Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 11549–11552, December 1995
Evolution

Microsatellite variability and genetic distances

(stepwise-mutation model/multiple loci/moments of distances/diffusion approximation/size of mutations)

LEV A. ZHIVOTOVSKY* AND MARCUS W. FELDMAN†

*Vavilov Institute of General Genetics, Russian Academy of Sciences, 3 Gubkin Street, Moscow, 117809, Russia; and †Department of Biological Sciences, Stanford University, Stanford, CA 94305

Communicated by Paul R. Ehrlich, Stanford University, Stanford, CA, August 7, 1995

ABSTRACT **We analyze the within- and between-population dynamics of the distribution of the number of repeats at multiple microsatellite DNA loci subject to stepwise mutation. Analytical expressions for moments up to the fourth order within a locus and the variance of between-locus variance at mutation-drift equilibrium have been obtained. These statistics may be used to test the appropriateness of the one-step mutation model and to detect between-locus variation in the mutation rate. Published data are compatible with the one-step mutation model, although they do not reject the two-step model. Using both multinomial sampling and diffusion approximations for the analysis of the genetic distance introduced by Goldstein *et al.* [Goldstein, D. B., Linares, A. R., Cavalli-Sforza, L. L. & Feldman, M. W. (1995) *Proc. Natl. Acad. Sci. USA* 92, 6723–6727], we show that this distance follows a χ^2 distribution with degrees of freedom equal to the number of loci when there is no variation in mutation rates among the loci. In the presence of such variation, the variance of the distance is obtained. We conclude that the number of microsatellite loci required for the construction of phylogenetic trees with reliable branch lengths may be several hundred. Also, mutations that change repeat scores by several units, even though extremely rare, may dramatically influence estimates of population parameters.**

1. Introduction

Microsatellite DNA is a special class of tandem repeat loci that involves a base motif of 1–6 bp repeated up to ≈100 times (1). Microsatellite loci are extremely polymorphic, with up to dozens of alleles at each locus and mutation rates as high as 10^{-3} (2, 3) or 10^{-4} (4, 5), and for these reasons, they are appropriate for use in molecular taxonomy, evolution, and population genetics (6). In population and evolutionary genetics, microsatellites are powerful because, besides the frequencies of the alleles, the repeat score for an allele may be viewed as a quantitative trait. This approach is reminiscent of quantitative genetics and already has been used to estimate the central moments of the number of repeats in human populations (7) and to evaluate genetic distances between populations (8, 9) and the corresponding F statistics (10). Except for these studies and earlier basic work in which the population distribution of allele sizes and their second moments under the one-step mutation model and random drift was analyzed (11, 12), little has been published about higher order moments. We study here the between-locus variation, in particular the distribution of the distances, including central moments of up to the fourth order.

2. Within-Population Variability

2.1. Mutation. Consider a mutation process operating on repeat numbers under which each allele may mutate to any other allele. Let $c = j - i$ be the change in repeat number due to mutation of the allele that carries j repeats to the allele with i repeats; $c < 0$ if it decreases, $c > 0$ if it increases. We assume that the probability, v_c, of a mutational change by c in the number of repeats does not depend on the allele mutated. The total mutation rate is $v = \Sigma_{c\neq 0} v_c$. Hereafter, we assume that mutation is not directional, in the sense that the mean change in repeat numbers among newly arisen mutations is zero; $\bar{v} = \Sigma_c v_c c = 0$.

The one- and two-step stepwise mutation models were generalized (10) by introducing the value $v\sigma_m^2$, where σ_m^2 is the variance of changes in repeats among the new mutations. For simplicity we denote it by the symbol w;

$$w = \sum_c v_c c^2 = v\sigma_m^2.$$

The one-step stepwise mutation scheme (studied in refs. 8–12) is a special case of this mutation model, with $w = v$. If there are mutational events that change repeat numbers by two or more, then $w > v$. For example, for the two-step model, $w = v + 3(v_2 + v_{-2})$, and for the three-step model, $w = v + 3(v_2 + v_{-2}) + 8(v_3 + v_{-3})$.

It also is useful to introduce the skewness and kurtosis of the mutation process, which we denote respectively by $s = \Sigma_c v_c c^3$, and $k = \Sigma_c v_c c^4$, assuming that $\bar{v} = 0$. Note that $k = w$ for the one-step mutation scheme, otherwise $k > w$. In particular, $k = w + 12(v_2 + v_{-2})$ for the two-step model, and $k = w + 12(v_2 + v_{-2}) + 72(v_3 + v_{-3})$ for the three-step model.

2.2. Population Parameters. At time t, let p_i be the frequency of allele A_i that carries i repeats. Following refs. 8 and 9, we may analyze the variation at this locus as a quantitative trait, the number of repeats. To this end, introduce the first four central sample moments of the allele frequency distribution $\{p_i\}$—namely, the mean of repeat numbers r, the variance V, the skewness S, and the kurtosis K:

$$r = \sum_i ip_i, \qquad V = \sum_i p_i(i-r)^2,$$
$$S = \sum_i p_i(i-r)^3, \quad K = \sum_i p_i(i-r)^4. \qquad [1]$$

Note that the expected squared difference between the scores of two alleles randomly drawn from the population is $2V$.

Following mutation, the expected frequency of allele A_i changes to $\tilde{p}_i$,

$$\tilde{p}_i = \sum_c v_{i-c}p_c. \qquad [2]$$

It follows from Eqs. **1** and **2** that, among gametes produced by the parental generation, the first four central moments of the distribution $\{\tilde{p}_i\}$ are

$$\tilde{r} = r, \tilde{V} = w + V, \tilde{S} = s + S, \tilde{K} = k + K + 6wV. \qquad [3]$$

2.3. Mutation-Gene Drift Equilibrium. The progeny generation comprises N haploid individuals (gametes) randomly chosen from among those produced by the parents, with

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

frequencies $\tilde{p}_i$ from Eq. **2**. Following ref. 12, this multinomial sampling procedure reduces $\tilde{V}$ by the factor $1 - 1/N$. Therefore, the expectation of the mean and variance of repeat numbers (see Eq. **1**) in the progeny generation, r' and V' respectively, must satisfy

$$\mathcal{E}_m\{r'\} = r,\ \mathcal{E}_m\{V'\} = \left(1 - \frac{1}{N}\right)(V + w), \qquad [4]$$

where $\mathcal{E}_m$ denotes the expectation operator with respect to multinomial sampling, and the primes refer to the allele frequency distribution at time $t + 1$.

From Eq. **4**, taking the expectation with respect to the initial generation and iterating, we see that ultimately the expected population variance approaches its mutation-drift equilibrium,

$$\hat{V} = \mathcal{E}\{V\} = (N - 1)w, \qquad [5]$$

where ε is the expectation operator with respect to the initial generation (9, 10, 12). We use the circumflexes ("hats") to label expected values at equilibrium.

Hereafter, we assume that w, s, and k are of the order of $1/N$, with $\zeta = (N - 1)s$ and $\kappa = (N - 1)k$.

For the skewness and kurtosis of the allele frequencies distribution, from Eqs. **3** and **4** and equations 27.4.2 and 27.5.1 of ref. 13, we have

PROPOSITION 1. *Among progeny*

$$\mathcal{E}_m\{S'\} \approx \left(1 - \frac{3}{N}\right)(S + s),$$

$$\mathcal{E}_m\{K'\} \approx K + \frac{1}{N}(-4K + 6\hat{V}V + \kappa + 6V^2), \qquad [6]$$

$$\mathcal{E}_m\{V'^2\} \approx V^2 + \frac{1}{N}(K - 3V^2 + 2\hat{V}V),$$

neglecting terms $\mathbb{O}(N^{-2})$.

Expressions for the central moments of the mean number of repeats expected among progeny follow from the moments of the mean value (ref. 13, p. 345):

$$\mathcal{E}_m\{r'\} = r,\ \mathcal{E}_m\{(r' - r)^2\} = \frac{1}{N}(w + V),\ \mathcal{E}_m\{(r' - r)^3\} = \frac{1}{N^2}(s + S),\ \mathcal{E}_m\{(r' - r)^4\} = \frac{3V^2}{N^2} + \mathbb{O}(N^{-3}). \qquad [7]$$

From Eq. **6**, taking expectations as for Eq. **5**, we obtain equilibria for the skewness, kurtosis, and variance of variances, $\widehat{\mathrm{Var}}\{V\} = \mathcal{E}\{V^2\} - \hat{V}^2$, in the population which we state as

RESULT 1. *The expected values of the central moments in a population at mutation-drift equilibrium are:*

$$\hat{V} = (N - 1)w,\ \widehat{Var}\{V\} \approx \frac{1}{3}\left(4\hat{V}^2 + \frac{\kappa}{2}\right),$$

$$\hat{S} \approx \frac{\zeta}{3}, \qquad \hat{K} \approx 5\hat{V}^2 + \frac{\kappa}{2}. \qquad [8]$$

The value of κ may significantly exceed $\hat{V}$ and thus influence the expected values of $\widehat{\mathrm{Var}}\{V\}$ and $\hat{K}$ if the mutation process allows multiple steps. It follows from the definition that

$$\kappa = \hat{V}\frac{\sum_c f_c c^4}{\sum_c f_c c^2}, \qquad [9]$$

where $f_1, f_2, f_3, \ldots$ are the fractions of new mutations that change the number of repeats by ± 1, ± 2, ± 3, etc., with $\Sigma_c f_c = 1$. Obviously, $f_c = v_c/v$. Formulae similar to those of *Result 1* for $\hat{V}$ and $\widehat{\mathrm{Var}}\{V\}$ have been independently derived (14) by using a coalescence argument with additional assumptions on the mutation probabilities.

2.4. Multiple Loci. Let p_{ij} be the frequency of a gamete carrying i repeats at one locus and j repeats at a second, with $p_{i.}$ and $p_{.j}$ the corresponding marginal (allele) frequencies at these loci:

$$p_{i.} = \sum_j p_{ij}, \quad p_{.j} = \sum_i p_{ij}.$$

Suppose that mutations at different loci arise independently, so that the mutation process may be described in terms of $v_{1,c}$ and $v_{2,c}$, the rates of mutation that change the repeat numbers by c at locus 1 and locus 2, respectively. The expected gamete frequencies are

$$\tilde{p}_{ij} = \sum_c \sum_h p_{ch} v_{1,i-c} v_{2,j-h} \text{ (after mutation)},$$

$$\tilde{p}^r_{ij} = (1 - R)\tilde{p}_{ij} + R\tilde{p}_{i.}\tilde{p}_{.j} \text{ (after recombination)}. \qquad [10]$$

Analogously to the variance in Eq. **1**, define the covariance between these loci with respect to the number of repeats in the parental generation as:

$$C_{12} = \sum_i \sum_j (i - r_1)(j - r_2)p_{ij}, \qquad [11]$$

where r_1 and r_2 are the means at loci 1 and 2, respectively. Also define the covariance $\tilde{C}^r_{12}$ after mutation and recombination but before sampling by the same Eq. **11** in which p_{ij} are replaced by $\tilde{p}^r_{ij}$. It follows from Eqs. **10** and **11** that $\tilde{C}^r_{12} = (1 - R)C_{12}$. Now, as before, let the "prime" denote the progeny generation, and $\mathcal{E}_m$ be the expectation in this generation with respect to a multinomial sample of size N from the gamete frequencies $\tilde{p}^r_{ij}$ (see Eq. **10**). In particular, $\mathcal{E}_m\{p'_{ij}\} = \tilde{p}^r_{ij}$. Then we can prove that $\mathcal{E}_m\{C'_{12}\} = (1 - R)(1 - 1/N)C_{12}$, so that $\mathcal{E}\{C_{12}\}$ is zero at mutation-drift equilibrium:

PROPOSITION 2. *The mean repeat numbers at different loci are not correlated with each other at mutation-drift equilibrium.*

It can also be proved that the variances at these loci, $V_1 = \Sigma_i(i - r_1)^2 p_{i.}$ and $V_2 = \Sigma_j(j - r_2)^2 p_{.j}$, have negligible correlation at mutation-drift equilibrium: $\mathcal{E}\{V_1V_2\} - \hat{V}_1\hat{V}_2 = \mathbb{O}(N^{-1})$, where $\hat{V}_1 = \mathcal{E}\{V_1\} = (N - 1)w_1$, $\hat{V}_2 = \mathcal{E}\{V_2\} = (N - 1)w_2$ are the variances for these two loci at equilibrium. We conjecture that this covariance is actually $\mathbb{O}(N^{-2})$. Therefore, correlation among variances of repeat numbers at different microsatellite loci can be ignored in the statistical analysis. Thus if $V_1, V_2, \ldots, V_L$ are variances at L microsatellite loci calculated according to Eq. **1**, with mean $\bar{V} = \Sigma V_l/L$, we may consider the between-locus variance of the variances,

$$\widehat{\mathrm{Var}}_L\{V\} = \frac{1}{L - 1}\left[\sum_{l=1}^{L} V_l^2 - L\bar{V}^2\right], \qquad [12]$$

as an estimate of the expected value $\widehat{\mathrm{Var}}\{V\}$ and obtain

RESULT 2. *In a population at mutation-drift equilibrium,* $\widehat{Var}\{V\}$ *given in Eq.* **8** *provides the predicted variance of repeat-score variances across loci,* $Var_L\{V\}$, *if the mutation rates do not vary among the loci.*

3. Genetic Distance Between Populations

3.1. Single Locus. In this section, let r_X, r_Y, and V_X, V_Y be the means and variances of the repeat scores at one locus in two genetically isolated populations labeled by X and Y. In ref. 9, the quantity $(r_X - r_Y)^2$ was denoted by $(\delta\mu)^2$ and used as a

distance between two populations. In the analysis here, for notational ease, we use the symbol Δ for this distance and call it the "squared mean difference" (SMD),

$$\text{SMD:}\Delta = (r_X - r_Y)^2. \quad [13]$$

The new notation is more convenient for the study of higher moments. From Eq. **7** we immediately obtain the expected value of Δ', the distance between these populations in the progeny generation (after mutation and multinomial sampling):

$$\mathcal{E}_m\{\Delta'\} = \Delta + \frac{1}{N}(V_X + w) + \frac{1}{N}(V_Y + w). \quad [14]$$

Suppose that both populations were derived from a single ancestral population of size N that was at mutation-drift equilibrium and let these populations also be of size N and at the same equilibrium. Then, from Eqs. **5** and **14**, we obtain

$$\mathcal{E}_m\{\Delta'\} = \Delta + 2w, \quad [15]$$

so that $\mathcal{E}\{\Delta\} = 2wt$, where, as before, $\mathcal{E}$ refers to expectation with respect to the initial generation (9, 15). Further, $\mathrm{Var}_m\{\Delta'\} = \mathcal{E}_m\{\Delta'^2\} - (\mathcal{E}_m\{\Delta'\})^2$. Decomposing $(r'_X - r'_Y)^4$ and using Eq. **7**, we can show that $\mathcal{E}_m\{\Delta'^2\} \approx \Delta^2 + 12w\Delta + 20w^2 + k/N$, and, using Eq. **15**, obtain a formula for the variance of the distance given that in the previous generation,

$$\mathrm{Var}_m\{\Delta'\} \approx 8w\Delta + 16w^2 + \frac{k}{N} \approx 8w\Delta + \mathcal{O}(N^{-2}). \quad [16]$$

To find the distribution of genetic distances after t generations, we neglect terms of the order of N^{-2} and use a diffusion approach (ref. 16, pp. 177–179). Under the assumption that $Nw \to \beta$ as $N \to \infty$, we may approximate the discrete changes in Δ by a diffusion process with the infinitesimal parameters $\mu = 2\beta$, the drift coefficient; and $\sigma^2(\Delta) = 8\beta\Delta$, the variance coefficient; and time $\tau = t/N$, where t is the real time in generations. Since $|\Delta' - \Delta| = |[(r'_X - r_X) - (r'_Y - r_Y)]\,[(r'_X + r_X) - (r'_Y + r_Y)]| < \text{const}\,|(r'_X - r_X) - (r'_Y - r_Y)|$ on some finite interval of Δ, and since the fourth central moment of r' is $\mathcal{O}(N^{-2})$ (Eq. **7**), $N\mathcal{E}\{(\Delta' - \Delta)^4\}$ approaches zero as N increases. Therefore, this process satisfies the Dynkin condition and is a diffusion (ref. 16, p. 165). The corresponding transition probability density function $p(\Delta|\tau N, x)$ satisfies the Kolmogorov backward differential equation

$$\frac{\partial p}{\partial \tau} = \frac{1}{2}\sigma^2(x)\frac{\partial^2 p}{\partial x^2} + \mu\frac{\partial p}{\partial x}, \quad [17]$$

where $p(\Delta|\tau N, x)$ is the probability density of the distance Δ at time $t = \tau N$ given $\Delta = x$ at $t = 0$. The transformation $\Delta = 2\beta y^2$ transforms the standard Brownian process, described by the heat equation $\partial p/\partial t = (1/2)\,\partial^2 p/\partial y^2$, to Eq. **17** (ref. 16, p. 173). Since the Brownian motion process gives rise to the Gaussian distribution (e.g., ref. 16, p. 217), this transformation produces a solution in terms of natural time, t (in generations), and the mutation parameter, w. In fact, we have

PROPOSITION 3. *Let two independent populations have the distance between them* $\Delta_0 = (r_X - r_Y)^2$ *at the initial time* t = 0, *and be at mutation-drift equilibrium. Then the probability density function of the distance* Δ *between them after* t *generations is*

$$p(\Delta|t,\Delta_0) = \frac{1}{\sqrt{\Delta}}\frac{1}{\sqrt{4\pi wt}}\exp\left\{-\frac{(\sqrt{\Delta} - \sqrt{\Delta_0})^2}{4wt}\right\}. \quad [18]$$

The following consequence of *Proposition 3* is useful for application:

RESULT 3. *Suppose that two populations split from an ancestral population at mutation-drift equilibrium and remain at this equilibrium. Then, after* t *generations of independent evolution,* $\Delta/2wt$ *has a* χ^2 *distribution with* 1 *degree of freedom.*

A related result, that $r_X - r_Y$ is normally distributed with variance $2wt$, has been obtained independently (15).

Define the variance of the distance with respect to the initial generation: $\mathrm{Var}\{\Delta\} = \mathcal{E}\{\Delta^2\} - [\mathcal{E}\{\Delta\}]^2$.

COROLLARY 1. *Under these conditions, the expected distance,* $\mathcal{E}\{\Delta\}$, *and its variance,* $Var\{\Delta\} = \mathcal{E}\{[\Delta - 2wt]^2\}$, *after* t *generations are*

$$\mathcal{E}\{\Delta\} = wt, \; Var\{\Delta\} \approx 8w^2t^2. \quad [19]$$

This result for the mean in Eq. **19** has been obtained (9) for the one-step mutation model.

Note. *Proposition 3*, and therefore *Result 3*, hold if the population size is sufficiently large that the diffusion provides a good approximation to the discrete time dynamic for Δ (16). This may require a population of several hundred individuals. Nevertheless, Eq. **19** is valid for any N and can be derived directly from Eqs. **15** and **16**, with $\mathrm{Var}\{\Delta\} = 8w^2t^2 + (8w^2 + k/N)t$, which can be approximated by $8w^2t^2$ after sufficient time.

3.2. Arbitrary Number of Loci. The theory given above may be extended to include multiple microsatellite loci. Consider L microsatellite loci, with mutation rates $v_1, v_2, \ldots, v_L$ and weighted mutation rates $w_1, w_2, \ldots, w_L$ whose mean and variance are denoted by

$$\bar{w} = \frac{1}{L}\sum_{l=1}^{L} w_l, \; \sigma_w^2 = \frac{1}{L}\sum_{l=1}^{L} w_l^2 - \bar{w}^2. \quad [20]$$

Let $\Delta_1, \Delta_2, \ldots, \Delta_L$ be SMD distances between two populations for these loci. It can be proved that for any $m \neq l$, $N\mathcal{E}\{(\Delta'_l - \Delta_l)(\Delta'_m - \Delta_m)\}$ approaches zero as N increases. Hence, we may write the backward Kolmogorov equation for the multidimensional distribution of the distances, $p(\Delta_1, \ldots, \Delta_L|t, x_1, \ldots, x_L)$, with the initial conditions $x_1, \ldots, x_L$, and no covariance terms (ref. 17, p. 332) as:

$$\frac{\partial p}{\partial t} = \frac{1}{2}\sum_{l=1}^{L}\sigma_l^2\frac{\partial^2 p}{\partial x_l^2} + \sum_{l=1}^{L}\mu_l\frac{\partial p}{\partial x_l}, \quad [21]$$

where

$$\mu_l = 2\beta_l, \; \sigma_l^2 = 8\beta_l x_l. \quad [22]$$

This equation has a solution that is the product of marginal probability densities,

$$p(\cdot) = p_1(\Delta_1|t, \Delta_{01})p_2(\Delta_2|t, \Delta_{02})\ldots p_L(\Delta_L|t, \Delta_{0L}), \quad [23]$$

each of which takes the form of Eq. **18** with w and Δ_0 replaced by w_l and Δ_{0l}, respectively, where Δ_{0l} is the initial distance between these populations at locus l at time $t = 0$. Thus, we have

PROPOSITION 4. *At mutation-drift equilibrium, the SMD distances at microsatellite loci are weakly correlated with each other, and the distribution of each takes the form in Eq.* **18**. *In particular, if at* t = 0 *the populations separated from an ancestral population at equilibrium, then each ratio* $\Delta_l/2w_lt$ *follows the* χ^2 *distribution with* 1 *degree of freedom.*

Define the total distance between two populations as the sum of the single-locus distances,

$$\Delta = \sum_{l=1}^{L}\Delta_l. \quad [24]$$

From *Proposition 4* and the additivity of χ^2, we immediately obtain

RESULT 4. *Suppose that two populations are descended from a common ancestor at mutation-drift equilibrium* t *generations ago.* If the mutation rates are the same for all loci, then $\Delta/2wt$ approximately follows the χ^2 distribution with L degrees of freedom.

We emphasize that even if mutation rates differ among loci, from *Proposition 4* or from *Proposition 2* with *Corollary 1*, we have

COROLLARY 2. *The expected total distance* (*Eq.* **24**) *between populations* t *generations after they split and its variance at mutation-drift equilibrium are*

$$\mathcal{E}\{\Delta\} = 2L\bar{w}t, \quad Var\{\Delta\} \approx 8L(\bar{w}^2+\sigma_w^2)t^2. \qquad [25]$$

4. Discussion

We have provided here an analytical approach that uses both multinomial sampling and a diffusion approximation to obtain analytical formulas for the variance, kurtosis, genetic distance, etc., that confirm the results of the simulation study for the one- and two-step mutation models (table 1 of ref. 7) that used the coalescent process.

Consider the data on 86 microsatellite loci in humans listed in table 4 of ref. 7. [These data were taken from Centre d'Étude du Polymorphisme Humain (Paris) families and so cannot really be considered to represent an isolated randomly mating population of constant size.] Here, the variance of allele sizes, averaged over the loci, was $\bar{V} \approx 4.88$. Taking this value as the equilibrium variance $\hat{V}$, we can use *Results 1* and *2* to calculate the expected values of the variance of within-locus variances and the kurtosis. Assuming the one-step mutation model with equal mutation rates at the loci, for which $\kappa = \hat{V}$ ($f_1 = 1$ and $f_c = 0$ for all $c > 1$, in Eq. **9**), these values are $\widehat{\mathrm{Var}}\{V\} \approx 32.6$ and $\hat{K} \approx 121.5$, which should be compared to the values estimated from data, namely $\mathrm{Var}_L\{V\} \approx 31.8$ and the mean of the kurtosis estimates across the loci, $\bar{K} = \Sigma_l/L \approx 125.2$, respectively. Such close correspondence between data and theoretical estimates is rare in statistical applications of this kind. It confirms the previous conclusions that the one-step model satisfactorily describes these data and analogous conclusions of refs. 7 and 18. This does not mean, however, that other models are rejected by these data. For instance, if we assume a two-step model with 90% one-step and 10% two-step mutations ($f_1 = 0.9, f_2 = 0.1$, so that $\kappa = 1.92\hat{V}$; see Eq. **9**), then $\widehat{\mathrm{Var}}\{V\}$ and $\hat{K}$ from Eq. **8** become 33.4 and 123.7, respectively, which are still close to the estimated values. Therefore, such a two-step mutation process is not excluded by this analysis; many models may be compatible with these data.

We emphasize that the term κ, which characterizes the fourth central moment for the number of repeats among new mutations, may contribute to the estimates of the population parameters. Even very rare multiple-step mutations may dramatically increase the within-locus kurtosis, K, and the between-locus variance of variances in repeat numbers, $\mathrm{Var}_L\{V\}$. For instance, suppose all new mutations change the repeat numbers by 1 except for a fraction 0.0001, which changes repeat scores by 10 ($f_1 = 0.9999, f_{10} = 0.0001$, in terms of Eq. **9**). Then $\kappa \approx 96\hat{V}$. Thus, the statistics $\mathrm{Var}(V)$ and the kurtosis K may be particularly useful in helping to distinguish the one-step from multiple-step models. Indeed, from *Result 1*, the greater the fraction of mutations that change allele sizes by 2 or more repeats, the bigger is κ and thus the bigger is the kurtosis and also the between-locus variance of variances. Of course, an appropriate statistical procedure is needed for more formal inferences.

In the above discussion, we have concentrated on the squared mean difference, $(\delta\mu)^2$ [in the analysis above we use the symbol Δ for $(\delta\mu)^2$]. We have found that this distance follows a χ^2-distribution with L degrees of freedom, if it is based on L microsatellite loci and if the mutation rates do not vary among microsatellite loci. With heterogeneity in the mutation rates, we have obtained Eq. **25** for its variance. Using *Corollary 2* of *Result 4*, we can estimate how many loci are required for reliable estimates of the divergence time and thus for the construction of phylogenetic trees. Indeed, let t be the actual time passed after two populations split, $(\delta\bar{\mu})^2$ be the estimate of genetic distance between these populations, and $\bar{t}$ be the estimate of the divergence time. From Eq. **25**, $\bar{t} = (\delta\bar{\mu})^2/2L\bar{w}$, where $\bar{w}$ is the average of w values. It follows from Eq. **25** that the standard deviation of the estimated time $\bar{t}$ is $t\sqrt{2/L}$, which increases linearly with time, so that the earlier diverged taxa are estimated with a greater absolute error. The relative error in the estimate of the divergence time between two taxa, calculated as the ratio of the standard deviation of the estimated time and the real divergence time, $\sqrt{\mathrm{Var}\{t\}}/t$, does not depend on time and is just $\sqrt{2/L}$. Thus, to produce a relative error of 10% requires 200 loci. Therefore, reliable estimates of divergence time and the branch lengths in phylogenetic trees involving many taxa may require several hundred loci. Such a conclusion does not appear to be a specific property of microsatellite data and applies to other distances. This may contribute to frequent failure to predict time of divergence obtained from molecular data (ref. 19, pp. 508–514).

We are indebted to an anonymous reviewer for a most careful analysis of the assumptions involved in our treatment. These comments significantly improved the manuscript. This work was partly supported by National Institutes of Health Grants GM28016 and GM28428.

1. Tautz, D. (1993) in *DNA Fingerprinting: State of the Science*, eds. Pena, S. D. J., Chakraborty, R., Epplen, J. T. & Jeffreys, A. J. (Birkhäuser, Basel), pp. 21–28.
2. Jeffreys, A. J., Royle, N. J., Wilson, V. & Wong, Z. (1988) *Nature (London)* **322,** 278–281.
3. Kelley, R., Gibbs, M., Collick, A. & Jeffreys, A. F. (1991) *Proc. R. Soc. London B* **245,** 235–245.
4. Levinson, G. & Gutman, G. A. (1987) *Mol. Biol. Evol.* **4,** 203–221.
5. Henderson, S. T. & Petes, T. D. (1992) *Mol. Cell. Biol.* **12,** 2749–2757.
6. Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R. & Cavalli-Sforza, L. L. (1994) *Nature (London)* **368,** 455–457.
7. Valdes, A. M., Slatkin, M. & Freimer, N. B. (1993) *Genetics* **133,** 737–749.
8. Goldstein, D. B., Linares, A. R., Cavalli-Sforza, L. L. & Feldman, M. W. (1995) *Genetics* **139,** 463–471.
9. Goldstein, D. B., Linares, A. R., Cavalli-Sforza, L. L. & Feldman, M. W. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 6723–6727.
10. Slatkin, M. (1995) *Genetics* **139,** 457–462.
11. Ohta, T. & Kimura, K. (1973) *Genet. Res.* **22,** 201–204.
12. Moran, P. A. P. (1975) *Theor. Popul. Biol.* **8,** 318–330.
13. Cramer, G. (1946) *Mathematical Methods of Statistics* (Princeton Univ. Press, Princeton, NJ), 9th Ed.
14. Roe, A. (1994) Ph.D. dissertation (Queen Mary College, London, U.K.).
15. Garza, J. C., Slatkin, M. & Freimer, N. B. (1995) *Mol. Biol. Evol.*, in press.
16. Karlin, S. & Taylor, H. M. (1981) *A Second Course in Stochastic Processes* (Academic, New York).
17. Feller, W. (1966) *An Introduction to Probability Theory and Its Applications* (Wiley, New York), Vol. 2.
18. Shriver, M. D., Jin, L., Chakraborty, R. & Boerwinkle, E. (1993) *Genetics* **134,** 983–993.
19. Hillis, D. M. & Moritz, C. (1990) in *Molecular Systematics*, eds. Hillis, D. M. & Moritz, C. (Sinauer, Sunderland, MA), pp. 502–515.