Estimation of Genetic Distance and Coefficient of Gene Diversity from Single-Probe Multilocus DNA Fingerprinting Data

Li Jin and Ranajit Chakraborty
Center for Demographic and Population Genetics, University of Texas Houston Health Science Center

DNA fingerprinting exhibits multilocus genotypes of individuals, detected by the use of a single multilocus probe. Consequently, population data on DNA fingerprinting do not provide a complete characterization of the genetic variation in terms of allele-frequency distributions, since neither the number of loci nor the locus affiliation of alleles is directly observable. Yet DNA fingerprinting has been proved to be a cost-effective method of detecting hypervariable polymorphisms in several organisms, where the traditional loci fail to detect enough variation for microevolutionary studies. In the present paper we demonstrate that the above-mentioned features of DNA fingerprinting data do not cause any serious problem when they are used in evolutionary studies. Bias-corrected estimators of Nei's standard and minimum genetic distances are derived, and, by an application of this theory to data on seven short tandem repeat loci in three major human populations, it is shown that these modified measures of genetic distances based on DNA fingerprint patterns are quite close to Nei's distances based on locus-specific allele frequencies. Empirical as well as theoretical support of the adequacy of such genetic distances from DNA fingerprinting data is also discussed, and it indicates that the technical limitations of DNA fingerprinting should not deter the use of the method for short-term evolutionary studies.

Introduction

The polymorphic loci known as variable number of tandem repeat (VNTR) loci are highly polymorphic in humans and in many other organisms (Burke and Bruford 1987; Nakamura et al. 1987; Gilbert et al. 1990; Georges et al. 1991; Ely et al. 1992). They are characterized by a large number of alleles per locus and high heterozygosity. Consequently, the VNTR loci, as a new class of genetic marker, have proved to be extremely useful in gene mapping (Nakamura et al. 1987; Hearne et al. 1992), forensic identification of individuals (Jeffreys et al. 1985c; Chakraborty and Kidd 1991), determination of relatedness of individuals (Jeffreys et al. 1985a, 1991; Chakraborty and Jin 1993a, 1993b), and evolutionary studies of closely related populations or species (Gilbert et al. 1990; Jin and Chakraborty, accepted).

A DNA fingerprint is the pattern revealed by a single multilocus probe (MLP) that simultaneously detects many minisatellite loci in the genome, by Southern blot analysis (Jeffreys et al. 1985b, 1988; Wong et al. 1986, 1987; Nakamura et al. 1987). Minisatellite loci are the VNTR loci whose repeat units are > 15 bp long. The genetic basis of a DNA fingerprint revealed by an MLP is that the probe is homologous to a large number of chromosomally dispersed genomic loci, many of which exhibit considerable genetic polymorphism (Nakamura et al. 1987; Armour et al. 1990). The simultaneous screening of many hypervariable loci makes DNA fingerprinting by MLPs efficient and cost-effective.

Several efforts have been made to utilize MLP data for evolutionary studies, since the first demonstration of their utility in genetic studies (Jeffreys et al. 1985b). However, it is noteworthy that the existing statistical measures, including Nei's genetic distances (Nei 1987), are not directly applicable to MLP data, since the number of loci detected by an MLP is unknown and since the allele frequencies at each locus are undetermined. Lynch (1990, 1991) proposed an empirical similarity measure on the basis of allele sharing. A very similar approach was used by Yuhki and O'Brien (1990) and Gilbert et al. (1990, 1991) in their analyses. Unfortunately, the lack of a theoretical study of the relationship of those genetic distance measures with the time of divergence makes the approach less appealing. However, it has been shown that measures of genetic distance from some summary statistics based on allele sharing can be

Key words: DNA fingerprint, genetic distances, coefficient of gene diversity, population genetics.

Address for correspondence and reprints: Dr. Ranajit Chakraborty, Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, P.O. Box 20334, Houston, Texas 77225.

© 1994 by The University of Chicago. All rights reserved.
0737-4038/94/1101-0012$02.00
used for evolutionary studies (Chakraborty and Jin 1993b; Jin and Chakraborty, accepted), since the mutation-drift expectations of such genetic distances theoretically retain a linear relationship with the time of divergence of two populations studied (t) when t is not very large (say, \( t \leq N_e \) generations, where \( N_e \) is the effective size of the population).

A statistic proposed by Stephens et al. (1992) to estimate heterozygosity shed some light on the utility of traditional genetic distance measures based on allele-frequency distributions for DNA fingerprint data. They showed that the specific distribution of allele frequencies for each locus is not needed in estimating the average heterozygosity. Therefore, the estimation of average heterozygosity is possible because the frequency of each allele can be estimated from the frequency of the corresponding band. However, their estimator is biased. Jin and Chakraborty (1993) therefore proposed bias-corrected estimators of the average heterozygosity and of the number of loci, from DNA fingerprint band-frequency data in a population sample.

In the present paper, we argue that genetic distance measures (e.g., Nei’s minimum genetic distance and Nei’s standard genetic distance) can be written in a form in which the specific distributions of allele frequencies for each locus are not needed, and that, in turn, these genetic distances can be estimated from the frequencies of the bands corresponding to the alleles. With the same logic, Nei’s (1973) coefficient of gene diversity, \( G_{ST} \), can also be estimated. The application of various genetic distance measures to a group of short tandem repeat (STR) loci (Edwards et al. 1992) indicates that Nei’s genetic distances do not differ from their modified version for DNA fingerprint data, on the basis of MLP. The technical limitations of MLP data are also discussed.

Theory

The Alternative Expressions of Nei’s Genetic Distances

Nei’s (1972) minimum genetic distance is defined by

\[
D_m = [(J_X + J_Y)/2] - J_{XY} ,
\]

and the standard genetic distance (Nei 1972) is defined by

\[
D_s = [(\log J_X + \log J_Y)/2] - \log J_{XY} ,
\]

where

\[
J_X = \sum_{i=1}^{L} \sum_{l=1}^{n_i} p_{lx(l)} / L ,
\]

\[
J_Y = \sum_{i=1}^{L} \sum_{l=1}^{n_i} p_{ly(l)} / L ,
\]

\[
J_{XY} = \sum_{i=1}^{L} \sum_{l=1}^{n_i} p_{lx(l)} p_{ly(l)} / L ,
\]

in which \( p_{lx(l)} \) and \( p_{ly(l)} \) are the frequencies of the \( i \)th allele of the \( l \)th locus from populations \( X \) and \( Y \), respectively. \( L \) is the number of loci, and \( n_i \) is the number of alleles at the \( i \)th locus. Note that \( J_X, J_Y, \) and \( J_{XY} \) can also be written as

\[
J_X = \sum_{k=1}^{A} p_{kX} / L ,
\]

\[
J_Y = \sum_{k=1}^{A} p_{kY} / L ,
\]

\[
J_{XY} = \sum_{k=1}^{A} p_{kX} p_{kY} / L ,
\]

where \( p_{kX} \) and \( p_{kY} \) are the frequencies of the \( k \)th allele among all alleles, regardless of their locus affiliation, and \( A \) is the total number of alleles at \( L \) loci considered (\( A = \sum_{l=1}^{L} n_l \)).

The Estimation of Number of Loci and Heterozygosity

Since all bands (alleles) observed are effectively dominant, the expectation of the relative frequency of occurrence of the \( k \)th band (allele), \( s_k \), is given by

\[
E(s_k) = 2p_k - p_k^2 ,
\]

where \( p_k \) is the frequency of the \( k \)th allele among all alleles for one particular population, and Hardy-Weinberg equilibrium is assumed (Stephens et al. 1992). Furthermore, by solving equation (4), \( p_k \) can be estimated by

\[
\hat{p}_k = 1 - \sqrt{1 - s_k} .
\]

It was suggested by Jin and Chakraborty (1993) that the average heterozygosity in a population (\( H \)) and the number of loci (\( L \)) for single-probe multilocus DNA fingerprints can be estimated (bias-corrected) by
\[
\hat{L} = L_m + \sum_{k=1}^{A_p} \left( 1 - \frac{s_k}{1 - s_k} \right) - \frac{1}{8n} \sum_{k=1}^{A_p} \frac{s_k}{1 - s_k},
\]

and

\[
\hat{H} = \left( \sum_{k=1}^{A} \frac{s_k}{L} \right) - 1,
\]

where \(A_p\) is the total number of alleles at all polymorphic loci, \(L_m\) is the number of monomorphic loci, and \(n\) is the number of individuals sampled from the population. Since \(J = 1 - H\), the bias-corrected estimates of \(J_x\) and \(J_y\) can be obtained from equation (7).

The Estimation of Genetic Distance for DNA Fingerprinting

Following a similar logic, we now derive a bias-corrected estimate of \(J_{xy}\) which may lead to bias-corrected estimates of Nei’s genetic distances (for both minimum distance and standard distance).

The probability \((J_{xy})\) that two alleles, one each chosen from of the two populations, are identical can be estimated by substituting \(p_k’s\) by using equation (5), yielding

\[
\hat{J}_{xy} = \frac{\sum \hat{p}_{kx}\hat{p}_{ky}}{\hat{L}}
\]

\[
= \sum_{k=1}^{A} \frac{(1 - \sqrt{1 - s_{kx})(1 - \sqrt{1 - s_{ky}})}}{\hat{L}},
\]

where \(s_{kx}\) and \(s_{ky}\) are the \(s_k’s\) in populations \(X\) and \(Y\), respectively, and \(\hat{L}\) is the estimated number of loci.

It can be shown that

\[
E[\hat{J}_{xy}] \approx J_{xy} + \sum_{k=1}^{A} \left[ \frac{s_{kx}}{1 - s_{kx}} \left( \frac{1 - \sqrt{1 - s_{ky}}}{n_x} \right) \right]
\]

\[
+ \left[ \frac{s_{ky}}{1 - s_{ky}} \left( \frac{1 - \sqrt{1 - s_{kx}}}{n_y} \right) \right] \frac{1}{8\hat{L}},
\]

where \(n_x\) and \(n_y\) are the number of individuals sampled from populations \(X\) and \(Y\), respectively. The approximation

\[
E[(1 - \sqrt{1 - s_{kx})(1 - \sqrt{1 - s_{ky}})}]
\]

\[
\approx p_{kx}p_{ky} + V(s_{kx})p_{kx}(1 - p_{kx})^{-3}/8
\]

\[
+ V(s_{ky})p_{ky}(1 - p_{ky})^{-3}/8,
\]

which is based on the Taylor series expansion, and the binomial distribution of \(s_{kx}\) and \(s_{ky}\), leading to

\[
V(s_{kx}) = E(s_{kx})[1 - E(s_{kx})]/n_x,
\]

and

\[
v(s_{ky}) = E(s_{ky})[1 - E(s_{ky})]/n_y,
\]

are used in deriving equation (9).

This yields a bias-corrected estimate of \(J_{xy}\),

\[
\hat{J}_{xy} = \hat{J}_{xy} - \Delta,
\]

where

\[
\Delta = \sum_{k=1}^{A} \left[ \frac{s_{kx}}{1 - s_{kx}} \left( 1 - \sqrt{1 - s_{ky}} \right)/n_x \right]
\]

\[
+ \frac{s_{ky}}{1 - s_{ky}} \left( 1 - \sqrt{1 - s_{kx}} \right)/n_y \right] / 8\hat{L},
\]

with \(\hat{L}\) estimated by taking the average of the two population-specific estimates of \(L\) obtained from equation (6).

Therefore, the bias-corrected estimates of Nei’s minimum distance and standard distance can be obtained by substituting \(\hat{J}_{xy}, \hat{J}_x,\) and \(\hat{J}_y\) in equations (1) and (3). These become

\[
\hat{D}_{mf} = [(\hat{J}_x + \hat{J}_y)/2] - \hat{J}_{xy},
\]

and

\[
\hat{D}_{sf} = [(\log \hat{J}_x + \log \hat{J}_y)/2] - \log \hat{J}_{xy}.
\]

The Estimation of Coefficient of Gene Diversity

Nei (1973) defined the coefficient of gene diversity, \(G_{ST}\), as

\[
G_{ST} = \frac{D_{ST}}{H_S + D_{ST}},
\]

where \(D_{ST} = \sum_k \sum_i D_{ki}/s^2\), in which \(D_{ki}\) is Nei’s (1972) minimum genetic distance between the \(kth\) and \(ith\) populations, and \(H_S = \sum_k H_k/s\), in which \(H_k\) is the heterozygosity of \(kth\) population. By using equation (13), \(G_{ST}\) can be estimated by

\[
\hat{G}_{ST} = \frac{\sum_k \sum_i \hat{D}_{ki}}{s \sum_k H_k + \sum_k \sum_i \hat{D}_{ki}},
\]

where equations (13) and (7) are used for estimating \(\hat{D}_{ki}\) and \(\hat{H}_k\), respectively.

Numerical Results

Recently, Edwards et al. (1991) described several STR loci, each of which demonstrates considerable de-
Table 1
Estimation of the Number of Loci and Heterozygosity for STR Loci

<table>
<thead>
<tr>
<th>METHOD AND POPULATION</th>
<th>ESTIMATE OF NO. OF LOCI</th>
<th>ESTIMATES OF HETEROZYGOSITY FROM MULTILOCUS DATA</th>
<th>ESTIMATES OF HETEROZYGOSITY FROM ALLELE FREQUENCIES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Five locus:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>5.04</td>
<td>0.635</td>
<td>0.638</td>
</tr>
<tr>
<td>Blacks</td>
<td>5.03</td>
<td>0.733</td>
<td>0.730</td>
</tr>
<tr>
<td>Asians</td>
<td>4.71</td>
<td>0.628</td>
<td>0.607</td>
</tr>
<tr>
<td>Seven locus:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>6.88</td>
<td>0.700</td>
<td>0.694</td>
</tr>
<tr>
<td>Blacks</td>
<td>6.88</td>
<td>0.762</td>
<td>0.762</td>
</tr>
<tr>
<td>Asians</td>
<td>6.84</td>
<td>0.669</td>
<td>0.654</td>
</tr>
</tbody>
</table>

Degrees of polymorphism within populations. The population genetic characteristics of five of these loci (TH01, RENA4, FARB, HPRTB, and ARA) are described elsewhere (Edwards et al. 1992) in Caucasians (200 individuals), American blacks (200 individuals), and Asians (80 individuals) currently residing in Houston. Two more STR loci (CD4 and PLA2Al) have been typed recently for the same individuals from the populations mentioned above (H. Hammond, personal communication).

For each individual from whom genotype data at all loci are available, an MLP DNA fingerprinting pattern can be generated by merging the genotype data of the seven STR loci together, so that each individual would exhibit 7 (for those who are homozygous at all loci) to 14 distinct bands (alleles, for those who are heterozygous at all loci). Therefore, such merged data mimics a seven-locus DNA fingerprinting profile for each individual sampled. Table 1 shows the estimates of the number of loci (by using eq. [6]) and average heterozygosity (by using eq. [7]), which may be contrasted with the true number of loci and the unbiased estimate (Nei 1978) of the average heterozygosity based on locus-specific allele-frequency data. We present such comparisons with two types of data merging (five locus and seven locus), since two (HPRTB and ARA) of the seven STR loci scored in this survey of three populations are X linked. Therefore, the five-locus genotype data consist of genotypic information on all individuals (male and female) sampled (373 total), while the seven-locus data consist of genotypic information on the females only (yielding 151 individuals total). For both comparisons, the estimates of $H$ from the multilocus genotype data (mimicking DNA fingerprinting) are reasonably close to the estimates based on locus-specific allele frequencies, and the error of prediction of the true number of loci is also small. The estimates for the Asian sample exhibit a greater deviation. This is probably because the sample size for this case is the smallest (71 for the five-locus data, and 27 for the seven-locus data). Of course, since the Asian sample is a composite of individuals of several different national populations (e.g., Taiwanese, Filipino, Japanese, Chinese, and Vietnamese), possible substructuring effects may also influence such departures. It should be recalled, however, that Edwards et al. (1992) did not find any significant departure from the Hardy-Weinberg expectations of locus-specific genotype frequencies in this sample.

The computations of the modified distances ($D_{sf}$ from eq. [14] and $D_{mf}$ from eq. [13]) are shown in table 2 for these merged DNA profiles. For comparison, we also computed the bias-corrected Nei’s standard distance and minimum distance estimates (Nei 1978) on the basis of the locus-specific allele frequencies, as reported by Edwards et al. (1992). A fifth estimate of genetic distance, $D_i$, also shown in this table, is based on allele sharing between individuals, $D_i = 1 - [n_b/(n_{w1} + n_{w2})]$, where $n_{w1}$ and $n_{w2}$ are the average numbers of allele shared between individuals within the two populations, and $n_b$ is the average number of shared alleles between two individuals drawn one each from the two populations (Jin and Chakraborty, accepted). These values ($n_{w1}$, $n_{w2}$, and $n_b$) were computed by taking averages over all possible pairs of individuals in the sample. We considered this measure of genetic distance, since Jin and Chakraborty (accepted) have shown that the evolutionary dynamics of $D_i$ is similar to that of Nei’s standard genetic distance ($D_m$) on the basis of locus-specific allele frequencies when the time of divergence between two populations is small, even though the computation of the statistic $D_i$ does not require the information of locus affiliation of alleles. Since Nei’s minimum distance measures ($D_m$ and $D_{mf}$) have an asymptotic value of

Table 2
Estimates of Nei’s Distances Obtained from Allele-Frequency Data and from DNA Fingerprint Data for STR Loci

<table>
<thead>
<tr>
<th>POPULATIONS COMPARED</th>
<th>GENETIC DISTANCES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$D_{sf}$</td>
</tr>
<tr>
<td>Five locus:</td>
<td></td>
</tr>
<tr>
<td>C-B</td>
<td>0.1560</td>
</tr>
<tr>
<td>C-A</td>
<td>0.1133</td>
</tr>
<tr>
<td>B-A</td>
<td>0.1510</td>
</tr>
<tr>
<td>Seven locus:</td>
<td></td>
</tr>
<tr>
<td>C-B</td>
<td>0.1483</td>
</tr>
<tr>
<td>C-A</td>
<td>0.0921</td>
</tr>
<tr>
<td>B-A</td>
<td>0.1401</td>
</tr>
</tbody>
</table>

NOTE. C = Caucasian; B = American black; and A = Asian.
Therefore, we conclude that the incomplete information (i.e., the Dm,f) and Y, do not share any allele at all loci), we divided (J_ + J_y)/2 (which occurs when the two populations, X and Y, do not share any allele at all loci), we divided these measures by the average homozygosity, (J_X + J_r)/2, in the two populations, so that the genetic distance values have a range of 0 to 1.

As in table 1, the genetic distances between the three pairs of populations were computed separately for merged genotype data for five loci and seven loci. These numerical illustrations indicate that the multilocus genotype–based estimates of genetic distances (D_jf and D_m,f) are close to the respective estimates based on locus-specific allele frequencies (D_j and D_m), even though the former ones utilize the frequency of each band (allele) from DNA fingerprint patterns without the recognition of locus affiliation of bands (alleles). The band (allele) sharing–based estimate (D_j) is also close to the estimate of Nei’s standard distance (D_s, based on locus-specific allele frequencies) or its modified version (D_jf, based on the multilocus data), which is consistent with the prediction of Jin and Chakraborty (accepted). Therefore, we conclude that the incomplete information (i.e., the unknown number of loci and the lack of knowledge of locus affiliation of alleles) in DNA fingerprinting data has little effect on the estimation of genetic distances between populations. The bias corrections, however, are critical, since without them (i.e., when the last term of eq. [6] and the term Δ of eq. [11] are ignored), the genetic distances would have been overestimated to the extent of 5%–37.5% of the amounts shown in table 2.

The coefficient of gene diversity, G_{STR}, estimated from the DNA fingerprinting data by using equation (16), for these three populations becomes 6.46% for the five-locus data and 6.00% for the seven-locus data. Those values are virtually identical (6.40% and 5.95%, respectively) to the estimates from locus-specific allele frequencies, using Nei and Chesser’s (1983) method.

**Discussion**

We have shown elsewhere that a bias correction for estimating the average heterozygosity and number of loci from DNA fingerprinting data is important (Jin and Chakraborty 1993), and, because of the sampling property of s_k and its relationship with allele frequencies (eq. [4]), the method of bias correction suggested by Nei (1978) and invoked by Stephens et al. (1992) must be modified in this context. In the present paper we observe that the same modified method of bias correction applies to the estimation of genetic distances as well, and in this case the bias correction is even more important. For example, in the numerical examples of the STR loci for the three major human populations given above, the bias of genetic distance estimates is 5%–37.5%, when the last term of equation (6) and the term Δ of equation (11) are not used. Therefore, we conclude that, even with our approximations, the need for bias correction is even more critical in the estimation of genetic distances, as compared with the estimation of the average heterozygosity and number of loci alone.

The close correspondence between the estimates based on locus-specific allele frequencies and the ones from multilocus data demonstrates that, even though the DNA fingerprinting patterns neither provide information with regard to the total number of loci underlying such DNA profiles nor recognize the locus affiliation of alleles, they may be used to estimate most of the important summary measures of genetic variation for evolutionary studies (e.g., the average heterozygosity, genetic distance, and the coefficient of gene diversity). The technical limitations of DNA fingerprinting do not seriously affect the predictability of these measures of genetic variation, as shown in the illustration above.

The inference regarding the average heterozygosity and genetic distance, based on DNA fingerprinting data, requires two additional assumptions that are not needed when each locus is individually characterized. First, the Hardy-Weinberg rule is assumed for each population, when the allele frequencies at the underlying loci in DNA fingerprinting data are estimated from the relative frequencies of each band within populations (s_k; see eq. [5]). This assumption has been a subject of considerable debate in recent years (Lander 1989; Cohen 1990; Le- wontin and Hart 1991). But, it has been widely demonstrated, in terms of traditional genetic markers (Mourant et al. 1976), as well as with hypervariable polymorphic loci (Devlin et al. 1990; Chakraborty and Kidd 1991; Chakraborty et al. 1991, 1993; Deka et al. 1992, and accepted; Devlin and Risch 1992; Edwards et al. 1992; Weir 1992), that even the cosmopolitan populations that are demographically substructured do not exhibit any substantial departure from the Hardy-Weinberg expectation of genotype frequencies. Population substructure affects the estimation of the total number of segregating alleles within a population, and, even there, in the presence of population substructure, an apparent excess of only the rare alleles is observed (Chakraborty et al. 1988; Chakraborty 1990). Since the rare alleles do not contribute substantially either in the evaluation of the average heterozygosity or in the estimation of genetic distance (Nei 1978), we contend that the presence of population substructure should not greatly affect the estimation of allele frequencies from the relative frequencies of specific fingerprinting bands (s_k). The number of individuals sampled is probably more important, since, when the sample size is small, the imprecision of the estimate of s_k values may accumulate more error than any real effect of substructuring.

Second, in our illustration of merged genotype data we treated the different alleles from different loci as sep-
arate bands; hence, the alleles are equated to bands. In contrast, in real DNA fingerprint data, comigration of DNA fragments in electrophoresis, resulting from alleles at different loci, may violate this assumption. While empirical data on the extent of comigration in DNA fingerprinting are not yet available, there are some studies suggesting that the assumption of no comigration may not be a serious problem in using DNA fingerprinting data for evolutionary studies. For example, in a position-by-position analysis of fragment sizes, Krawczak and Bockel (1992) and Bockel et al. (1992) showed that the problem of comigration may be examined by postulating “position-specific genetic factors,” F, an unobservable variable. When \( F > 0 \), an individual’s DNA fingerprint pattern would exhibit the presence of bands at that position. The relative frequencies of the presence of bands at specific positions (x, in the terminology of Bockel et al. 1992), in turn, predict how large the values of \( F \) can be in a DNA fingerprint database from population surveys. In general, values of \( x \) do not exceed 0.15–0.25, so that the probability of \( F > 2 \) is <0.035, under the assumption that \( F \) is distributed as a Poisson variate. This suggests that comigration of alleles at different loci is not a very common phenomenon. Furthermore, there is no evidence that alleles of any specific frequency would be more likely to comigrate than the ones that form distinctly different bands.

DNA fingerprinting data and the data generated by several probes must be used with some reservations for evolutionary studies for other reasons as well (e.g., see Kidd et al. 1991). For example, since polymorphisms at such loci are due to copy-number variation of tandemly repeated units and since the exact mechanism of production of new alleles is not yet known, their utility for studying genetic affinities of populations has been questioned. However, although new genetic variability at such loci seems to be generated by complex processes, such as strand slippage during DNA replication or non-homologous sister-chromatid exchange, recent studies suggested that different classes of VNTR loci have patterns of genetic variation that agree consistently with the two main mutation-drift models of population genetics. There is empirical (Levinson and Gutman 1987a, 1987b; Kwiatkowski et al. 1992; Schlötterer and Tautz 1992; Bowcock et al. 1993; Mahtani and Willard 1993) as well as theoretical (Shriver et al. 1993; Valdes et al. 1993) support of the assumption that the hypervariable loci whose repeat units are short (1–5 bp, called “microsatellite,” or “STR,” loci) follow the predictions of the stepwise-mutation model (Ohta and Kimura 1973). In contrast, minisatellite loci with repeat units ≥15 bp follow the prediction of the infinite-allele model (Clark 1987; Jeffreys et al. 1988; Flint et al. 1989). Furthermore, even though a full dissection of all loci underlying the MLPs used for DNA fingerprinting is yet to be made, it is known that these loci are of minisatellite nature (Jeffreys et al. 1985b; Wong et al. 1987; Armour et al. 1990) and that they are located on chromosomal bands dispersed in the genome that are identifiable from in situ hybridization of metaphase chromosomes (Royle et al. 1988; Zischler et al. 1989). Therefore, the evolutionary dynamics of measures of genetic distances \( (D_m, D_f) \) that are based on the DNA fingerprint should have properties identical to the locus-specific, allele frequency-based measures of Nei’s minimum \( (D_m) \) and standard \( (D_f) \) distances under the infinite-allele model (Li and Nei 1975).

In summary, the present work demonstrates that, even though the DNA fingerprinting data give neither a full characterization of all loci nor the allele frequency distributions at individual loci, they are useful for evolutionary studies, and Nei’s genetic distance can be used for such purposes. Since the hypervariability at such loci is caused by a high mutation rate (Jeffreys et al. 1988), DNA fingerprinting data may be useful only for studying short-term evolution. The abundance of minisatellite loci detected by DNA fingerprinting has been noted in birds (Burke and Bruford 1987; Hillel et al. 1989; Burke et al. 1991), marmosets (Dixson et al. 1988), dogs and cats (Jeffreys and Morton 1987), foxes (Gilbert et al. 1990), lions (Gilbert et al. 1991), and many other animals (see citations in Kelly et al. 1989). Therefore, the use of the method developed in this paper will not be restricted to microevolutionary studies in humans alone. Indeed, for some of the species listed above, genetic distances based on classical genetic markers are almost useless for studying evolutionary relationships between closely related populations and species, because of an extremely low level of genetic variation. In contrast, DNA fingerprint variation in highly inbred populations is seen to be discriminatory enough to validate the history of establishment of such populations (Gilbert et al. 1990).

**Acknowledgments**

This research was supported by U.S. Public Health Service research grants GM 41399 and GM 45816 from the U.S. National Institutes of Health and by grant 92-IJ-CX-K024 from the National Institute of Justice. We thank Dr. M. Nei and an anonymous reviewer for their constructive suggestions that improved the presentation of the results.

**LITERATURE CITED**

126 Jin and Chakraborty


MASATOSHI NEI, reviewing editor

Received June 17, 1993

Accepted September 13, 1993