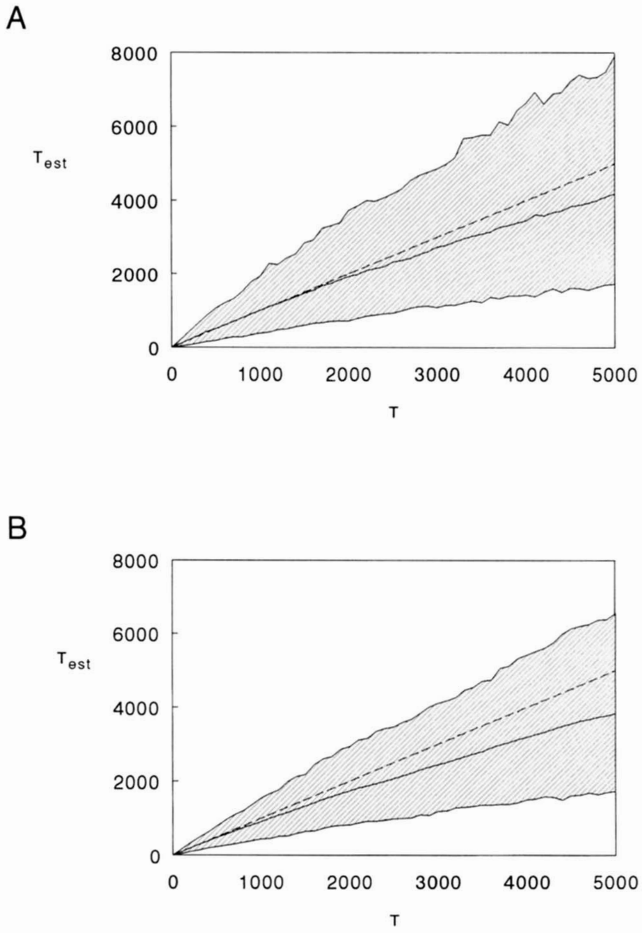

FIGURE 4.—Reliability of estimated divergence time τ_{est}. At $\tau = 0$, a haploid equilibrium population of size $2N$ was divided into two populations of size $2N$. In (A) $2N = 200$, and in (B) $2N = 20{,}000$. τ was estimated on the basis of Equation 14 by averaging D_1 and $\bar{D}_0$ over 15 loci. The dashed straight line corresponds to the real divergence time, the solid line to the mean estimate $\overline{\tau_{est}}$ over 560 simulations. The hatched area resulted by discarding the 5% largest and the 5% smallest estimates. Hence, this area corresponds to a 90% confidence interval of the estimate τ_{est} over 15 microsatellite loci.

whereas Equation 11 is certainly not applicable. Notice that Equation 13 is more general since it does not include a specific value for $\hat{D}_0$. In fact, $\hat{D}_0$ is not equal to $4N\mu$ but has to be calculated as indicated in the APPENDIX.

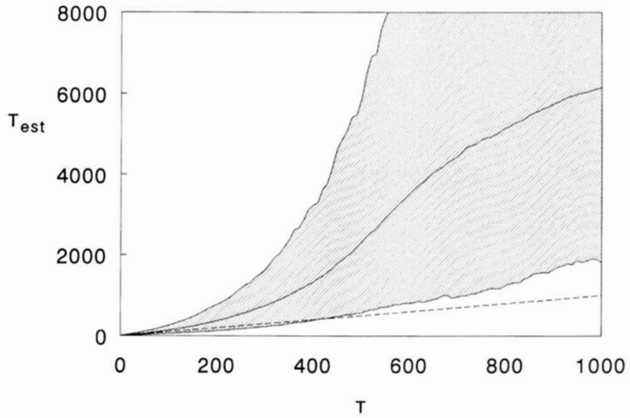

FIGURE 5.—The same situation as in Figure 4, but now two diverging populations of size $2N = 200$ originate from an equilibrium parent population of size $2N = 20{,}000$. Even in a short time perspective, τ_{est} is no longer a reliable estimator of τ.

Dependence on demographic history: Up to now we have assumed that the two diverging populations are of the same size as the parent population. This assumption simplifies the theoretical analysis, but it may be quite unrealistic. For example, population subdivision may result from the founding of small island populations from a large mainland population, or it may be associated with bottlenecks caused by external disturbance. Although a thorough study of such historical demographic scenarios is beyond the scope of this paper, we want to demonstrate that demographic events can be essential for the conclusions that are drawn from genetic distance measures. To illustrate this, we show that the conditions at the moment of population subdivision may have dramatic consequences for the estimation of divergence time.

We consider the simple scenario that two small diverging populations of constant size $2N = 200$ result from a large equilibrium population of size 20,000. Figure 5 shows that in this case Equations 14 and 15 no longer provide an unbiased estimator of divergence time. The differences between replicates are enormous and the mean value τ_{est} overestimates τ by a wide margin. Notice that after about only 400 generations, the true value τ does no longer fall within the 90% range of the estimates.

This discrepancy between τ and τ_{est} can be explained by the fact that D_0 does not remain constant when the population size changes. In the parent population of size $2N = 20{,}000$, $\hat{D}_0 \approx 11$ (see Figure 1). This value is much higher than the equilibrium value $\hat{D}_0 \approx 0.4$ of the daughter populations of size $2N = 200$. Hence D_0 will decrease from ~11 to ~0.4 in each daughter population, and Equation 14 will systematically overestimate the value of τ.

We conclude that the conditions at the time of population separation may be of utmost importance for the dynamics of genetic distance measures. Estimates of divergence time always reflect assumptions on the history of a population and other demographic events. However, although playing a crucial role, demographic history is typically not known in practice.

Reconstruction of phylogenetic relationships: GOLDSTEIN *et al.* (1995a) studied the reliability of D_1 and allele sharing D_{AS} for the reconstruction of phylogenetic relationships. To this end, they considered evolution along a three-taxon tree of variable length. In their simulations, an equilibrium population of $2N$ haploid individuals was divided in two populations of size $2N$. After b generations, one of the daughter populations was split again, and after b more generations the genetic

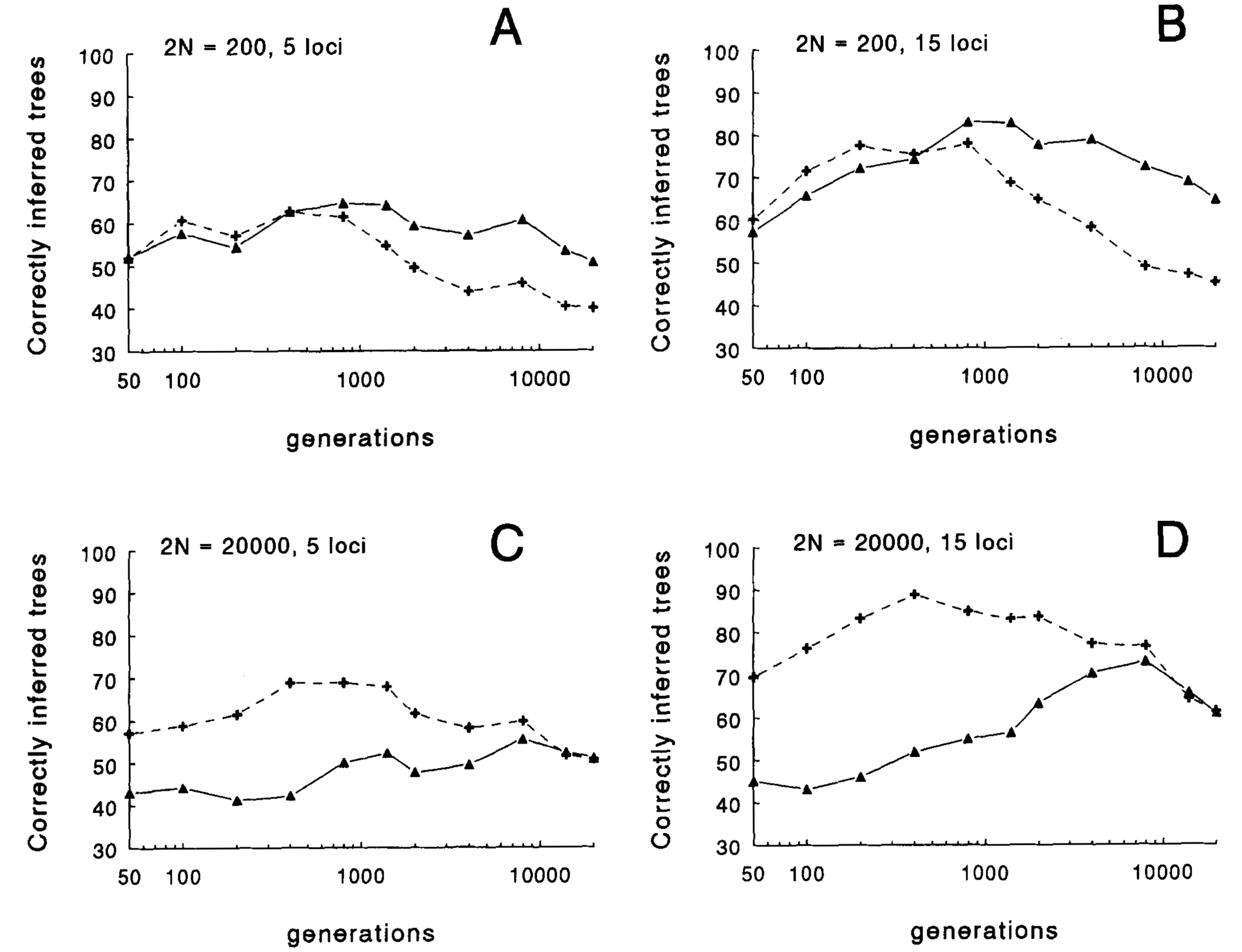

FIGURE 6.—Reliability of D_1 and D_{AS} in recovering the correct phylogeny. The percentage of correctly inferred phylogenetic trees using D_1 (▲) and D_{AS} (+) is plotted against tree length $2b$ on a logarithmic scale. Population sizes are $2N = 200$ (A and B) or $2N = 20,000$ (C and D), the number of loci is five (A and C) or 15 (B and D). B corresponds to Figure 2 of GOLDSTEIN *et al.* (1995a). The values shown are the results of 560 independent simulations. Notice that a random guess would infer 33% correct trees.

distances between the resulting three populations were compared. The proper phylogenetic relationship was inferred by assuming that the pair of populations with the smallest distance had been separated last. In a population with $2N = 200$, using 15 loci, GOLDSTEIN *et al.* found that the distance measure D_1 was superior to D_{AS} for $2b > 500$ generations.

To investigate the robustness of these results in the context of a limited number of alleles, we studied the same model as GOLDSTEIN *et al.* (1995a). However, as before, we restricted the number of alleles ($M = 10$) and we used different population sizes ($2N = 200$ and $2N = 20,000$). Moreover, we considered five loci as well as 15 loci. Figure 6 illustrates some of our results. Even for these highly simplistic trees the percentage of correctly inferred trees is quite low. In case of five loci, this percentage is only ~60%, while a random guess would render 33%. Notice that the estimates based on D_1 get less accurate in larger populations, as expected on the basis of our earlier arguments. The same arguments lead to the prediction that D_{AS}, too, should perform worse in larger populations. However, for the population sizes considered here, this was not the case. In contrast, the estimates based on D_{AS} were more accurate in the larger population. As a consequence, D_{AS} was superior to D_1 in the larger population.

We also studied a broader range of phylogenetic trees by looking at the percentage of proper estimation for trees with variable branch lengths. We defined b_1 as the number of generations between first and second branching and b_2 as the number of generations between second branching and the moment of measuring genetic distances. The results of our simulations are shown in Figure 7. Clearly, a large divergence time between the first and the second branching and a short time after the second branching ($b_1 \gg b_2$), renders a high percentage of correct estimates. On the other hand, if both population subdivisions took place shortly

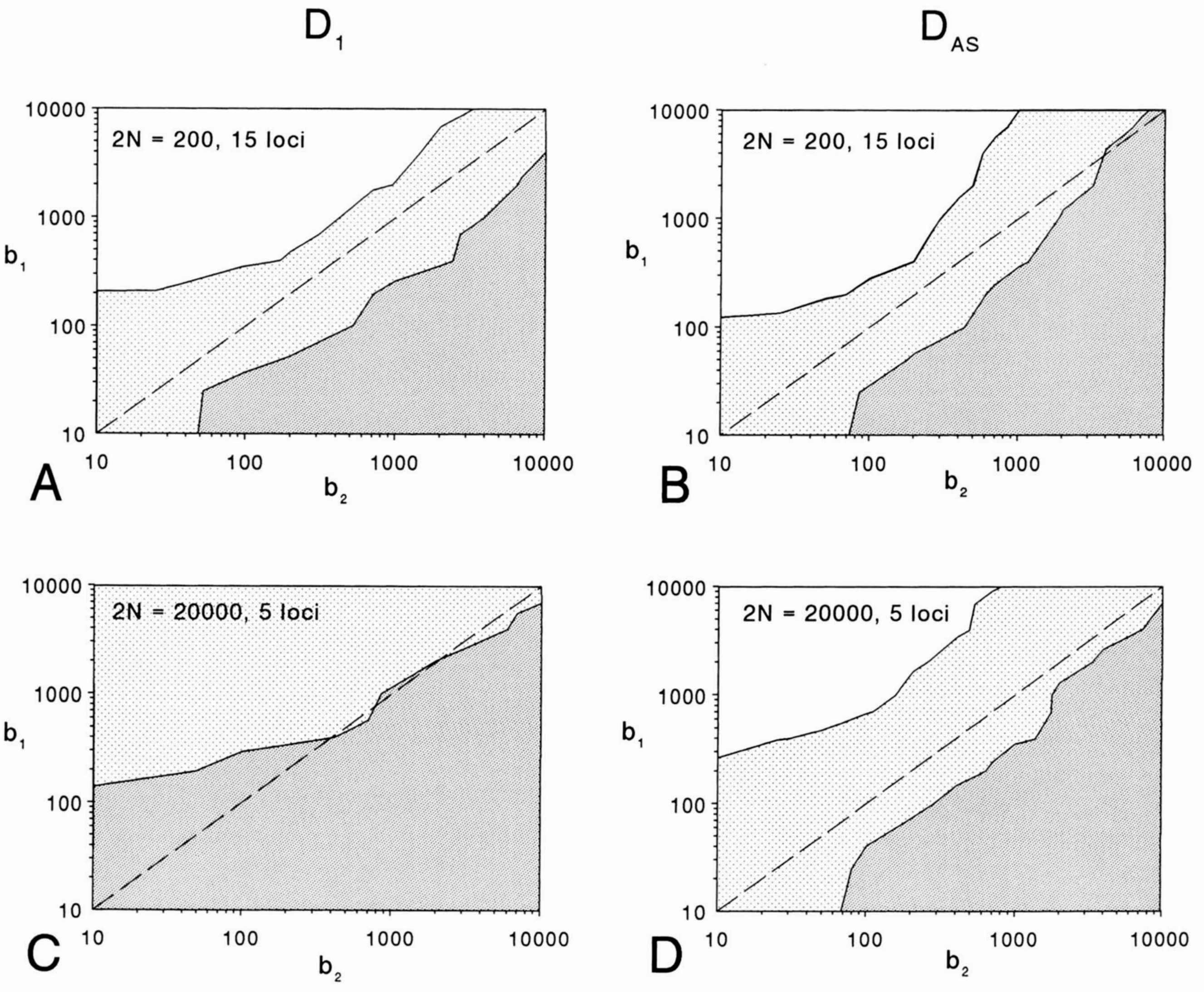

FIGURE 7.—Contour plots indicating the reliability of D_1 (A and C) and D_{AS} (B and D) in recovering the correct phylogeny for varying branch lengths, b_1 and b_2. A and B correspond to 15 loci and a population size $2N = 200$, C and D to five loci and $2N = 20{,}000$. In the white area, the percentage of correct estimates is high ($P > 95\%$), in the light gray area, it is intermediate ($50\% < P < 95\%$), in the dark area, it is low ($P < 50\%$). The dashed diagonal line corresponds to the case of equal branch lengths as in Figure 6.

after each other and the genetic distance was measured after a long time ($b_1 \ll b_2$), phylogeny reconstruction gets very unreliable. Again, the most striking result is the finding that D_1 is not a good estimator when the population size is large and the number of available microsatellite loci is small.

In contrast to the results of GOLDSTEIN *et al.* (1995a), it appears that even in a longer time perspective, D_{AS} is superior to D_1, at least in larger populations. GOLDSTEIN *et al.* explain their findings by their observation that D_{AS}, while being superior with respect to variance, is inferior to D_1 with respect to expectation. The latter may not be the case when the number of alleles is limited.

We have argued above that *all* genetic distance measures, including D_1 and D_{AS}, should perform worse in larger populations if the number of alleles is limited. Here we observed this effect for D_1, but not for D_{AS}. We expect, however, that the performance of D_{AS} will deteriorate for larger population sizes than those considered here.

EVALUATION OF MODEL ASSUMPTIONS

The simulations presented above were all based on a rather specific stepwise mutation model with a relatively small number ($M = 10$) of possible alleles and a relatively high mutation rate ($\mu = 10^{-3}$). Our results clearly illustrate the general point that constraints on allele size and the resulting possibility of back mutations have a homogenizing effect in that different populations are shifted toward the same limit distribution of alleles. It is, however, not our intention to advocate the model and the parameters chosen as the most adequate choice for studying mutation pressure at microsatellite loci. In our opinion, any such choice would be premature since at present the mutation process at microsatellite loci is only poorly understood (*e.g.*, Ellegren *et al.* 1995; Rubinsztein *et al.* 1995; Amos and Rubinsztein 1996). To get an impression of the robustness of our results, we will nevertheless consider different parameter values and an alternative model for the mutation process.

Dependence on M: Up to now, the maximal number of alleles was restricted to $M = 10$. In fact, about 10 alleles are quite often found in empirical studies. Moreover, a value of $D_0 = 20$ is quite typical for natural populations, indicating a value of $M \approx 11$ according to Equation 5 (D. Goldstein, personal communication). However, the actual number of alleles found will certainly underestimate the maximum M, and $\hat{D}_0$ will typically be smaller than $(M^2 - 1)/6$ in moderately sized populations (see Figure 1B). In other words, M might actually be larger than 10.

To study the consequences of a larger range of allele sizes, we repeated our simulations by putting the maximal number of alleles to $M = 20$. Some of the results are presented in Figures 8 and 9. Figure 8A shows that in large populations the equilibrium value of D_0 does again differ from the infinite-alleles expectation $\hat{D}_0 = 4N\mu$. In contrast to $M = 10$ (Figure 1B), the discrepancy is not yet present at $N\mu = 1$, but it becomes visible around $N\mu = 5$. Figure 8B shows that the time course of D_1 differs again significantly from the infinite-alleles expectation 11. Compared with $M = 10$ (Figure 3), however, the upper limit is much larger ($\hat{D}_1 = 66.5$), and it is not approached within 20,000 generations. Accordingly, D_1 increases almost linearly for a much larger number of generations. As a consequence, the bias in the estimate of the divergence time between two populations remains small, even in a somewhat longer time perspective (compare Figures 4 and 9). Still, however, the standard error of the estimate is very large and extreme care should be taken when applying such an unbiased but unprecise estimator in practice.

Dependence on μ: For a given set of microsatellite loci the mutation frequency μ is typically not known, and it may even be variable between and within loci. Although mutation rates ranging from 10^{-2} to 10^{-5} are often reported in the literature (Levinson and Gutman 1987; Jeffreys *et al.* 1988; Kelly *et al.* 1991; Edwards *et al.* 1992; Weber and Wong 1993), the value of 10^{-3} chosen in our simulations may be too high. The precise value of μ will be important for the rate of genetic differentiation: a higher value of μ increases the relative importance of mutation pressure and it therefore reduces the diversifying action of genetic drift.

The effect of the mutation rate on our simulation results is, however, highly predictable. All calculations as well as simulations with other values of μ indicate that it is only the product $N\mu$ that matters and not the mutation rate μ *per se*. Hence all results reported thus far for the mutation rate $\mu = 10^{-3}$ directly apply to the mutation rate $\mu = 10^{-4}$ if the population size N is replaced by $10N$. In particular, the dependence of the equilibrium value of D_0 on the population size in Figure 1B holds for all values of μ if the abscissa is relabeled in terms of $N\mu$, as in Figure 8A.

An alternative model: For illustrative purposes, we have modified the stepwise mutation model as little as possible, by just imposing strict limits upon the range of allele sizes. Of course, a fixed range of allele sizes

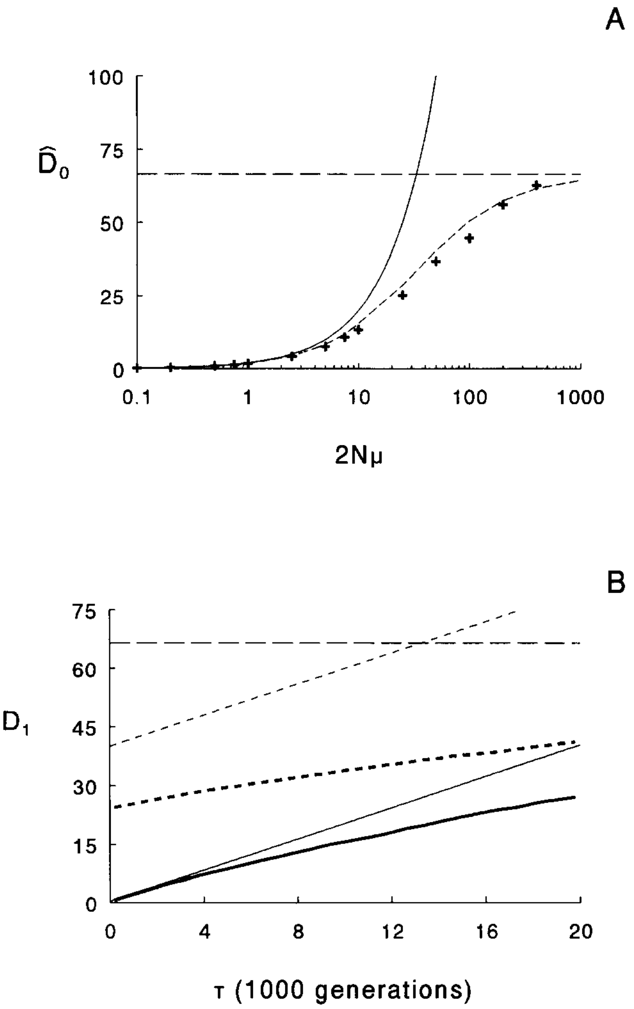

Figure 8.—The average squared differences in allele size within ($\hat{D}_0$) and between (D_1) populations when the maximal number of alleles is $M = 20$. (A) As in Figure 1B, the equilibrium value of $\hat{D}_0$ is plotted against population size (but notice that it is expressed in terms of $2N\mu$). +, the mean outcome of computer simulations based on the stepwise mutation model with $M = 20$ alleles; – – –, the analytical prediction (see appendix); and ———, the prediction of the infinite-alleles version of the stepwise mutation model. For $M = 20$, the limiting value of $\hat{D}_0$ is 66.5. (B) As in Figure 3, the change in D_1 is shown as a function of the time that has elapsed since the splitting of two populations of size $2N$ (———, $2N = 200$; – – –, $2N = 20{,}000$). The thin lines correspond to the predictions of the infinite alleles version of the stepwise mutation model, the bold lines represent the outcome of computer simulations based on the finite alleles version with $M = 20$. Notice that the limiting value $\hat{D}_1 = 66.5$ is not approached within 20,000 generations.

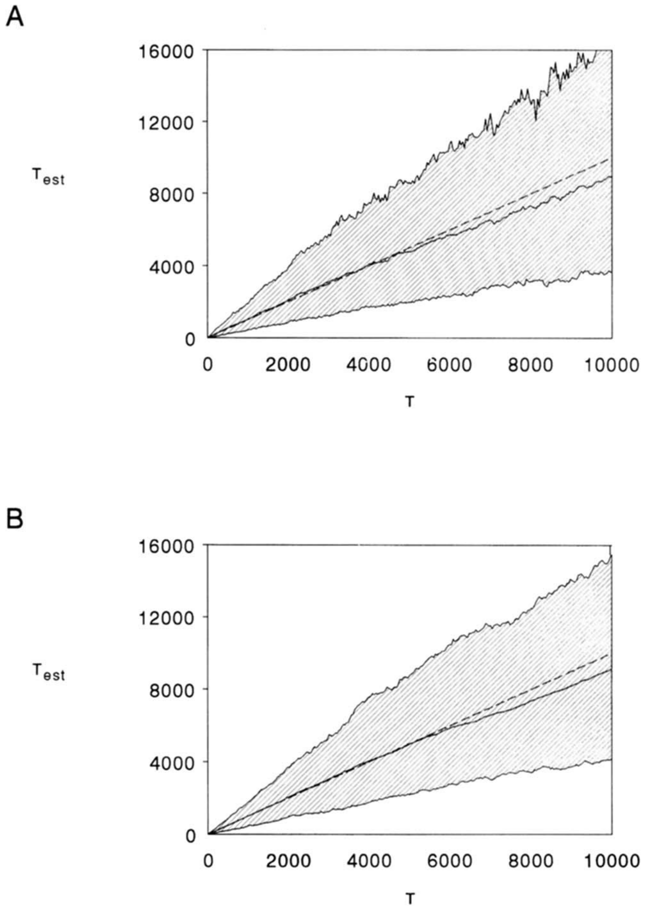

FIGURE 9.—Reliability of the estimator τ_{est} for $M = 20$. Compared with Figure 4, the range of divergence times τ (abscissa) is doubled. The mean estimate $\overline{\tau_{est}}$ (solid line) and its 90% confidence interval (hatched area) show that τ_{est} is less biased, but also less precise than for $M = 10$.

with constant mutation rates along the whole range can hardly be considered realistic. We would like to stress, however, that our main conclusions do not depend on the details of the model: the crucial point is that genetic differentiation due to genetic drift will be limited whenever the limit distribution of the mutation process has a finite variance.

To illustrate this general principle, we present another simple model, which basically allows an infinite range of allele sizes but which does not lead to a uniform limit distribution. Assume that—as in the "classical" stepwise mutation model—the probability for an "upward" mutation (increasing the length of the allele by one repeat) is constant and equal to $\nu = \mu/2$. In contrast, longer alleles have a higher probability to disintegrate than shorter alleles, *i.e.*, the probability δ_j of a "downward" mutation (reducing the length of an allele A_{j+1} by one repeat) is positively related with j.

Let us for simplicity assume that δ_j is directly proportional to $j\nu$

$$\delta_j = \frac{j}{\alpha}\,\nu \qquad (16)$$

where α is a constant of proportionality. Notice that upward mutations have a higher probability than downward mutations ($\nu > \delta_j$) whenever $j < \alpha$ while the opposite is true whenever $j > \alpha$. With these assumptions, it is easy to see that the limit distribution f^* of the mutation process satisfies the relation

$$f^*_{i+1} = \frac{\alpha}{i}\,f^*_i. \qquad (17)$$

Hence f^* is essentially a Poisson distribution:

$$f^*_{i+1} = \frac{\alpha^i}{i!}\,e^{-\alpha}$$

with mean allele size $E(i) = \alpha + 1$ and variance $\mathrm{var}(i) = \alpha$.

Notice that the model does not impose a limit on allele size, but that the variance is nevertheless limited. It is instructive to compare the behavior of this specific infinite-alleles version of the stepwise mutation model with that of the finite-alleles version considered earlier. Recall that the finite-alleles model has a uniform limit distribution over the M alleles, with variance $(M^2 - 1)/12$. To compare the finite-alleles model with $M = 10$ with the infinite-alleles model characterized by (16), we chose $\alpha = (M^2 - 1)/12 = 8.25$. With this value of α, we found that $\hat{D}_0$ has not only the same maximal value, but also that the increase of $\hat{D}_0$ with population size (Figure 1B) is almost identical in both models. The same holds true for other aspects of the models. For example, the estimation of divergence time leads to results that are indistinguishable from those in Figure 4. However, the number of alleles present may increase far beyond $M = 10$ in the model without limits on allele size.

The infinite-alleles model specified by (16) is probably not more realistic than the finite-alleles model considered earlier. However, the comparison of both models illustrates the important point that our results are not artifacts of a specific model, but consequences of far more general principles.

CONCLUSIONS

In view of their abundance and high degree of polymorphism, microsatellites are highly promising for analyzing the genetic and demographic structure of populations. However, a major drawback of microsatellites is the limited range of allele sizes combined with the high mutation rate. As a consequence, the potential for genetic divergence is limited, and genetic information specific for a population may easily be lost due to mutation. Already in moderately sized populations, the homogenizing force of mutation pressure can dominate the diversifying force of random genetic drift. Only in small populations and in a short term perspective, genetic differentiation at microsatellite loci is likely to occur.

To illustrate these general points, we have investigated a finite alleles version of the stepwise mutation model. Typically, we assumed that maximally $M = 10$ alleles occur at each locus, that the mutation frequency is $\mu = 10^{-3}$, and that only 15 microsatellite loci are available. We arrived at the following conclusions:

In small populations ($N < 500$, say), the constraints on allele size do not appear to be relevant. In large populations ($N > 5000$, say), however, the predictions of the finite alleles model differ significantly from those of infinite alleles model (Figure 1). For more general parameter combinations, Equation 6 gives an indication for the population size above which the infinite alleles approximation of the stepwise mutation model will become unreliable.

Only in small populations and in a short term perspective ($N < 1000$ and $\tau < 2000$, say), the increase of the genetic distance D_1 with divergence time is properly described by the linear Equation 11 of GOLDSTEIN *et al.* (1995a) (Figure 3). With a constrained number of alleles, the range of linearity of D_1 is always limited. For the time range where D_1 is approximately linear, the time change of D_1 is reasonably well described by Equation 13. In contrast to the infinite alleles model, $\hat{D}_0$ is smaller than $4N\mu$.

D_1 and R_{ST} can be used to estimate the divergence time of populations. The bias in the estimate is surprisingly small, even in relatively large populations ($2N = 20{,}000$, Figure 4). However, the standard error of the estimates is very large if the estimate is based on only a moderate number of microsatellite loci.

Genetic distance is highly sensitive to the demographic history of diverging populations. As a consequence, the reliability of estimates of population parameters will be strongly affected by bottlenecks or fluctuations in population size. For example, divergence time will be systematically overestimated if two divergent island populations resulted from a larger mainland population (Figure 5).

Average squared distance D_1 and allele sharing D_{AS} both perform badly in the reconstruction of phylogenetic relationships (Figures 6 and 7). Remarkably, D_{AS}, a distance measure only based on allele frequency and not on allele size, may be superior to D_1, especially in large populations and in a (relatively) short time perspective. In any case, many microsatellite loci (much more than 15, say) are required to correctly infer a given phylogenetic relationship. This finding is also supported by the results of ZHIVOTOVSKY and FELDMAN (1995), who study the distribution of genetic distances in an infinite alleles version of the stepwise mutation model.

Qualitatively, our conclusions do not depend on the specific choice of the model parameters M and μ. However, if more than about $M = 10$ allele sizes are feasible and/or if the mutation rate is smaller than $\mu = 10^{-3}$, notable discrepancies between the finite- and the infinite-alleles version of the stepwise mutation model only occur at larger population sizes.

Our results do not reflect specific aspects of the model assumptions but rather general principles underlying the interplay of genetic drift and mutation. In fact, virtually all results were qualitatively and quantitatively reproduced by a stepwise mutation model with infinitely many alleles but inhomogeneous mutation rates. The homogenizing force of mutation pressure gets important whenever mutation tends to shift the population to a limit distribution with finite variance.

Summarizing we conclude that genetic distance measures based on the infinite alleles version of the stepwise mutation model may perform fairly well in a short-term perspective and in the context of small populations of constant size. Beyond this context microsatellite data should be treated with care since there is a considerable danger of misinterpretation.

We thank KUKE BIJLSMA, MICHIEL VAN BOVEN, ODILIA VELTEROP, LOUIS VAN DE ZANDE, and particularly DAVID GOLDSTEIN for helpful comments on the manuscript. These investigations were supported by the Life Science Foundation (SLW), which is subsidized by the Netherlands Organization for Scientific Research (NWO).

LITERATURE CITED

AMOS, W., and D. C. RUBINSZTEIN, 1996 Microsatellites *are* subject to directional evolution. Nat. Genet. (in press).

BOWCOCK, A., A. RUIZ LINARES, J. TOMFORDHE, E. MINCH, J. R. KIDD *et al.*, 1994 High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368:** 455–457.

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.

ELLEGREN, H., C. R. PRIMMER and B. C. SHELDON, 1995 Microsatellite evolution: directionality or bias in locus selection? Nat. Genet. **11:** 360–362.

GARZA, J. C., M. SLATKIN and N. B. FREIMER, 1995 Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol. Biol. Evol. **12:** 594–603.

GOLDSTEIN, D. B., A. RUIZ LINARES, L. L. CAVALLI-SFORZA and M. W. FELDMAN, 1995a An evaluation of genetic distances for use with microsatellite loci. Genetics **139:** 463–471.

GOLDSTEIN, D. B., A. RUIZ LINARES, L. L. CAVALLI-SFORZA and M. W. FELDMAN, 1995b Genetic absolute dating based on microsatellites and the origin of modern humans. Proc. Natl. Acad. Sci. USA **92:** 6723–6727.

JEFFREYS, A. J., N. J. ROYLE, V. WILSON and Z. WONG, 1988 Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. Nature **322:** 278–281.

KELLY, R., M. GIBBS, A. COLLICK and A. J. JEFFREYS, 1991 Spontaneous mutation at the hypervariable mouse minisatellite locus Ms6-hm: flanking DNA sequence and analysis of germline and early somatic events. Proc. R. Soc. Lond. Ser. B Biol. Sci. **245:** 235–245.

KEMENY, J. G., and J. L. SNELL, 1976 *Finite Markov Chains.* Springer, Berlin.

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.

LEVINSON, G., and G. A. GUTMAN, 1987 Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4:** 203–224.

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

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.

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

TAUTZ, D., 1993 Notes on the definition and nomenclature of tandemly repetitive DNA sequences, pp. 21–28 in *DNA Fingerprinting: State of the Science*, edited by S. D. J. PENA, R. CHAKRABORTY, J. T. EPPLEN and A. J. JEFFREYS. Birkhäuser Verlag, Basel.

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

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

WRIGHT, S., 1951 The genetical structure of populations. Ann. Eugenics **15:** 323–354

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

Communicating editor: W. J. EWENS

APPENDIX

In this APPENDIX we show how to calculate the expected equilibrium value of the average squared difference in allele size, $\hat{D}_0$, in case of the M-allele version of the stepwise mutation model. We make use of a method developed by OHTA and KIMURA (1973). It is easy to see that D_0 can be expressed in the form

$$D_0 = \sum_i i^2 C_i \qquad \text{(A1)}$$

where C_i is the sum of the product of the frequencies of alleles i repeats apart:

$$C_i = \sum_j f_j f_{j+i}. \qquad \text{(A2)}$$

Hence, to calculate $\hat{D}_0$, it is sufficient to derive the M values $C_0, C_1, \ldots, C_{M-1}$ in equilibrium. OHTA and KIMURA (1973) show for the infinite alleles version of the stepwise mutation model that the C_i satisfy the following differential equations:

$$\frac{dC_0}{dT} = 4N\mu[C_1 - C_0] + 1 - C_0 \qquad \text{(A3a)}$$

$$\frac{dC_i}{dT} = 2N\mu[C_{i-1} + C_{i+1} - 2C_i] - C_i \quad \text{for} \quad i > 0. \qquad \text{(A3b)}$$

(Time is measured in units of $2N$ generations, so $T = t/2N$.)

With a finite number of M alleles, i ranges from 0 to $M - 1$. Moreover, some mutations assumed in the derivation of (A3) do not occur since the allele of minimal (maximal) size cannot mutate to an allele of smaller (larger) size. Consider, for example, the $M - i$ pairs of alleles that are i repeats apart (as in C_i). Then, mutations that enlarge the allele size cannot occur at one of the alleles in each of the two pairs containing the smallest and the largest allele. Now assume that the allele frequencies are uniformly distributed, as in the limit distribution. Then, mutation to a larger allele is not possible in a fraction $1/(M - i)$ of those alleles joined together in the parameter C_i. As a consequence, system (A3) has to be changed to:

$$\frac{dC_0}{dT} = 4N\mu\left[C_1 - \left(1 - \frac{1}{M}\right)C_0\right] + 1 - C_0 \qquad \text{(A4a)}$$

$$\frac{dC_i}{dT} = 2N\mu\left[\left(1 - \frac{1}{M - i + 1}\right)C_{i-1} + C_{i+1} - \left(2 - \frac{1}{M - i}\right)C_i\right] - C_i$$

$$\text{for} \quad 0 < i < M - 1 \qquad \text{(A4b)}$$

In equilibrium, all dC_i/dT are equal to 0, yielding a system of M linear equations for the unknowns C_i. Insertion of the solution of this system into (A1) yields the equilibrium value of D_0. Figures 1B and 8A show that computer simulations are in good agreement with this prediction.