Dynamics of Repeat Polymorphisms Under a Forward-Backward Mutation Model: Within- and Between-Population Variability at Microsatellite Loci

Marek Kimmel,* Ranajit Chakraborty,† David N. Stivers† and Ranjan Deka‡

**Department of Statistics, Rice University, Houston, Texas 77251, †Human Genetics Center, University of Texas Health Science Center, Houston, Texas 77225 and ‡Department of Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania 15261*

Manuscript received October 25, 1995
Accepted for publication January 27, 1996

ABSTRACT

Suggested molecular mechanisms for the generation of new tandem repeats of simple sequences indicate that the microsatellite loci evolve via some form of forward-backward mutation. We provide a mathematical basis for suggesting a measure of genetic distance between populations based on microsatellite variation. Our results indicate that such a genetic distance measure can remain proportional to the divergence time of populations even when the forward-backward mutations produce variable and/or directionally biased alleles size changes. If the population size and the rate of mutation remain constant, then the measure will be proportional to the time of divergence of populations. This genetic distance is expressed in terms of a ratio of components of variance of allele sizes, based on expressions developed for studying population dynamics of quantitative traits. Application of this measure to data on 18 microsatellite loci in nine human populations leads to evolutionary trees consistent with the known ethnohistory of the populations.

THE study of human genome diversity for inferring the history of human genetic differentiation has been a focus of attention of biological anthropological investigations since the discovery of the first polymorphic marker in the human genome. Traditional serological and immunological markers used for this purpose generally do not provide a high resolution for distinguishing populations with close historical connections. This is so because genetic distances between populations are generally small, except for the major racial groups. This causes statistical error in phylogenetic reconstruction of the history of genetic differentiations of world populations. The advent of hypervariable DNA polymorphisms has the potential for increasing the accuracy of such studies, as has been empirically demonstrated (BOWCOCK *et al.* 1994; DEKA *et al.* 1995a,b). Hypervariable tandem repeat markers have the potential for being particularly efficient for this purpose, because they offer a greater number of segregating alleles. As a consequence, the extent of genetic diversity within and between populations (in absolute scale) is larger for these loci than for the traditional loci.

In order that a measure of genetic distance be useful for such studies, it is necessary for it to be a known monotonic increasing function of time of divergence (*i.e.*, the absence of gene flow) between populations. This criterion raises concerns regarding the utility of hypervariable tandem repeat markers in evolutionary studies, since alleles at such loci evolve by molecular processes that involve both contraction and expansion of repeat sizes (WEBER and WONG 1993; JEFFREYS *et al.* 1994). As a consequence, alleles of similar sizes are not necessarily evolutionarily related (KIDD *et al.* 1991). Furthermore, the mathematical relationship of the expectations of traditional measures of genetic distances with the time of divergence have specific underlying assumptions regarding mutation processes that may not hold for tandem repeat loci (SHRIVER *et al.* 1993; VALDES *et al.* 1993; DI RIENZO *et al.* 1994).

While the mechanism of mutational changes at tandem repeat loci is not precisely known at a molecular level, empirical observations indicate that the allele size changes can be approximated by a forward-backward random walk model. The population dynamics of specific forms of such mutation models has been studied in the context of within- and between-population genetic variation at protein-enzyme loci (WEHRHAHN 1975; LI 1976; CHAKRABORTY and NEI 1982). In these models, mutations were frequently assumed to introduce symmetric changes of one and/or two steps in either direction. CHAKRABORTY and NEI (1982) considered a general multi-step random walk model using a bidirectional binomial as a specific example. Mathematically, the same model applies to tandem repeat loci, with allelic states defined by size (number of repeat units) instead of the charge of the protein molecule.

In an application of the model of CHAKRABORTY and NEI (1982), the index of between- *vs.* within-population variance of allelic states was introduced and proven to be proportional to the time of divergence of populations in the absence of migration, selection and population size fluctuations.

Corresponding author: Ranajit Chakraborty, Human Genetics Center, University of Texas Health Science Center, P.O. Box 20334, Houston, TX 77225.

More recently, SLATKIN (1995) considered a similar index for studying gene differentiation at tandem repeat loci in a substructured population. SLATKIN further showed that the variance components in his index relate to the within- and between-population diversity measures proposed by GOLDSTEIN *et al.* (1995a,b). Also, SHRIVER *et al.* (1995) suggested a revised measure of genetic distance that they argue is appropriate for microsatellite tandem repeat markers and that is proportional to the time of divergence between populations under some specific conditions of forward-backward mutation of alleles.

In this paper our goal is to demonstrate that SLATKIN's (1995) index, which is identical with an index introduced by CHAKRABORTY and NEI (1982), has properties more general than those indicated in the literature; namely, that the time linearity of the index holds for stepwise mutations with arbitrary distributions of changes of allele size. Therefore, the index is applicable for studying genetic differentiation in the presence of allele-size expansion or contraction bias in mutations, such as those postulated by RUBINSZTEIN *et al.* (1995). We show this by using both a direct population dynamics approach and the theory of coalescence for a generalized stepwise mutation model.

We also illustrate that the CHAKRABORTY-NEI-SLATKIN index, applied to a new set of microsatellite loci typed in several of human populations of African, Oriental and Indoeuropean origin, leads to interpopulation distances consistent with the accepted ancestry of human populations. The importance of the model assumptions, including constancy of population size and mutation rate, are discussed in light of the above theoretical and empirical observations.

THEORY

We show that under any general forward-backward mutation model, the within-population allelic variation in a finite population reaches a state of equilibrium when variation is measured in terms of the distribution of allele size differences in genotypes of diploid individuals. In addition, when the population sizes of two diverging populations remain constant over time, we show that the ratio of between- *vs.* within-population variance of allele size differences is linearly related to the time of divergence and is independent of mutation rate.

Within-population variability: Consider a population of diploid individuals and a locus with a denumerable set of alleles indexed by integer numbers. The within-population component of genetic variance

$$E\left[\sum_{i=1}^{2N}(X_i-\bar{X})^2/(2N-1)\right],$$

where $E(\cdot)$ denotes the expectation of a random variable, and X_i is the size of the allele in the ith chromosome present, is equal to $V_t/2$, where

$$V_t = E[(X_i - X_j)^2], \tag{1}$$

and X_i, X_j are the sizes of two alleles randomly selected from the population. X_i and X_j are time-dependent random variables, *i.e.*, $X_i = X_i(t)$ and $X_j = X_j(t)$, but for notational simplicity the argument t is suppressed, since the time dependence is always clear from the context. In V_t, subscript t denotes chronological time (in units of generations) counted from a convenient reference point. We consider the time evolution of V_t in a stepwise mutation model with sampling from the finite allele pool. We assume the following:

- In each generation, the genotypes of all individuals are sampled with replacement from the $2N$ chromosomes present in the previous generation (FISHER-WRIGHT model, EWENS 1979).
- Each chromosome independently is subject, with probability ν per generation, to a mutation that replaces an allele of size X with an allele of size $X + U$, where U is an integer-valued random variable with probability generating function

$$\varphi(s) = \sum_{u=-\infty}^{\infty} s^u \Pr[U = u] = E(s^u), \tag{2}$$

defined for s in the neighborhood of 1.

The version of the model in CHAKRABORTY and NEI (1982) considers the binomial special case of $\varphi(s)$. The generalization below is straightforward.

Based on WEHRHAHN (1975) and CHAKRABORTY and NEI (1982), the probability generating function

$$P(s, t) = E(s^{X_i - X_j})$$

of $X_i - X_j$ is given by

$$P(s, t) = P(s, 0)\exp[-a(s)t] + \frac{1}{2Na(s)}\{1 - \exp[-a(s)t]\}, \tag{3}$$

where

$$a(s) = \frac{1}{2N} - 2\nu[\psi(s) - 1], \tag{4}$$

and $\psi(s) = [\varphi(s) + \varphi(1/s)]/2$ is the symmetrized form of $\varphi(s)$. In the vicinity of $s = 1$ for which $a(s)$ is positive, the solution of this equation converges, as $t \to \infty$, toward

$$P(s) = \left(2N\left\{\frac{1}{2N} - 2\nu[\psi(s) - 1]\right\}\right)^{-1}, \tag{5}$$

which can be represented as

$$P(s) = \frac{1-p}{1 - p\psi(s)}, \tag{6}$$

where

$$p = \frac{4N\nu}{1 + 4N\nu}. \tag{7}$$

Under equilibrium conditions, the same result can be derived using the coalescent approach (APPENDIX). Before passing to variances, let us note that $\psi(s)$, the symmetrized form of the probability generating function $\varphi(s)$, does not embody any assumptions imposed on our formulation. On the contrary, its presence in the expressions (4)–(6) is a direct consequence of considering the difference of sizes of two randomly selected alleles ($X_i - X_j$), which is a random variable with a symmetric distribution. The most important consequence of this fact is that all the results we obtain are valid for general asymmetric (directionally biased) mutation mechanisms and not only for the symmetric single-step special case.

The variance of $X_i - X_j$,

$$V_t = \frac{\partial^2}{\partial s^2} P(s, t)\,|_{s=1}$$

$$= V_\infty + (V_0 - V_\infty) \exp[-t/(2N)], \quad (8)$$

where $V_\infty = (4N\nu)\psi''(1)$, is equivalent to (8) in CHAKRABORTY and NEI (1982). The within-population variance of allelic size X_i is $V_t/2$ in generation t.

Between-population variability: We begin by calculating the probability generating function of the random variables $Z_{1i} - Z_{2j}$, the difference of sizes of two alleles randomly selected from two subpopulations (say, 1 and 2), which resulted from a split at time 0 in an ancestral population.

The probability generating function $D(s, t)$ of $Z_{1i} - Z_{2j}$ at time t is equal to

$$D(s, t) = W_0(s) R(s, t), \quad (9)$$

where $W_0(s)$ is the probability generating function of the size difference between randomly selected alleles in the ancestral population at $t = 0$, while $R(s, t)$ represents the change in that difference during the time interval $[0, t]$. Based on the model assumptions, we obtain

$$R(s, t) = \exp\{(2\nu t)[\psi(s) - 1]\}, \quad (10)$$

which can be interpreted as the Poisson distribution of the number of mutation events compounded with the random size of the mutation events. If we denote $D_t = \mathrm{Var}(Z_{1i} - Z_{2j})$, then the above yields

$$D_t = V_0 + R''(1, t) = V_0 + (2\nu t)\psi''(1). \quad (11)$$

In the formulation of variance components analysis, Z_{1i} and Z_{2j} can be represented as

$$Z_{1i} = Y_1 + X_{1i},$$

$$Z_{2j} = Y_2 + X_{2j}.$$

Y_1 and Y_2 are exchangeable random variables representing the between-population variability. Likewise, X_{mn} are exchangeable random variables, independent of Y_1 and Y_2, representing the within-population variability in populations 1 and 2 [for an extended discussion of components of the genetic variance, see *e.g.*, COCKERHAM and WEIR (1987) and references therein]. Therefore,

$$\mathrm{Var}(Z_{1i} - Z_{2j}) = \mathrm{Var}(Y_1 - Y_2) + \mathrm{Var}(X_{1i} - X_{2j}). \quad (12)$$

We know that $\mathrm{Var}(Z_{1i} - Z_{2j}) = D_t$, (see Equation 11) and $\mathrm{Var}(X_{1i} - X_{2j}) = V_t$ (see Equation 1). The between-population variance at time t is equal to $B_t/2$ where $B_t = \mathrm{Var}(Y_1 - Y_2)$. Using Equation 12 we obtain

$$B_t = (D_t - V_t) = (V_0 - V_\infty)\{1 - \exp[-t/(2N)]\} + 2\nu t\psi''(1), \quad (13)$$

which is asymptotically equivalent to $2\nu t\psi''(1)$. If the ancestral population was at equilibrium at time $t = 0$ (*i.e.*, if $V_0 = V_\infty$), then

$$B_t = 2\nu t\psi''(1). \quad (14)$$

Under the same condition, $V_t = V_\infty$, and (8) and (14) yield the expression

$$2B_t/V_t = t/N. \quad (15)$$

This is the index introduced by CHAKRABORTY and NEI (1982), which is linear as a function of time of divergence of populations 1 and 2.

Relationship to the T_R index: SLATKIN (1995), based on a coalescence argument, introduced an index

$$T_R = 4R_{ST}/(1 - R_{ST}), \quad (16)$$

where $R_{ST} = (\bar{S} - S_W)/\bar{S}$, $\bar{S}$ is twice the estimated total variance of allele size in two populations pooled together, and S_W is twice the average of the estimated total variance of allele size within each population.

Let us suppose that n_1 and n_2 chromosomes have been sampled from each of the two populations, respectively. In the standard notation of analysis of variance (SOKAL and ROHLF 1981),

$$\bar{S} = 2 \cdot SS_{\mathrm{tot}}/(n_1 + n_2 - 1) = 2 \cdot MS_{\mathrm{tot}}, \quad (17)$$

$$S_W = 2 \cdot SS_{\mathrm{with}}/(n_1 + n_2 - 2) = 2 \cdot MS_{\mathrm{with}}. \quad (18)$$

It is known that $E(MS_{\mathrm{with}}) = V_t/2$ and $E(MS_{\mathrm{betw}}) = V_t/2 + n_0 B_t/2$, where n_0 is the harmonic mean of n_1 and n_2 in our notation for the components of population variance. This leads first to

$$E(S_W) = V_t, \quad (19)$$

and then to

$$E(\bar{S}) = [2/(n_1 + n_2 - 1)][E(SS_{\mathrm{betw}}) + E(SS_{\mathrm{with}})] = V_t + n_0 B_t/(n_1 + n_2 - 1), \quad (20)$$

which leads to

$$E(\bar{S} - S_W) = n_0 B_t/(n_1 + n_2 - 1) = 2n_1 n_2 B_t/[(n_1 + n_2 - 1)(n_1 + n_2)], \quad (21)$$

which is approximately equal to $B_t/2$ when both sample

TABLE 1

Number of chromosomes sampled at loci for the populations analyzed

Locus	Population[a]								
	CP	GR	BW	UP	CN	JP	SO	BE	BB
D13S71	138	196	50	72	98	48	220	100	50
D13S193	152	188	50	70	54	50	212	52	50
D13S124	156	196	50	72	98	44	278	100	48
FLT1	154	190	50	72	54	52	234	54	50
D13S121	156	192	50	64	50	50	222	52	50
D13S118	154	192	46	72	54	50	228	54	38
D13S197	154	188	50	70	52	52	216	54	46
D13S122	154	182	50	66	54	50	226	54	50
PLA2A	98	102	124	70	102	96	104	98	70
THO1	98	102	118	70	102	96	102	98	64
CSF1R	98	102	118	66	102	96	102	98	68
F13A1	70	98	100	72	102	94	78	100	84
CYP19	72	90	98	72	102	96	116	100	84
LPL	72	84	100	72	102	78	100	92	84
DM-CTG	160	104	100	72	102	100	106	100	84
SCA	94	92	96	52	50	72	86	98	80
DRPLA	92	100	100	68	100	90	98	100	84
HD-CAG	158	96	100	72	100	112	120	100	84

[a] The population names are abbreviated as follows: GR, German; CP, unrelated Caucasian from CEPH pedigree panel; BW, Brazilian White; UP, Uttar Pradesh; CN, Chinese; JP, Japanese; SO, Sokoto Nigerian; BE, Benin; BB, Brazilian Black.

sizes are large and are of comparable magnitude. Therefore, based on the definition of Slatkin's index in (16), and (20) and (21) above, we have

$$E(T_R) \cong 4E(\bar{S} - S_W)/E(S_W) \cong 2B_t/V_t. \quad (22)$$

This shows that Slatkin's result can be derived from CHAKRABORTY and NEI (1982) formulation.

If a population at equilibrium under mutation-drift balance splits into more than two subpopulations of identical size equal to that of the ancestral population, (16)–(22) hold with appropriate changes of n_1, n_2, ..., and n_0 (SOKAL and ROHLF 1981), so that (22) is still applicable. As before, V_t and B_t are the within- and between-populations components of variance of differences in allele sizes.

APPLICATION TO DATA ON REPEAT LOCI IN NINE HUMAN POPULATIONS

Recently, DEKA *et al.* (1995a,b) surveyed for worldwide genetic variation at eight dinucleotide (FLT1, D13S118, D13S121, D13S71, D13S122, D13S197, D13S193 and D13S124), five trinucleotide (PLA2A, DM, SCA, DRPLA and HD) and five tetranucleotide (THO1, CSF1R, F13A1, CYP19 and LPL) repeat loci. From these surveys we selected allele size distributions from nine populations (unrelated Caucasians from the CEPH panel; German; Brazilian Whites; Brahmins from Uttar Pradesh, India; Sokoto from Nigeria; Benin; Brazilian Blacks; Japanese and Chinese) for the present application. The anthropological description of the sampled populations are given in the original surveys (DEKA *et al.* 1995a,b). In Table 1 we present the sample sizes (number of chromosomes sampled) from each of their populations for 18 loci.

For each locus, the locus-specific distances for a pair of populations were estimated by $4(\bar{S} - S_W)/S_W$, in which $\bar{S}$ and S_W were computed from (17) and (18). The average distance matrix was computed by taking the simple arithmetic mean of the locus-specific distance matrices over the 18 loci. Table 2 shows the average distance matrix for all pairs of populations. The neighbor-joining dendrogram (SAITOU and NEI 1987) of this distance matrix is shown in Figure 1. As can be seen from this dendrogram, the average distances over all 18 loci group the populations by their major racial characteristics. The four Caucasian populations (CEPH, German, Brazilian Whites and Uttar Pradesh Brahmins) cluster together; other disjoint clusters are formed by the two Mongoloid populations (Chinese and Japanese) and the three populations of African ancestry (Sokoto Nigerians, Benin and Brazilian Blacks). The Caucasoid and Mongoloid populations cluster together first; the Africans are furthest apart.

DISCUSSION AND CONCLUSIONS

The theory developed above shows that SLATKIN's (1995) index of genetic distance for tandem repeat loci is identical to the one proposed by CHAKRABORTY and NEI (1982). While Slatkin's index has been derived under the assumption of the expected allele size change equal to 0, we demonstrate that the distribution of allelic changes by stepwise mutations can be arbitrary. This is caused by the fact that both variance components (B_t and V_t) in this derivation can be expressed

TABLE 2

The distance matrix

GR	**0.000**							
BW	**0.000**	**0.000**						
UP	**0.034**	**0.040**	**0.033**					
CN	0.238	0.217	0.160	0.150				
JP	0.187	0.159	0.138	0.116	**0.029**			
SO	0.317	0.342	0.231	0.228	0.194	0.183		
BE	0.349	0.363	0.327	0.313	0.269	0.255	**0.039**	
BB	0.152	0.144	0.144	0.125	0.118	0.087	**0.042**	**0.044**
	CP	GR	BW	UP	CN	JP	SO	BE

The distance index has been computed for all pairs of the nine populations. Submatrices of distances among members of major racial groups are set in boldface type. Abbreviations as in Table 1.

as expectations of squared differences of allele sizes (Eqs. 1 and 12). Thus, a directional bias of size changes by mutations does not affect the linearity of the relationship of the expected distance with the time of divergence of populations.

Our derivations are carried out at a more general level than those of SLATKIN (1995). He derived the mutation-drift equilibrium expectations of the first two moments, while we obtain the transient (Eqs. 3 and 9) as well as asymptotic (equilibrium) (Equation 6) expressions for the distribution of allele size differences, characterized by their probability generating functions. As a consequence, our theory, see for example (8) and (11), can also be used to study within- and between-population dynamics of allele size variation in the absence of mutation drift balance.

For example, if the ancestral population was completely homozygous for the locus in question, *i.e.*, if $V_0 = 0$, then based on (8) and (13) we obtain that our index is equal to $2B_t/V_t = x/[1 - \exp(-x/2)] - 2$, where $x = t/N$. Hence, $2B_t/V_t$ is less than x by a factor that depends on time of divergence expressed in units of the effective population size. If $t = 5N$, then $2B_t/V_t$ is less than x by a factor of 0.69; if $t = 20N$, then $2B_t/V_t$ is less than x by a factor of 0.90. This effect will lead to underestimation of t if equilibrium in the ancestral population is assumed, but in fact it did not exist. Careful analysis of the consequences is beyond the scope of this paper.

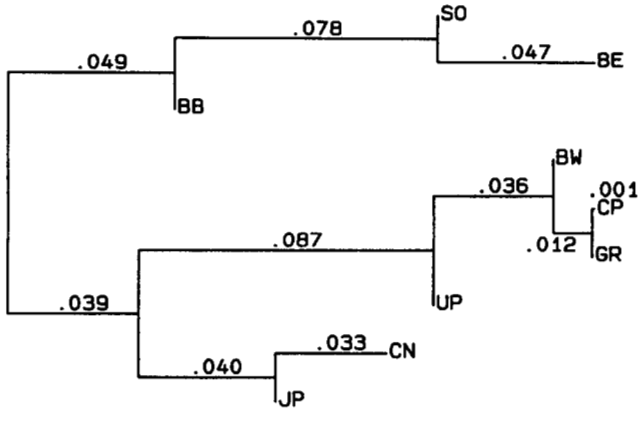

FIGURE 1.—The neighbor-joining dendrogram of the nine populations based on the distance indices in Table 1.

In the case of within-population variability the asymptotic but not transient result can also be obtained using a coalescence approach (APPENDIX). We should note that our approach can also be used to prove the time-linearity of indices of GOLDSTEIN *et al.* (1995), without their assumptions of single-step symmetric mutations.

The estimators S_W and S_B (and consequently $\overline{S}$), as given by Eqs. 9a, 9b and 10 in SLATKIN (1995), are unbiased estimators of the respective parameters only when an equal number of alleles are sampled from each population. In contrast, the variance components estimators (Equations 17 and 18) used in this work are unbiased in the general case as well.

Since our index is a ratio of components of variance, its sample value can be negative if $\overline{S} < S_W$. This may occur when two populations are genetically close and within each of them the genetic variation is considerable. Since distances between populations cannot be negative, we suggest using zero in situations when negative values are obtained. This does not mean that the populations are identical, but that their differences are dominated by statistical noise.

The application of the genetic distance measure shown above indicates that the index satisfactorily groups populations according to their known ethnohistoric clusters. However, caution has to be exercised in applications when populations analyzed include some that have been historically small, or known to have gone through recent bottlenecks during their history. The genetic distance indices among them as well as those between any of these and the larger populations may not conform to linearity. This is as expected, since the constancy of the population size (N) is a critical assumption of our derivation as well as that of SLATKIN (1995).

In principle at least, the population dynamics approach [CHAKRABORTY and NEI (1982); also see Equation 3] can be extended to any population for which

the size fluctuations over generations can be analytically specified. Technically, one way to accomplish this is to set $N = N(t)$ in the differential or difference equation leading to relationship (3). This was done for a special case of symmetric binomial $\varphi(s)$ by CHAKRABORTY and NEI (1977). Time linearity of $2B_t/V_t$ cannot be guaranteed any more, but for an assumed or estimated model of time change of $N(t)$ [*e.g.*, logistic growth of $N(t)$], the time of divergence can still be analytically related to $2B_t/V_t$.

We should also note that pooling tandem repeat data over loci may create some problems depending on how the pooling is carried out. In this presentation, we computed the distance indices separately for each locus, and then took the arithmetic mean over all loci. In contrast, one could also consider estimates for S_W and $\bar{S} - S_W$ based on data pooled over all loci, and then construct the index of the genetic distance. For 18 loci, considered in aggregate, this alternative approach yields a dendrogram almost identical to the one shown in Figure 1. However, this may not be true in general. For example, when the eight dinucleotide, eight trinucleotide or five tetranucleotide loci were considered separately, these two approaches yielded considerable differences in the distance matrices as well as in the resulting dendrograms. At this stage the actual causes of such discrepancies cannot be identified, but the sampling variances of the estimates of S_W and $\bar{S} - S_W$ are likely to contribute to this. Under the model assumptions the ratio of $(\bar{S} - S_W)/S_W$ estimates a parameter that is independent of the mutation rate at a locus as well as the distribution of allelic size changes caused by mutation. This fact can be used as a rationale of taking a simple average of the distance indices over loci. A more detailed theoretical treatment of this problem requires evaluation of both the stochastic and the contemporary sampling variance of the ratio estimates proposed.

We thank Professor OLLE NERMAN of the University of Gotheborg for providing insights underlying the derivation in the APPENDIX. This work was supported by grants GM-41399 (R.C. and D.S.) and GM-45861 (R.D. and R.C.) from the National Institutes of Health, and DMS 9203436 and DMS 9409909 (M.K.) from the National Science Foundation and by the Keck's Center for Computational Biology at the Rice University (M.K.). Part of this work was carried out during M.K.'s visit at the University of Gotheborg in September 1995.

LITERATURE CITED

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

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

CHAKRABORTY, R., and M. NEI, 1982 Genetic differentiation of quantitative characters between populations of species: I. Mutation and random genetic drift. Genet. Res. Camb. **39:** 303–314.

COCKERHAM, C. C., and B. S. WEIR, 1987 Correlations, descent measures: drift with migration and mutation. Proc. Natl. Acad. Sci. USA **84:** 8512–8514.

DEKA, R., L. JIN, M. D. SHRIVER, L. M. YU, S. DECROO *et al.*, 1995a Population genetics of dinucleotide $(dC\text{-}dA)_n\cdot(dG\text{-}dT)_n$ polymorphisms in world populations. Am. J. Hum. Genet. **56:** 461–474.

DEKA, R., M. D. SHRIVER, L. M. YU, R. E. FERRELL, and R. CHAKRABORTY, 1995b Intra- and inter-population diversity at short tandem repeat loci in diverse populations of the world. Electrophoresis **16:** 1559–1564.

DI RIENZO, A., A. C. PETERSON, J. C. GARZA, A. M. VALDES, M. SLATKIN *et al.*, 1994 Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA **91:** 3166–3170.

EWENS, W. J., 1979 *Mathematical Population Genetics.* Springer, New York.

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

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

JEFFREYS, A. J., K. TAMAKI, A. MACLEOD, D. G. MONCKTON, D. L. NEIL *et al.*, 1994 Complex gene conversion events in germline mutation at human minisatellites. Nature Genet. **6:** 136–145.

KIDD, J. R., F. L. BLACK, K. M. WEISS, I. BALAZS and K. K. KIDD, 1991 Studies of three Amerindian populations using nuclear DNA polymorphisms. Hum. Biol. **63:** 775–794.

LI, W.-H., 1976 Electrophoretic identity of proteins in a finite population and genetic distance between taxa. Genet. Res. Camb. **28:** 119–127.

RUBINSZTEIN, D. C., W. AMOS, J. LEGGO, S. GOODBURN, S. JAIN *et al.*, 1995 Microsatellite evolution—evidence for directionality and variation in rate between species. Nature Genet. **10:** 337–343.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

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.

SHRIVER, M. D., L. JIN, E. BOERWINKLE, R. DEKA, R. E. FERRELL *et al.*, 1995 A novel measure of genetic distance for highly polymorphic tandem repeat loci. Mol. Biol. Evol. **12:** 914–920.

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

SOKAL, R. R., and F. J. ROHLF, 1981 *Biometry.* Freeman, New York.

TAVARE, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Pop. Biol. **26:** 119–164.

TAVARE, S., 1995 Calibrating the clock: using stochastic processes to measure the rate of evolution, pp. 114–152 in *Calculating the Secrets of Life,* edited by E. S. LANDER and M. S. WATERMAN. National Academy Press, Washington, DC.

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.

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

Communicating editor: W. J. EWENS

APPENDIX: COALESCENT DERIVATION OF THE PROBABILITY GENERATING FUNCTION $P(s)$

Background mathematics and relevant references for this appendix can be found in either of the two reviews, TAVARE (1984) or TAVARE (1995).

For any two alleles X_i and X_j drawn from the population at equilibrium, the time τ to coalescence is exponentially distributed with parameter $1/(2N)$. Conditional on τ, the number of mutation events in $[-\tau, 0]$, in both alleles taken jointly, is Poisson with parame-

ter $2\nu\tau$. Therefore, the probability generating function of the number of mutation events is equal to

$$\begin{aligned}\alpha(s) &= \int_0^\infty \alpha(s|\tau) f(\tau)\, d\tau \\ &= \int_0^\infty \exp[2\nu\tau(s-1)]\, \frac{1}{2N} \exp[-\tau/(2N)]\, d\tau \\ &= \frac{1-p}{1-ps},\end{aligned}$$

with p as in Equation 7. This corresponds to the geometric distribution. The mutation process in each allele separately can be viewed as a random walk with step size being a random variable with probability generating function $\varphi(s)$. The contribution of each mutation event to $X_i - X_j$ is a random variable with probability generating function $\psi(s) = [\varphi(s) + \varphi(1/s)]/2$, since mutations occur alternately on randomly chosen alleles. Therefore, the probability generating function of $X_i - X_j$ is equal to

$$P(s) = \alpha[\psi(s)] = \frac{1-p}{1-p\psi(s)},$$

consistent with (6).