THEORETICAL POPULATION BIOLOGY **50**, 325–344 (1996)
ARTICLE NO. 0034

Statistics for Microsatellite Variation Based on Coalescence*

Jonathan K. Pritchard and Marcus W. Feldman

Department of Biological Sciences, Stanford University, Stanford, California 94305

Received September 1995

The stepwise mutation model, which was at one time chiefly of interest in studying the evolution of protein charge-states, has recently undergone a resurgence of interest with the new popularity of microsatellites as phylogenetic markers. In this paper we describe a method which makes it possible to transfer many population genetics results from the standard infinite sites model to the stepwise mutation model. We study in detail the properties of pairwise differences in microsatellite repeat number between randomly chosen alleles. We show that the problem of finding the expected squared distance between two individuals and finding the variance of the squared distance can be reduced for a wide range of population models to finding the mean and mean square coalescence times. In many cases the distributions of coalescence times have already been studied for infinite site problems. In this study we show how to calculate these quantities for several population models. We also calculate the variance in mean squared pairwise distance (an estimator of mutation rate $\times$ population size) for samples of arbitrary size and show that this variance does not approach zero as the sample size increases. We can also use our method to study alleles at linked microsatellite loci. We suggest a metric which quantifies the level of association between loci—effectively a measure of linkage disequilibrium. It is shown that there can be linkage disequilibrium between partially linked loci at mutation–drift equilibrium.

INTRODUCTION

In recent years, microsatellites have become prominent among the types of genetic markers used for studying phylogenetic relationships among closely related individuals or populations. The principal reason for this popularity is that there are often large numbers of alleles at a single locus, even within a population. Microsatellites, also know as VNTRs (variable numbers of tandem repeats), are a class of tandem repeat, in which a short DNA segment (usually from two to five bases) is repeated up to about 100 times (Tautz, 1993). The character which is scored experimentally is the

* This research was supported in part by NIH Grant GM 28428 to M. W. Feldman. J. K. Pritchard is also supported by a Howard Hughes Medical Institute Predoctoral Fellowship. The authors are also grateful to Professor Peter Donnelly for alerting them to the unpublished thesis by Adrian Roe.

0040-5809/96 $18.00

number of repeats. The typically high levels of diversity stem from an exceedingly high mutation rate (recently estimated to be 5.6×10^{-4} for a series of 15 microsatellite loci in humans by Weber and Wong, 1993). The mutational process in microsatellites seems to involve slippage of the DNA strands during replication, usually by a single repeat (Schlötterer and Tautz, 1992). This suggests that the mutation process might be modelled by a stepwise model that was originally proposed (Ohta and Kimura, 1973) for electrophoretic variants, although generally regarded as inappropriate for that purpose (Nei, 1987). The stepwise mutation property has the consequence that much of standard population genetics theory is not immediately applicable to microsatellite data.

Formal analysis of the mathematics and statistics of the stepwise model, begun by Ohta and Kimura (1973) and Moran (1975) has recently been applied to data on microsatellite polymorphisms in human populations reported by Valdes *et al.* (1993) and Shriver *et al.* (1993). Two approaches have been taken. The first uses coalescent theory and might be called a "backward approach" since it focuses on the history of lineages within and between populations (Valdes *et al.*, 1993; Di Rienzo *et al.*, 1994; Garza *et al.*, 1995; Slatkin, 1995), while the second, a "forward approach," uses recursions based on the dynamics of the multinomial sampling process (Goldstein *et al.* 1995a,b; Zhivotovsky and Feldman, 1995). This paper extends the theory based on the coalescent approach.

Coalescent modeling (see reviews in Hudson, 1990; Hudson, 1993; Tajima, 1993) takes advantage of the property that when homologous DNA sequences from different individuals are sampled, they must share a common ancestor. The time elapsed since the most recent common ancestor of a pair of homologous sequences can be viewed as a random variable. The exact distribution of this random variable depends on the precise nature of the biological model in question. It is also possible to study the genealogy for a sample of m randomly chosen sequences; here both the topology of the tree and the times of each of the nodes are random. Significantly, the genealogy only reflects ancestry, and variation among the sequences can be thought of as being superimposed onto a given genealogical tree. The variation among the sequences results from mutations which are considered to be "sprinkled" onto a genealogy according to a Poisson process. The number of mutations which occur on each branch of the tree is Poisson distributed, with the mean number proportional to the branch length. Given a particular genealogy, it is straightforward to obtain expected values for many of the useful measures of variation. Roughly speaking, these expected values must be averaged over all possible genealogies, weighted by the likelihood of each genealogy.

An important characteristic of the coalescent approach is that it is generally based on the idea of studying the properties of a limited sample,

rather than the whole population. This means that coalescent models are often useful for the analysis of experimental data. The coalescent approach is also valuable as a framework for conducting extremely rapid Monte Carlo simulations (cf. Hudson, 1990) of neutral evolution at the within- or between-populations level.

In the past, coalescent modeling has mainly been applied to infinite sites or infinite allele models (but see also Valdes *et al*., 1993; Di Rienzo *et al*., 1994; Garza *et al*., 1995; Slatkin, 1995). The goal of this paper is to extend the coalescent analysis in the framework of the stepwise mutation model to the analysis of higher order and mixed moments that may be useful for the statistical analysis of experimental data.

One possible area of application is in the analysis of the distance measures used for comparisons between populations. Goldstein *et al*. (1995a, b) have used a pair of related distance measures. The first measure computes for a locus the average over all pairs of individuals of the squared difference in repeat score between them. This distance is labelled D_0 when all alleles come from a single population, and D_1 when the comparisons are between alleles drawn from two distinct populations. The second distance (called $(\delta\mu)^2$) is for a comparison of samples from two populations. In this case the distance is the square of the difference between the mean repeat number in each population. While the results of Zhivotovsky and Feldman (1995) are of greatest utility in the analysis of $(\delta\mu)^2$, the results in the present study apply to D_0 and D_1, with $(\delta\mu)^2 = D_1 - D_0$.

Coalescent Model for Comparison of two Sequences

In this paper, we consider a random-mating population composed of N haploid (or equivalently $N/2$ diploid) individuals. Unless otherwise stated, there is no gene flow from outside populations. Throughout this paper, we will also adopt a strict stepwise mutation model, with symmetric mutation probabilities, in which an allele with i repeats can mutate only to $i-1$ or $i+1$ repeats, each with probability $\mu/2$. Furthermore, it is assumed that the repeat number can range over all the integers. As pointed out by Moran (1975), in a stepwise mutation model the distribution of repeat scores does not converge, but the variance in repeat number within a population does. Consequently we will be interested in the differences between individuals, rather than the actual repeat numbers *per se*.

In our model, the difference in repeat number between two individuals depends on three random processes which will be denoted s, n, and t, each of which represent random variables. The three variables are defined as follows:

1. t: The time since the most recent common ancestor of the two microsatellite sequences. This time is usually taken to have an exponential

probability distribution, with a mean of N generations ($1/N$ is the probability of two individuals having a common ancestor in the previous generation in a random mating population (see Tajima, 1983, for further discussion)).

2. n: the total number of mutations which have occurred on the two lineages in the time t since the common ancestor. We assume that mutations occur according to a Poisson process, at the rate μ per lineage per generation. Then, conditional on t, n will be Poisson distributed, with mean $2\mu t$.

3. s: the difference in repeat number between the two individuals. This difference s can be thought of as the distance travelled from the origin in a symmetric random walk of n steps, with equal probability of moving up or down at each step.

In fact, the distribution of s depends on n, which in turn depends on t; consequently we write s as $s(n(t))$. Using $s(n(t))$, it is fairly straightforward to calculate the moments of the distribution of distances between a randomly chosen pair of individuals. It will then be possible to extend this to more complex population structures, to samples of more than two individuals, and to pairs of linked microsatellite loci.

MOMENTS OF $s(n(t))$ IN A SINGLE UNSTRUCTURED POPULATION

The moment generating function for the distance s travelled in a symmetric random walk of n steps is given in Révész (1990, p. 13) as

$$M(x) = E[e^{xs}] = \left[\frac{e^x + e^{-x}}{2}\right]^n.$$

We can calculate the expected value of s^m by differentiating the expression m times with respect to x and then setting $x = 0$. As expected from considerations of symmetry, we have $E[s|n] = E[s^3|n] = 0$. It turns out that $E[s^2|n] = n$ and $E[s^4|n] = 3n^2 - 2n$. We will also need the first and second moments of n, which is Poisson distributed with mean $2\mu t$ ($E[n|t] = 2\mu t$ and $E[n^2|t] = 2\mu t(1 + 2\mu t)$) and of t, which is exponentially distributed ($E[t] = N$ and $E[t^2] = 2N^2$). Then it is possible to calculate the expected values of $s^2(n(t))$ and $s^4(n(t))$ by conditioning s on n and n on t:

$$E[s^2(n(t))] = \int_0^\infty \sum_{n=1}^{\infty} \sum_{s=-\infty}^{\infty} s^2 \Pr[s|n(t)] \Pr(n|t) \Pr(t)\, dt$$

$$= \int_0^\infty \sum_{n=0}^{\infty} n \Pr(n|t) \Pr(t)\, dt$$

$$= \int_0^\infty 2\mu t \Pr(t)\, dt$$

$$= 2\mu E[t] \tag{1}$$

$$= 2\mu N \tag{2}$$

and

$$E[s^4(n(t))] = \int_0^\infty \sum_{n=0}^{\infty} \sum_{s=-\infty}^{\infty} s^4 \Pr[s \mid n(t)] \Pr(n \mid t) \Pr(t)\, dt$$

$$= \int_0^\infty \sum_{n=0}^{\infty} (3n^2 - 2n) \Pr(n \mid t) \Pr(t)\, dt$$

$$= \int_0^\infty (2\mu t + 12\mu^2 t^2) \Pr(t)\, dt$$

$$= 2\mu E[t] + 12\mu^2 E[t^2] \tag{3}$$

$$= 2\mu N + 24\mu^2 N^2. \tag{4}$$

Now, with the second and fourth moments of s in hand, we can easily compute the variance of s^2:

$$\mathrm{Var}[s^2(n(t))] = E[s^4] - E[s^2]^2$$

$$= 2\mu E[t] + 12\mu^2 E[t^2] - 4\mu^2 E^2[t] \tag{5}$$

$$= 2\mu N + 20\mu^2 N^2. \tag{6}$$

Note that the expected squared distance (2) is twice the expectation of the equilibrium variance of the frequency distribution which was first obtained by Moran (1975). Slatkin (1995) used coalescent modeling to derive (2), while Goldstein *et al.* (1995) calculated the expectation of their D_0 (which is the same as $E[s^2]$) as $2\mu(N-1)$, based on a forward recursive model. Zhivotovsky and Feldman (1995) derived results using the forward recursive approach from which (6) can be obtained after making appropriate approximations. Both Slatkin (1995) and Goldstein *et al.* (1995) noted that the expected squared distance $E[s^2]$ grows linearly in time and, therefore, proposed that the observed average squared difference is the natural distance metric for use in phylogenetic reconstruction based on microsatellite data. Consequently, this study will focus on the expectation and variance of s^2.

Observe that Eqs. (1), (3), and (5) were derived without reference to the distribution of coalescence times. This means that for any model of population structure in which $E[t]$ and $E[t^2]$ is known, the moments of s can be

quickly calculated by substitution into (1) and (3), and the variance of s^2 by substitution into (5). For each of the population models which follow, we will calculate the expectation and variance of s^2.

Note that in the terminology of Goldstein *et al.* (1995a), we have just calculated the expected value of D_0, and also the variance of D_0 for a sample of two alleles. In the next section we do the same for D_1 (the distance between alleles in separate populations.) Later in this paper we also calculate the variance of D_0 for an arbitrary sample size.

MOMENTS OF $s(n(t))$ IN DIVIDED POPULATIONS

Here we calculate the second moment of $s(n(t))$ (i.e., D_1) and the variances of D_1 for two different models of divided populations. The first model involves a radiation of two populations from a single ancestral population with no subsequent migration. The second model examines two populations at migration-mutation-drift equilibrium.

Branching Populations; No Migration

In the first model, we consider two populations of haploid size N which branched from a single ancestral population (also of size N) at a time τ generations before the present. There is no migration between the populations following the split.

Under this model, the patterns of ancestry within a subpopulation are exactly as for the single population above. This is because each subpopulation is of size N following the split, and the ancestral population is also of size N. Therefore, the moments for comparisons within a subpopulation are given by (2) and (4), and the variance by (6).

The coalescence between subpopulations takes place entirely before the split, τ generations ago. Thus, the coalescence time t for comparisons between subpopulations is distributed as $(\mathrm{Exp}[1/N] + \tau)$, where $\mathrm{Exp}[1/N]$ refers to the exponential distribution with parameter $1/N$. Then it follows that $E[t] = N + \tau$ and $E[t^2] = 2N^2 + 2N\tau + \tau^2$. Substituting in (1) and (5), we have

$$E[s^2(n(t))] = 2\mu(N + \tau), \tag{7}$$

$$\mathrm{Var}[s^2(n(t))] = 2\mu(N + \tau) + 4\mu^2(5N^2 + 4N\tau + 2\tau^2). \tag{8}$$

Separate Populations at Migration-Drift Equilibrium

The second model is for two populations, each of size N, with a probability v ($v \neq 0$) per generation of each individual migrating from one population to the other. Here it is supposed that the populations have

been separated for a great length of time, so that they are at migration–mutation–drift equilibrium (i.e., the rate of divergence due to drift and mutation is exactly balanced by the rate of convergence due to migration). The ancestors of the sampled genes can move from one population to the other and only have the opportunity to coalesce when they are in the same population. Slatkin (1991) provided an analysis of expected coalescence times under such a model of population subdivision. Hey (1991) calculates both the mean and variance of coalescence times, using a continuous time Markov chain. His results are, strictly speaking, derived for haploid systems, but the adjustment for diploid organisms should be negligible. Hey's results provide the relevant moments of t. For comparisons within a subpopulation, he has $E[t] = 2N$ and $E[t^2] = 8N^2 + N/v$. Comparing between subpopulations, $E[t] = 2N + 1/2v$ and $E[t^2] = 8N^2 + 3N/v + 1/2v^2$. Substituting into (1) and (5), we derive the expectation of s^2 and its variance for comparisons between alleles chosen from the same subpopulation:

$$E[s^2(n(t))] = 4\mu N, \tag{9}$$

$$\mathrm{Var}[s^2(n(t))] = 4\mu N + 4\mu^2(20N^2 + 3N/v). \tag{10}$$

For comparisons of two individuals, one from each population, the results are

$$E[s^2(n(t))] = 2\mu(2N + 1/2v), \tag{11}$$

$$\mathrm{Var}[s^2(n(t))] = 2\mu(2N + 1/2v) + 4\mu^2(20N^2 + 7N/v + 5/4v^2). \tag{12}$$

Analogous results have been derived by M. W. Feldman *et al.* (unpublished) using the recursion approach.

Change in Population Size

We can also use this framework to compute the moments of $s(n(t))$ in a population whose size changes over time. This kind of problem has previously been examined for infinite sites models by several workers, including Tajima (1989). In addition, Chakraborty and Nei (1977) studied the effect of bottlenecks on average heterozygosity and genetic distance between populations under the stepwise mutation model.

Suppose that the haploid population size has been N_1 for the last τ generations, and before that it was N_2. Then we can find the expected values of t and t^2 by observing that the probability density function of t is exponentially distributed with parameter N_1^{-1} for $0 \leqslant t < \tau$ and exponentially distributed with parameter N_2^{-1} for $t \geqslant \tau$. Then the expected values

of t and t^2 can be calculated using the fact that the probability of $t > \tau$ is $e^{-\tau/N_1}$. Since the exponential distribution is memoryless, it follows that

$$E[t] = \int_0^\tau \frac{t}{N_1} e^{-t/N_1}\, dt + e^{-\tau/N_1} \int_\tau^\infty \frac{t}{N_2} e^{(-t-\tau)/N_2}\, dt,$$

$$E[t^2] = \int_0^\tau \frac{t^2}{N_1} e^{-t/N_1}\, dt + e^{-\tau/N_1} \int_\tau^\infty \frac{t^2}{N_2} e^{(-t-\tau)/N_2}\, dt.$$

Computing the integrals and substituting the values into (1) and (5), the expectation of s^2 and its variance are calculated as

$$E[s^2(n(t))] = 2\mu(N_1 + (N_2 - N_1)\, e^{-\tau/N_1}), \tag{13}$$

$$\begin{aligned}\mathrm{Var}[s^2(n(t))] = {} & 2\mu[N_1 + (N_2 - N_1)\, e^{-\tau/N_1}] \\ & + 4\mu^2[5N_1^2 + (6N_2(N_2 + \tau) - 2N_1(2N_1 + N_2 + 3\tau))\, e^{-\tau/N_1} \\ & - (N_2 - N_1)^2\, e^{-2\tau/N_1}].\end{aligned} \tag{14}$$

It is interesting to observe that in the limit, as N_1 goes to infinity, this model converges to the case of the branching populations with no migration.

Population Heterozygosity

An additional quantity of interest is the probability π that two chromosomes sampled at random from a population do not share the same microsatellite repeat number at some locus. This quantity is analogous to the heterozygosity of a gene, or nucleotide diversity at the DNA sequence level (e.g., see Nei, 1987). In terms of the theoretical framework described above, we clearly have $\pi = \Pr[s(n(t)) \neq 0]$. Observe that we can only have $s = 0$ when n, the number of mutations, is even, so that there are an equal number of steps (i, say) up and down. The probability of having both i steps up and down, given that $n = 2i$ mutations have occurred is simply

$$\binom{2i}{i}\left(\frac{1}{2}\right)^{2i}.$$

The probability of having exactly $2i$ mutations is the Poisson probability

$$\frac{(2t\mu)^{2i}\, e^{-2t\mu}}{(2i)!}.$$

Thus, for a given value of t, the probability that $s(n(t)) = 0$ is

$$\binom{2i}{i}\left(\frac{1}{2}\right)^{2i} \frac{(2t\mu)^{2i}\, e^{-2t\mu}}{(2i)!} = \frac{(t\mu)^{2i}}{(i!)^2}\, e^{-2t\mu}.$$

Then the total probability that $s(n(t))=0$ is found by summing over all i and, integrating over all t, weighted by the probability density function for t. So the heterozygosity π is given by

$$\pi = 1 - \sum_{i=0}^{\infty} \int_0^{\infty} \frac{(t\mu)^{2i}}{(i!)^2} e^{-2t\mu} \frac{e^{-t/N}}{N} dt. \tag{15}$$

The sum on the right side of (15) is the Laplace transform of the modified Bessel function (Abramowitz and Stegun, 1965, p. 1027 (1970 edition)). Thus the heterozygosity reduces to $1-(1+4N\mu)^{-1/2}$, which is the result originally obtained by Ohta and Kimura (1973).

VARIANCE OF s^2 FOR A SAMPLE TAKEN FROM A POPULATION

In order to estimate $E[s^2]=2\mu N$ for some microsatellite locus in a particular population, a natural approach would be to collect a sample of m chromosomes from the population, find the repeat number for each, and take the mean value of s^2 over the $\binom{m}{2}$ pairwise comparisons. Here we show how to find the variance of this estimate of $2\mu N$ for an arbitrary sample size. An interesting result is that the variance of the estimate does not approach zero as m becomes arbitrarily large. This property has also been noted by Tajima (1983) for a similar problem using the infinite sites model. The first steps of the argument used here will follow arguments used by Tajima in that study. In these calculations we assume the simplest population model: a single random mating population of haploid size N.

Let k_{ij} represent the actual value of the squared distance between the ith and jth chromosomes in the sample ($k_{i,j}=s^2_{i,j}(n_{i,j}(t_{i,j}))$). Then the estimate of the mean square distance based on the sample is defined by

$$\hat{k} = \sum_{i<j}\sum k_{ij} \binom{m}{2}^{-1}. \tag{16}$$

Following the framework of Tajima (1983) we have

$$E[\hat{k}] = \sum_i \sum_j E[k_{ij}] \binom{m}{2}^{-1} = 2\mu N \tag{17}$$

for $i<j$, so that $\hat{k}$ is an unbiased estimator. The variance of $\hat{k}$ is given by

$$\mathrm{Var}[\hat{k}] = E[\hat{k}^2] - E[\hat{k}]^2. \tag{18}$$

Observe that we have already discussed in detail the properties of $\hat{k}$ and $\hat{k}^2$ for the special case of $m=2$; the expectations of $\hat{k}$, $\hat{k}^2$, and $\mathrm{Var}[\hat{k}]$ are given by (2), (4), and (6), respectively, in the $m=2$ case in a single population.

Note that we can write $E[\hat{k}^2]$ as

$$\begin{aligned} E[\hat{k}^2] &= E\left[\left\{\sum_i \sum_j k_{ij}\right\}^2\right]\binom{m}{2}^{-2} \\ &= \left\{\sum_i \sum_j U_2 + \sum_i \sum_j \sum_r U_3 + \sum_i \sum_j \sum_r \sum_s U_4\right\}\binom{m}{2}^{-2} \\ &= \left\{U_2 + 2(m-2)\, U_3 + \binom{m-2}{2} U_4\right\}\binom{m}{2}^{-1}, \end{aligned} \tag{19}$$

where $i \neq j \neq r \neq s$, and $i < j$ and $r < s$ and

$$U_2 := E[k_{ij}^2] \tag{20a}$$

$$U_3 := E[k_{ij} \times k_{ir}] \tag{20b}$$

$$U_4 := E[k_{ij} \times k_{rs}]. \tag{20c}$$

In order to estimate $\mathrm{Var}[\hat{k}]$ we need U_2, U_3, and U_4. These correspond to studying genealogies with 2, 3, and 4 individuals, respectively. We have already studied the properties of trees with two individuals in detail; in fact, the value of U_2 is given by (4):

$$U_2 = 2\mu N + 24\mu^2 N^2. \tag{21}$$

The calculations for U_3 and U_4 (see Appendix 2) turn out to be more involved, but they give final values of

$$U_3 = \mu N + 12\mu^2 N^2, \tag{22}$$

$$U_4 = \frac{2\mu N + 28\mu^2 N^2}{3}. \tag{23}$$

Substituting these values of U_2, U_3, and U_4 into (19) it follows that

$$E[\hat{k}^2] = \frac{2\mu N[m(m+1)] + 4\mu^2 N^2[2+3m+7m^2]}{3(m^2-m)}. \tag{24}$$

So from (18), we have the variance of $\hat{k}$ as

$$\mathrm{Var}[\hat{k}] = \frac{2\mu N[m(m+1)] + 8\mu^2 N^2[1+3m+2m^2]}{3(m^2-m)}. \tag{25}$$

Notice that as the sample size m becomes large, the variance approaches

$$\frac{2\mu N}{3}+\frac{16\mu^2N^2}{3}, \tag{26}$$

which is just $U_4-E[k]^2$ and agrees with the value recently obtained by Zhivotovsky and Feldman (1995). Some results concerning higher order moments of variances in stepwise mutation models have been obtained previously by Roe (1992), who also used methods from coalescent theory.

COVARIANCE OF SQUARED DISTANCES AT LINKED LOCI

In this section we suggest a measure of the linkage disequilibrium between microsatellite loci which takes advantage of the stepwise mutation process. Instead of merely considering alleles as being the same or different, as is done when studying disequilibrium between point mutations, we can use the property that alleles which are similar in repeat number tend to be more closely related than alleles which are far apart. Suppose that two loci are completely linked and, therefore, share a common coalescence time. If this coalescence time is very long, then the squared distance is expected to be large at both loci; if it is short, then the squared distance is expected to be small at both loci. It therefore seems intuitively reasonable that the covariance or correlation in squared distance between two loci can be used as a metric of linkage disequilibrium. Here we show analytically for the covariance that this is in fact the case. This analysis is related to work by Weir (1992) and Chakraborty, Srinivasan, and Andrade (1993), who were interested in VNTR loci in connection with DNA fingerprinting. Those authors discuss tests of correlation in size between different loci, but do not consider the underlying mutation process or population genetics giving rise to the correlations.

Let $s_1^2(n_1(t_1))$ and $s_2^2(n_2(t_2))$ be the squared differences in repeat number between two chromosomes at microsatellite loci 1 and 2, respectively. Then the covariance of s_1^2 and s_2^2 is given by

$$\begin{aligned}\mathrm{Cov}[s_1^2(n_1(t_1)), s_2^2(n_2(t_2))] &= E[s_1^2(n_1(t_1))\times s_2^2(n_2(t_2))]\\ &\quad -E[s_1^2(n_1(t_1))]\,E[s_2^2(n_2(t_2))].\end{aligned}$$

Recall that $E[s^2]$ is given by (1). In Appendix 1, we show how to calculate $E[s_1^2\times s_2^2]$; its value is given by (A4). Then we have

$$\begin{aligned}\mathrm{Cov}[s_1^2(n_1(t_1)), s_2^2(n_2(t_2))] &= 4\mu^2E[t_1t_2]-4\mu^2E[t_1]\,E[t_2]\\ &= 4\mu^2\mathrm{Cov}[t_1, t_2].\end{aligned} \tag{27}$$

When there is complete linkage between loci 1 and 2, we have $t_1 \equiv t_2$. Then the covariance can readily be predicted for each of the population models described in the first part of the paper, merely by substitution of $E[t]$ and $E[t^2]$. For instance, in the simplest population model—one population of size N—we have $E[t] = N$ and $E[t^2] = 2N^2$. Then the covariance is

$$\mathrm{Cov}[s_1^2(n_1(t_1)), s_2^2(n_2(t_2))] = 8\mu^2N^2 - 4\mu^2N^2 = 4\mu^2N^2. \tag{28}$$

In the case of partial linkage, Hudson (1990) gives the following formula (based on theory developed by Griffiths (1981) and Hudson (1983)) for the correlation of t_1 and t_2 in a population:

$$\mathrm{Cor}[t_1, t_2] = \frac{R + 18}{R^2 + 13R + 18},$$

under usual diffusion approximations, where $R = 2Nc$ and c is the recombination fraction between the two loci. If we assume that t_1 and t_2 have identical distributions, then the variances of t_1 and t_2 are equal, so we can write $\mathrm{Cov}[t_1, t_2] = \mathrm{Cor}[t_1, t_2]\ \mathrm{Var}[t]$. The variance of t for the exponential distribution is N^2, so the result is that

$$\mathrm{Cov}[s_1^2, s_2^2] = 4\mu^2N^2 \frac{R + 18}{R^2 + 13R + 18}. \tag{29}$$

This shows that there is nonzero linkage disequilibrium between microsatellite loci at mutation–drift equilibrium. The strength of the association falls off roughly as the inverse of $2Nc$.

As a practical application, one may wish to know whether two microsatellite loci are in linkage disequilibrium in a population. In the light of these theoretical results, it would be reasonable to use the estimated covariance of squared distances at the two loci as a test statistic. Thus, with the haplotypes of a sample of m chromosomes drawn from the population, the estimated covariance would be calculated by finding the mean values of s_1^2, s_2^2, and $(s_1 s_2)^2$, over all $\binom{m}{2}$ pairwise comparisons. Since we do not know the distribution of the covariance at this time, a nonparametric test of significance is appropriate. The null hypothesis is that the level of association (i.e., covariance) between alleles at the two loci is not greater than would be expected by chance alone. We can estimate the properties of a distribution in which the two loci are independent of one another using some sort of permutation test. One such test would be to generate new data sets by randomly pairing alleles from the original data set (one allele from each locus, sampling without replacement), until a new set

of m haplotypes was generated. The covariance would be calculated for each new data set, and one would conclude that there is significant disequilibrium if the observed covariance is greater than the covariance in 95%, say, of the random data sets. This is similar to a test described by Chakraborty, Srinivasan, and Andrade (1993), who were interested in correlations of absolute allele size between VNTR loci, rather than of squared distances, as here.

One difficulty with this approach, pointed out by Zhivotovsky and Feldman (1995), is that the variance of estimators of distances using s^2 is likely to be very large, reducing the utility of confidence intervals for distances (and times) based on s^2.

DISCUSSION

The theoretical framework which has been established in this study is intended to accomplish a couple of different goals related to the analysis of microsatellite data. The first is to make it easier to compute the moments of microsatellite distances under a wide range of population models. The types of population models discussed here are partly intended as illustrations of the utility of the method and can easily be extended to other cases of special interest. Second the coalescent viewpoint combined with our $s(n(t))$ model immediately suggests the possibility of highly efficient computer simulations of microsatellite evolution. These can be used both for checking theoretical results and for studying cases which are analytically intractable. In the latter case, the results developed here may prove to be useful as special cases in which the simulation results can be checked. Furthermore, this work provides the foundation for the development of a number of statistical tests. There are, for instance, applications in which it is useful to place confidence intervals on $\hat{k}$ (e.g., Goldstein *et al.* 1996), or confidence intervals on the level of genetic exchange between populations (e.g., see discussion of R_{ST} and F_{ST} in Slatkin, 1995).

It must be remembered that the results presented here are predicated on a symmetric model of mutation. There is some evidence that this assumption is too restrictive for application to some bodies of data. Rubinsztein *et al.* (1995) suggest a tendency for mutation to higher copy number, while other workers have proposed a model in which mutation tends to be biased toward some focal value (Garza, Slatkin, and Freimer, 1995). Of course, such departures from symmetry would alter our results.

One possible use of microsatellite data is as an estimator of effective population size. The results obtained here concerning the variance of $\hat{k}$ for a sample taken from a population are instructive. It is apparent that the variance is quite considerable: if $N\mu \gg 1$, the coefficient of variation of

the estimate of $N\mu$ based on a large sample at a single locus is roughly $(4/3)^{1/2} \approx 115\%$. It is clearly necessary to use many loci to achieve a satisfactory confidence interval on $N\mu$. The good news is that the sample sizes at each locus need not be large. Again, for large $N\mu$, the variance on $\hat{k}$ when ten chromosomes are typed is just 1.28 times greater than the variance achieved with an infinite sample.

APPENDIX 1: THE EXPECTATION OF PRODUCTS OF $s(n(t))$

Here we show how to calculate the expected value of a product of the form

$$s_1^a(n_1(t_1)) \times s_2^b(n_2(t_2)) \times \cdots \times s_j^k(n_j(t_j)), \tag{A1}$$

where the subscripts of s refer to realizations of the evolution of a microsatellite on different branches of a genealogical tree, or on different trees (e.g., for different loci or for comparisons of different pairs of individuals). The superscripts are exponents. This result is used for two different applications in this paper. One of these is to calculate the covariance of s^2 at two linked loci. For that calculation we need to find $E[s_1^2(n_1(t_1)) \times s_2^2(n_2(t_2))]$. In such a situation, the natural assumption is that the coalescent times t_1 and t_2 are not necessarily independent of one another, but that mutations at one locus, or at one point in time do not affect the probability of mutation elsewhere. Formally, by independence of mutations we mean that once the values of the t_i are specified, the values of the s_i do not covary:

$$\mathrm{Cov}[(s_1^a(n_1(t_1)), s_2^b(n_2(t_2)), ..., s_j^k(n_j(t_j))) \,|\, (t_1, t_2, ..., t_j)] = 0. \tag{A2}$$

In the following calculation, uppercase letters S and T refer to particular realizations of the random processes s and t. Furthermore, for simplicity of notation, let us take $s(t)$ as a shorthand for $s(n(t))$. (Then we consider the probability of some specific distance $S(T)$ as being the probability of $S(n(T))$ summed over all n. This is justified by the assumption of independence of mutations in (A2).) So from the definition of a joint expectation we have

$$\begin{aligned} &E[s_1^a(t_1) \times s_2^b(t_2) \times \cdots \times s_j^k(t_j)] \\ &\quad = \int_{T_1} \int_{T_2} \cdots \int_{T_j} \sum_{S_1} \sum_{S_2} \cdots \sum_{S_n} S_1^a(T_1) \times S_2^b(T_2) \times \cdots \times S_j^k(T_j) \\ &\qquad \times \Pr[(S_1 \,|\, T_1), (S_2 \,|\, T_2), ..., (S_j \,|\, T_j), T_1, T_2, ..., T_j] \times dT_1\, dT_2 \cdots dT_j. \end{aligned}$$

By (A2) we can simplify this to

$$\int_{T_1}\int_{T_2}\cdots\int_{T_j}\sum_{S_1}\sum_{S_2}\cdots\sum_{S_j} S_1^a(T_1)\times S_2^b(T_2)\times\cdots\times S_j^k(T_j)$$
$$\times \Pr[S_1\,|\,T_1]\times\Pr[S_2\,|\,T_2]\times\cdots\times\Pr[S_j\,|\,T_j]$$
$$\times \Pr[T_1, T_2, ..., T_j]\times dT_1\,dT_2\cdots dT_j,$$

from which we can write

$$\int_{T_1}\int_{T_2}\cdots\int_{T_j}\left\{\sum_{S_1} S_1^a\Pr[S_1\,|\,T_1]\right\}\times\left\{\sum_{S_2} S_2^b\Pr[S_2\,|\,T_2]\right\}$$
$$\times\cdots\times\left\{\sum_{S_j} S_1^k\Pr[S_j\,|\,T_j]\right\}\times\Pr[T_1, T_2, ..., T_j]\times dT_1\,dT_2\cdots dT_j.$$

So this reduces to

$$E[s_1^a(t_1)\times s_2^b(t_2)\times\cdots\times s_j^k(t_j)]$$
$$=\int_{T_1}\int_{T_2}\cdots\int_{T_j} E\{s_1^a\,|\,T_1\}\times E\{s_2^b\,|\,T_2\}$$
$$\times\cdots\times E\{s_j^k\,|\,T_k\}\times\Pr[T_1, T_2, ..., T_j]\times dT_1\,dT_2\cdots dT_j. \qquad \text{(A3)}$$

One important consequence of (A3) is that if any of the s_i in the original product (A1) is raised to an odd power, the expectation of the entire product is zero. Furthermore, it is particularly easy to calculate the expectation of $s_1^2(n_1(t_1))\times s_2^2(n_2(t_2))$, whose value is needed for the covariance calculation described above:

$$E[s_1^2(n_1(t_1))\times s_2^2(n_2(t_2))]=\int_{T_1}\int_{T_2} E\{s_1^2\,|\,T_1\}\times E\{s_2^2\,|\,T_2\}$$
$$\times\Pr[T_1, T_2]\,dT_1\,dT_2$$
$$=\int_{T_1}\int_{T_2}(2\mu T_1)\times(2\mu T_2)\times\Pr[T_1, T_2]\,dT_1\,dT_2$$
$$=4\mu^2\int_{T_1}\int_{T_2} T_1 T_2\Pr[T_1, T_2]\,dT_1\,dT_2$$
$$=4\mu^2 E[t_1 t_2]. \qquad \text{(A4)}$$

APPENDIX 2: CALCULATIONS OF U_3 AND U_4

In order to compute the variance of s^2 for a sample, we need to calculate the quantities U_3 and U_4, defined in (20b) and (20c), respectively. We can calculate the value of U_3 by calculating the expected values of $k_{ij} \times k_{ir}$ separately for each possible branching pattern in a tree for three individuals. The possible patterns are shown in Fig. 1. Only one basic topology exists for a tree with three endpoints, but the way in which the branches are labelled produces different expected values of $k_{ij} \times k_{ir}$. One third of the six possible labellings have individual i as the outgroup (Fig. 1A), and two thirds have j or r as the outgroup (Fig. 1B). The values of $k_{ij} \times k_{ir}$ will be designated U_{3a} and U_{3b} respectively for these two different cases. Then the value of U_3 is a weighted average of these:

$$U_3 = U_{3a}/3 + 2U_{3b}/3. \tag{A5}$$

We designate by t_2 the elapsed time between the present (at which time there are three lineages) and the internal node (point at which two of the three lineages coalesce). The time between the internal node and the root of the tree (at which time the two remaining lineages coalesce) is t_1 (see Fig. 1). The coalescence time t_1 is simply the coalescence time for two lineages; as before this is exponentially distributed with $E[t_1] = N$ and $E[t^2] = 2N^2$. The coalescence time t_2 is the time required for three lineages to merge into two. There are $\binom{3}{2}$ ways in which this can occur, and so t_2 is exponentially distributed with $E[t] = N/3$ and $E[t^2] = 2N^2/9$. (See Tajima, 1983, for some of the assumptions which go into a model of this kind.)

We will use lower case letters to designate values of $s(n(t))$ on the branches of a tree (see Fig. 1). For U_3, x represents the value of s on the branch which leads exclusively to individual i (i.e., for U_{3a}, x is the difference in microsatellite repeat number between individual i and the

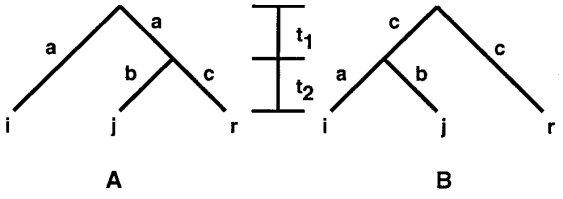

FIG. 1. Two possible branching patterns which must be considered in calculating U_3. Sampled individuals are represented by i, j, and r. The differences in microsatellite repeat number between the putative individual at the internal node and individuals i, j, and r are x, y, and z, respectively. The elapsed time between sampling and the internal node is t_2; the time between the internal node and the root of the tree is t_1.

putative ancestor at the node connecting j and r), y is the value on the branch leading to j, and z is the value on the branch leading to r. Note that when i is the outgroup, x is the distance on the branch which goes up through the root and down the other side to the internal node. For U_{3a}, the elapsed time on the branch for x is $2t_1 + t_2$, and the times on the branches for y and z are both t_2.

Then we can rewrite U_{3a} as

$$U_{3a} = E[k_{ij} \times k_{ir}] = E[(x+y)^2 (x+z)^2]. \tag{A6}$$

But when we expand the product $(x+y)^2 (x+z)^2$, we can make use of (A3), which allows us to neglect any terms with odd powers of x, y, or z. So we have

$$U_{3a} = E[x^4 + x^2y^2 + x^2z^2 + y^2z^2]. \tag{A7}$$

We can calculate the value of (A7) using (3), which provides the expected value of s^4 and (A4) which provides the expected value of $s_1^2 s_2^2$. Recall that in the derivations of (3) and (A4) the mutation rate was taken to be $2\mu t$, reflecting that there are two branches of length t. Now it will be more convenient to do the accounting for each branch separately, so in what follows, we will modify (3) and (A4) by removing a factor of 2 from the mutation rate, and taking t_{tot} as the *total* branch length:

$$E[s^4(n(t_{\text{tot}}))] = \mu E[t_{\text{tot}}] + 3\mu^2 E[t_{\text{tot}}^2] \tag{3'}$$

$$E[s_1^2(n_1(t_{1\text{tot}})) \times s_2^2(n_2(t_{2\text{tot}}))] = \mu^2 E[t_{1\text{tot}} t_{2\text{tot}}]. \tag{A4'}$$

Using (3′) and (A4′), and inserting the elapsed times of branches x, y, and z into (A7) we get

$$\begin{aligned} E_{3a} &= \mu E[2t_1 + t_2] + 3\mu^2 E[(2t_1 + t_2)^2] \\ &\quad + \mu^2 [E[(2t_1 + t_2)\, t_2] + E[(2t_1 + t_2)\, t_2] + E[t_2^2]] \\ &= \mu E[2t_1 + t_2] + \mu^2 E[12t_1^2 + 16t_1 t_2 + 6t_2^2]. \end{aligned} \tag{A8}$$

Then using the expected values of t_1, t_2, t_1^2, and t_2^2, and the independence of t_1 and t_2, we get

$$U_{3a} = \tfrac{7}{3}\mu N + \tfrac{92}{3}\mu^2 N^2. \tag{A9a}$$

Similarly, the value of U_{3b} can be shown to be

$$U_{3b} = \tfrac{1}{3}\mu N + \tfrac{8}{3}\mu^2 N^2. \tag{A9b}$$

Then combining these in the right proportions (see (A5)), the value of U_3 is

$$U_3 = \mu N + 12\mu^2 N^2. \tag{A9c}$$

Calculation of U_4

The calculation of U_4 is performed in much the same way as for U_3. There are now two different tree topologies. The first topology (Fig. 2A) occurs with probability $\frac{2}{3}$, and the second topology occurs with probability $\frac{1}{3}$ (Tajima, 1983). There are three distinct ways of picking pairs for the first topology (a, b, c) and two ways for the second topology (d and e). Counting the possible arrangements of i, j, r, and s it turns out that arrangements corresponding to d are twice as frequent as e). So a, b, c, and d each occur with frequency $\frac{2}{3} \times \frac{1}{3}$, and e with frequency $1/3^2$. Then

$$U_4 = \tfrac{2}{9}(U_{4a} + U_{4b} + U_{4c} + U_{4d}) + \tfrac{1}{9}U_{4e}. \tag{A10}$$

As before, the times between the nodes are labelled: t_1 is the time in which there are just two distinct lineages, t_2 is the time when there are three lineages, and t_3 is the time when there are four lineages (see Fig. 2). The moments of t_1 and t_2 are as given in the discussion for U_3; and t_3 is exponentially distributed with a mean of $N/\binom{4}{2} = N/6$, and $E[t^2] = N^2/18$.

The branches are labelled as shown in Fig. 2; the times for each branch in Fig. 2A are v: $2t_1 + t_2 + t_3$; w: t_2; x: $t_2 + t_3$; y and z: t_3. The times for the branches in Fig. 2B are v: $2t_1 + t_2$; w and x: $t_2 + t_3$; y and z: t_3.

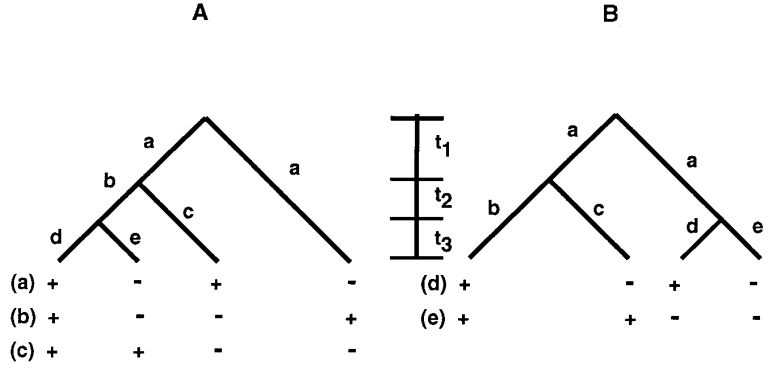

FIG. 2. U_4 equals $E[k_{ij}k_{rs}]$, in which individuals i, j, r, and s are all distinct. Including the two possible topologies (A and B), there are five distinct arrangements (designated (a) through (e)) of i, j, r, and s. The pairwise comparisons (e.g., k_{ij}) are between "+"'s or between "−"'s. Differences in microsatellite repeat number between nodes, or between nodes and tips of the tree are designated by v, w, x, y, and z as indicated. The elapsed times between coalescent events are given by t_1, t_2, and t_3 for the periods in which there are two, three, and four extant lineages, respectively.

Now the expected values of $k_{ij} \times k_{rs}$ can be calculated for each of these five cases in the same way as for U_3. For instance, for U_{4a} we have

$$\begin{aligned} U_{4a} &= (w+x+y)^2 (v+w+z)^2 \\ &= w^4 + v^2w^2 + v^2x^2 + v^2y^2 + w^2x^2 + w^2y^2 + w^2z^2 + x^2z^2 + y^2z^2 \\ &= \mu E[t_2] + \mu^2 E[6t_2^2 + 4t_3^2 + 4t_1t_2 + 4t_1t_3 + 8t_2t_3] \\ &= \tfrac{1}{3}\mu N + 4\mu^2 N^2. \end{aligned} \tag{A11a}$$

The other values are

$$U_{4b} = \tfrac{1}{3}\mu N + 4\mu^2 N^2 \tag{A11b}$$

$$U_{4c} = \tfrac{10}{9}\mu^2 N^2 \tag{A11c}$$

$$U_{4d} = \tfrac{7}{3}\mu N + \tfrac{98}{3}\mu^2 N^2 \tag{A11d}$$

$$U_{4e} = \tfrac{4}{9}\mu^2 N^2. \tag{A11e}$$

From (A10) we have the value of U_4 as

$$U_4 = \frac{2\mu N + 28\mu^2 N^2}{3}. \tag{A11f}$$

REFERENCES

Abramowitz, M., and Segun, I. A., 1965. "Handbook of Mathematical Functions," Dover, New York.

Chakraborty, R., and Nei, M., 1977. Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model, *Evolution* **31**, 347–356.

Chakraborty, R., Srinivasan, M. R., and M. de Andrade, 1993. Intraclass and interclass correlations of allele sizes within and between loci in DNA typing data, *Genetics* **133**, 411–419.

Di Rienzo, A., Peterson, A. C., Garza, J. C., Valdes, A. M., Slatkin, M., and Freimer, N. B., 1994. Mutational processes of simple-sequence repeat loci in human populations, *PNAS* **91**, 3166–3170.

Goldstein, D. B., Linares, A. R., Cavalli-Sforza, L. L., and Feldman, M. W., 1995a. An evaluation of genetic distances for use microsatellite loci, *Genetics* **139**, 463–471.

Goldstein, D. B., Linares, A. R., Cavalli-Sforza, L. L., and Feldman, M. W., 1995b. Genetic absolute dating based on microsatellites and the origin of modern humans, *PNAS* **92**, 6723–6727.

Goldstein, D. B., Zhivotovsky, L. A., Nayar, K., Linares, A. R., Cavalli-Sforza, L. L., and Feldman, M. W., 1996c. Statistical properties of the variation at linked microsatellite loci: Implications for the history of human *Y* chromosomes, *Mol. Biol. Evol.*

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

Griffiths, R. C., 1981. Neutral two-locus multiple allele models with recombination, *Theor. Popul. Biol.* **19**, 169–186.

Hey, J., 1991. A multi-dimensional coalescent process applied to multi-allelic selection models and migration models, *Theor. Popul. Biol.* **39**, 30–48.

Hudson, R. R., 1983. Properties of a neutral allele model with intragenic recombination, *Theor. Popul. Biol.* **23**, 183–201.

Hudson, R. R., 1990. Gene genealogies and the coalescent process, *in* "Oxford Surveys in Evolutionary Biology" (D. J. Futuyma and J. Anotonovics, Eds.), Vol. 7, pp. 1–44, Oxford Univ. Press, Oxford.

Hudson, R. R., 1993. The how and why of generating gene genealogies, *in* "Mechanisms of Molecular Evolution" (N. Takahata and A. G. Clark, Eds.), pp. 23–36, Sunderland, MA.

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

Nei, M., 1987. "Molecular Evolutionary Genetics," Columbia Univ. Press, New York.

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

Révész, P., 1990. "Random Walk in Random and Non-random Environments," World Scientific, Singapore.

Roe, A., 1992. "Correlations and Interactions in Random Walks and Population Genetics," Ph.D. thesis, University of London, London, UK.

Rubinsztein, D. C., Amos, W., Leggo, J., Goodburn, S., Jain, S., Li, S.-H., Margolis, R. L., Ross, C. A., and Ferguson-Smith, M. A., 1995. Microsatellite evolution—Evidence for directionality and variation in rate between species, *Nature Gen.* **10**, 337–343.

Schlötterer, C., and Tautz, D., 1992. Slippage synthesis of simple sequence DNA, *Nucleic. Acids Res.* **20**, 211–216.

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

Slatkin, M., 1991. Inbreeding coefficients and coalescence times, *Genet. Res. Cambridge* **58**, 167–175.

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

Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations, *Genetics* **105**, 437–460.

Tajima, F., 1989. The effect of change in population size on DNA polymorphism, *Genetics* **123**, 597–601.

Tajima, F., 1993. Measurement of DNA polymorphism, *in* "Mechanisms of Molecular Evolution" (N. Takahata and A. G. Clark, Eds.), pp. 37–59, Sinauer, Sunderland, MA.

Tautz, D., 1993. Notes on the definition and nomenclature of tandemly repetitive DNA sequences, *in* "DNA Fingerprinting: State of the Science" (S. D. J. Pena, R. Chakraborty, J. T. Epplen, and A. J. Jeffreys, Eds.), pp. 21–28, Birkhäuser, Basel.

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

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

Weir, B. S., 1992. Independence of VNTR alleles defined as fixed bins, *Genetics* **130**, 873–887.

Zhivotovsky, L. A., and Feldman, M. W., 1995. Microsatellite variability and genetic distance, *Proc. Natl. Acad. Sci. USA* **92**, 1549–1552.