Genet. Res., Camb. (1994), **63**, *pp.* 1–9 *With 2 text-figures*

Population dynamics of DNA fingerprint patterns within and between populations

LI JIN AND RANAJIT CHAKRABORTY*

Center for Demographic and Population Genetics, Graduate School of Biomedical Sciences, University of Texas at Houston Health Science Center, Houston, Texas

(*Received 11 May 1993 and in revised form 16 November 1993*)

Summary

DNA fingerprint is a pattern of a variable number of bands (DNA fragments) with different sizes on a Southern gel for each individual, generated by one or many VNTR loci. Genetic divergence between individuals within and between populations can be studied in terms of number of shared bands between individuals. Using a population genetic model we show that the expectations of measures of genetic distance between populations based on band sharing data from DNA fingerprint patterns are functions of composite parameters $M = 4N\nu$, and time of divergence (t) between populations, where N is the effective size of the populations, and ν, the mutation rate. The expected genetic distance remains linear with time of divergence at least up to N generations as long as the average heterozygosity at the DNA fingerprint loci remains at or below 90%. Neither incomplete knowledge of the allele frequencies at each locus, nor the unknown number of loci underlying DNA fingerprint pattern, compromise these evolutionary dynamics of DNA fingerprint patterns. Applications of this theory to data on three human populations, and review of literature indicate that co-migration of alleles, and the presence of syntenic loci underlying the fingerprint pattern have little impact of the reliability of evolutionary conclusions from DNA fingerprint studies.

1. Introduction

Evolutionary studies of closely related populations or species are efficient when the genetic alterations between the contrasting units (groups) are large enough to make them genetically somewhat distinct from each other. With a low substitution rate (e.g. 1×10^{-9} per site per generation; Nei, 1987), long segments of DNA sequences may be required for short-term evolutionary studies, which makes the approach of utilizing DNA sequences currently not quite efficient (Saitou & Nei, 1986). However, variable number of tandem repeat (VNTR) loci which are characterized by high mutation rates (up to 0·001, or 0·05 per locus per generation; Jeffreys *et al.* 1988) may lessen this problem to some extent. Due to the high rate of mutation leading to changes in the number of repeats of a short DNA sequence, VNTR loci are characterized by high heterozygosities and large numbers of alleles per locus (Nakamura *et al.* 1987; Wong *et al.* 1987). Because of their hypervariability, VNTR loci have been extremely useful in gene mapping, in forensic identification of individuals, in determining relatedness of individuals, and in evolutionary studies of genetically close populations or species.

DNA fingerprint is a pattern of a variable number of bands (DNA fragments) with different sizes on a Southern gel for each individual, generated by one or many VNTR loci (Jeffreys *et al.* 1985*a*, *b*, *c*; Wong *et al.* 1986, 1987). Such patterns, detected by hybridization with a single multilocus probe (MLP), or combination of patterns from hybridization with several single locus probes (SLPs), provide opportunities for microevolutionary studies. Population relationships are studied by using population frequency distributions generated by SLPs which have characteristics parallel to the traditional isozyme and RFLP loci (Chakraborty *et al.* 1992). A single MLP, which detects many hypervariable VNTR loci simultaneously in the genome, has features that are somewhat different from the SLPs. First, the number of loci underlying the MLP is unknown, and the allele frequencies at each locus are not available. This

* Corresponding author.

makes genetic distance computations troublesome, since all measures of genetic distances rely on the allele frequency data for each locus in all populations (Chakraborty & Rao, 1991). Second, the fingerprint data of the MLPs are often marred with technical limitations such as the incomplete resolution of bands of nearly similar sizes and co-migration of fragments produced by alleles at different loci. These technical problems introduce complications in the statistical interpretation of DNA fingerprint data from a single MLP. Although these issues have been mentioned by others (see, e.g., Lynch, 1988) to argue that standard population genetic principles may not apply to DNA fingerprinting data, there had been some attempts that demonstrate that, properly interpreted, DNA fingerprint data provide information regarding several critical population genetic parameters that are useful for evolutionary studies (Jeffreys *et al.* 1988; Stephens *et al.* 1992; Jin & Chakraborty, 1993). In addition, the MLPs have the merit of being efficient and cost effective in the sense that through their use it is easy to type many individuals at many loci in a short period of time, and with relatively low cost. Therefore, a formal statistical method of calibrating evolutionary changes from MLP data by taking into account their limitations is a worthwhile exercise.

Several efforts have been made to utilize MLP data since the first demonstration of the utility of MLPs in genetic studies (Jeffreys *et al.* 1985*a*, *b*, *c*). Lynch (1990, 1991) proposed an empirical similarity measurement for population genetic studies using DNA fingerprint data from MLPs. Although a very similar approach was followed by Yuhki & O'Brien (1990) and Gilbert *et al.* (1990, 1991), the lack of theoretical support makes it less appealing.

The purpose of this research is to provide a statistical basis for measurement of genetic distances between populations using the DNA fingerprint data, based on a population genetic model. First, we consider a model where the allele frequencies at each underlying locus are known, through which we study the population dynamics of summary measures such as number of shared bands between individuals within a population, as well as that of the number of shared bands between individuals of different populations. Drift expectations of such summary measures are studied under a specific mutation model (the infinite allele model; Wright, 1949; Kimura & Crow, 1964). Second, we show that the drift expectation of the number of shared alleles (bands) between individuals from two populations is a function of the composite parameter ($M = 4Nv$) and $t/2N$, where N is the effective population size, v, the mutation rate, and t is the divergence time of the two populations. These demonstrate that the genetic distance based on the number of shared alleles (bands) between individuals is approximately equivalent to Nei's genetic distance (Nei, 1972, 1987), even though the present approach does not explicitly utilize the entire allele frequency data from both populations. Numerical illustrations from the data on several SLP loci studied in three human racial groups are provided to indicate that the loss of information from this approach is not substantial in comparison with the evaluation of Nei's genetic distances from allele frequency data. Finally, we show that even though, for the analysis of SLP data, such data summarization based on the allele frequency distribution is not required, this approach allows an application of this theory to use fingerprint data from MLPs for population genetic studies. The basic assumption required for such application is to view the MLP as a collection of several SLPs where the co-migration of alleles at different loci are neglected as in other studies (Lynch, 1988, 1990, 1991; Li *et al.* 1993). Finally, we indicate the possible effects of co-migration and incomplete resolution of similar size fragments on the evaluation of genetic distances based on the number of shared bands (alleles).

2. Theory

(i) *The mutation model*

VNTR loci can be classified into three groups based on the size of the repeat unit: microsatellites (1–2 base pair (bp) repeat unit), short tandem repeat (STR, 3–5 bp repeat unit), and minisatellite (15–70 bp repeat unit) (Shriver *et al.* 1993). Although the exact molecular mechanisms of copy number alterations of core units at VNTR loci are still unknown, several mechanisms (e.g. replication slippage, unequal sister chromatid exchange) have been suggested based on the experimental as well as population genetic evidence. The minisatellite loci are presumably the major determinants of DNA fingerprint patterns revealed by a MLP. Allelic variations at such loci have been explained by unequal exchanges between long tandem repeat arrays. This results in a very large number of different sized alleles, as the infinite-allele model (IAM) assumes (Clark, 1989; Flint *et al.* 1989; Chakraborty *et al.* 1991; Shriver *et al.* 1993). In this research, we use the infinite-allele model as the underlying mutation mechanism of minisatellite VNTR loci to investigate the population dynamics of our summary statistics.

(ii) *The distribution of the number of shared alleles (bands) between individuals within and between populations*

Chakraborty & Jin (1993) showed that the number of shared alleles (bands) can be used as a summary measure to describe kinship relationships between individuals for DNA fingerprinting data generated either by a combination of several SLPs or by a MLP. They also showed that the number of shared alleles

(bands) between two individuals drawn randomly from a population (n_w) at l-th locus has the following distribution,

$$n_w = \begin{cases} 0 & 1-4a_2+4a_3-3a_4+2a_2^2, \\ 1 & 4a_2-4a_3+5a_4-4a_2^2, \\ 2 & -2a_4+2a_2^2, \end{cases} \quad (1)$$

where a_m is the sum of the m-th power of allele frequencies of l-th locus being considered, i.e.

$$a_m = \sum_{i=1}^{k} x_i^m(l), \quad (2)$$

where $x_i(l)$ is the frequency of i-th allele at l-th locus and k is the number of alleles at l-th locus. The drift expectation of n_w can thus be written as

$$E[n_w] = 4E(a_2)-4E(a_3)+E(a_4). \quad (3)$$

By following the same approach, the distribution of the number of shared alleles (bands) two individuals one of which is drawn from population 1 while the other is drawn from population 2, (n_b), at l-th locus is given by

$$n_b = \begin{cases} 0 & 1-4b_{11}+2b_{12}+2b_{21}-3b_{22}+2b_{11}^2, \\ 1 & 4b_{11}-2b_{12}-2b_{21}+5b_{22}-4b_{11}^2, \\ 2 & -2b_{22}+2b_{11}^2, \end{cases} \quad (4)$$

in which

$$b_{mn} = \sum_{i=1}^{k} x_i^m(l)\, y_i^n(l), \quad (5)$$

where $x_i(l)$, $y_i(l)$ are the allele frequencies of i-th allele ($i = 1, 2, \ldots, k$) at l-th locus from population 1 and population 2, respectively. The drift expectation of the number of shared alleles (bands) between population 1 and population 2, n_b, can be written as

$$E[n_b] = 4E(b_{11})-2E(b_{12})-2E(b_{21})+E(b_{22}). \quad (6)$$

(iii) *The dynamics of the number of shared alleles (bands) between individuals within populations*

Following Li & Nei (1975), we start with a k-allele model so that the results under the infinite-allele model are obtained by letting $k \to \infty$. Consider a randomly mating diploid population of effective size N. Assume that N is sufficiently large so that $1/N^2$ and higher powers of $1/N$ are negligible compared with $1/N$. Following Kimura (1968), we assume that there are k possible allelic states at a locus and each allele mutates at the rate of ν per generation to any one of the $k-1$ other allelic types with equal probability. All mutations are assumed neutral in this analysis.

Following Li & Nei (1975), let

$$\mu^{(t)}_{mn,\,pq} = E[x_i^m(t)\, x_j^n(t)\, y_i^p(t)\, y_j^q(t)]$$

be the (m, n, p, q)-th moments of $x_i(t)$, $x_j(t)$, $y_i(t)$, and $y_j(t)$. The allele frequencies $x_i(t)$, $x_j(t)$, $y_i(t)$, and $y_j(t)$ satisfy the following recurrence equations:

$$\begin{aligned} x_i(t+1) &= X_i(t)+\delta X_i(t), \\ x_j(t+1) &= X_j(t)+\delta X_j(t), \\ y_i(t+1) &= Y_i(t)+\delta Y_i(t), \\ y_j(t+1) &= Y_j(t)+\delta Y_j(t), \end{aligned}$$

where $\delta X_i(t)$ and $\delta Y_i(t)$ are the deviations of allele frequencies $X_i(t)$ and $Y_i(t)$ at t-th generation due to random drift, and

$$\begin{aligned} X_i(t) &= (1-c)\, x_i(t)+d, \\ X_j(t) &= (1-c)\, x_j(t)+d, \\ Y_i(t) &= (1-c)\, y_i(t)+d, \\ Y_j(t) &= (1-c)\, y_j(t)+d, \\ d &= \nu/(k-1), \\ c &= k\nu/(k-1), \end{aligned} \quad (7)$$

and

$$\begin{aligned} E[\delta X_i(t)] = E[\delta X_j(t)] = E[\delta Y_i(t)] = E[\delta Y_j(t)] &= 0, \\ E[\{\delta X_i(t)\}^2] &= X_i(t)\,[1-X_i(t)]/2N, \\ E[\{\delta Y_i(t)\}^2] &= Y_i(t)\,[1-Y_i(t)]/2N, \\ E[\delta X_i(t)\,\delta X_j(t)] &= -X_i(t)\,X_j(t)/2N, \\ E[\delta Y_i(t)\,\delta Y_j(t)] &= -Y_i(t)\,Y_j(t)/2N, \\ E[\delta X_i(t)\,\delta Y_i(t)] = E[\delta X_i(t)\,\delta Y_j(t)] &= 0. \end{aligned}$$

By approximating $\mu^{(t+1)}_{mn,\,pq}-\mu^{(t)}_{mn,\,pq}$ by $d\mu_{mn,\,pq}(t)/dt$, we have the following differential equation

$$\begin{aligned} \frac{d\mu^{(t)}_{mn,\,pq}}{dt} = \{&[(m+n)(A+m+n-1) \\ &+(p+q)(A+p+q-1)]\mu^{(t)}_{mn,\,pq} \\ &+m(B+m-1)\,\mu^{(t)}_{(m-1)\,n,\,pq} \\ &+n(B+n-1)\,\mu^{(t)}_{m(n-1),\,pq} \\ &+p(B+p-1)\,\mu^{(t)}_{mn,\,(p-1)\,q} \\ &+q(B+q-1)\,\mu^{(t)}_{mn,\,p(q-1)}\}/4N, \end{aligned} \quad (8)$$

where $A = 4Nc$, $B = 4Nd$, and the terms involving $1/N^2$ and higher-order terms are neglected. The solution of the equation (8) can be obtained step by step, starting from $\mu^{(t)}_{10,\,00}$, $\mu^{(t)}_{01,\,00}$, $\mu^{(t)}_{00,\,10}$, and $\mu^{(t)}_{00,\,01}$. Note that

$$\mu^{(t)}_{10,\,00} = \frac{1}{k}+\left[x_i(0)-\frac{1}{k}\right]e^{-ct}, \quad (9)$$

and $\mu^{(t)}_{01,\,00}$, $\mu^{(t)}_{00,\,10}$, $\mu^{(t)}_{00,\,01}$ can be obtained by replacing $x_i(0)$ with $x_j(0)$, $y_i(0)$, and $y_j(0)$, respectively.

Using the solutions of equation (8), $E[a_2(t)]$, $E[a_3(t)]$, $E[a_4(t)]$ can be expressed in terms of parameters t, ν, N, and M (see Appendix equations (A 1)–(A 3)), so

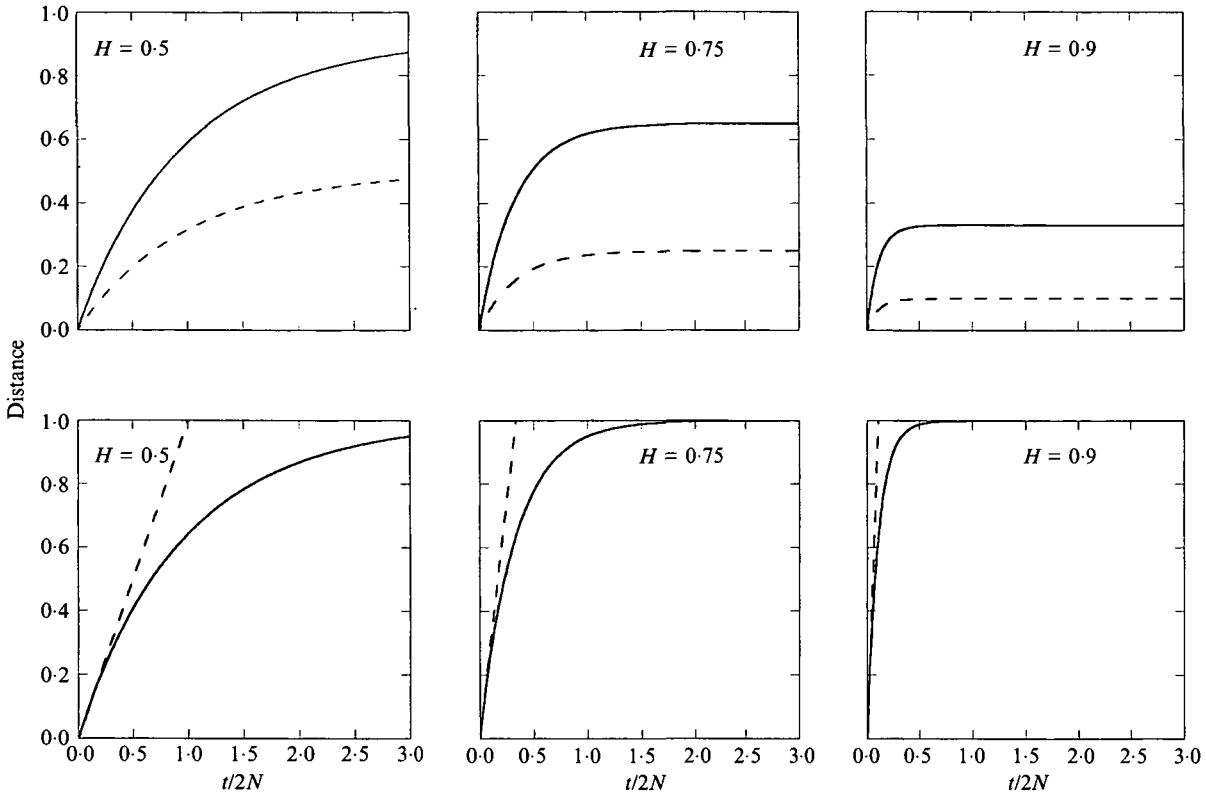

Fig. 1. Relationship between expectations of measures of genetic distance between populations and time of divergence ($t/2N$), in units of $2N$ generations for various values of average heterozygosity (H). In the top three panels, the solid lines represent the expected genetic distance based on DNA fingerprinting band (allele) sharing data, $E[D(t)]$ (equation (15), and the dashed lines are expectations of Nei's minimum distance (equation (16)), while in the bottom three panels $E[D_t(t)]$ (equation (18)) are shown by solid lines, and Nei's standard distance by dashed lines.

that the expectation of the number of shared alleles (bands) between individuals within population ($n_w(t)$) at t-th generation after divergence is given by

$$E[n_w(t)] = \left[a_4(0) - \frac{12a_3(0)}{M+6} + \frac{36a_2(0)}{(M+5)(M+6)} - \frac{18}{(M+3)(M+5)(M+6)}\right] e^{-(4\nu+3/N)t} - \frac{4(M+3)}{M+6} \times \left[a_3(0) - \frac{6a_2(0)}{M+4} - \frac{4}{(M+2)(M+4)}\right] e^{-(3\nu+3/2N)t} + \frac{4(M^2+3M-1)}{(M+4)(M+5)}\left[a_2(0) - \frac{1}{M+1}\right] e^{-(2\nu+1/2N)t} + \frac{2(2M^2+6M+3)}{(M+1)(M+2)(M+3)}. \quad (10)$$

If the initial population (i.e., $t = 0$) is at equilibrium, we have

$$a_2(0) = 1/(M+1),$$
$$a_3(0) = 2/(M+1)(M+2),$$
$$a_4(0) = 6/(M+1)(M+2)(M+3),$$

as obtained by Li and Nei (1975). Even otherwise, at $t \to \infty$, equation (10) obtains the limit

$$E[n_w(\infty)] = \frac{4M^2+12M+6}{(M+1)(M+2)(M+3)}. \quad (11)$$

(iv) *The dynamics of the number of shared alleles (bands) between individuals between populations*

With the same approach, the expectation of the number of shared alleles (bands) between individuals one of which comes from population 1 and the other from population 2 can be obtained by replacing the expectation of b_{mn} of equation (6) by the solutions of equation (8) (see Appendix equations (A 4)–(A 7)). This becomes

$$E[n_b(t)] = \left[b_{22}(0) - \frac{2\{b_{12}(0)+b_{21}(0)\}}{M+2} + \frac{4b_{11}(0)}{(M+2)^2}\right] \times e^{-(4\nu+1/N)t} - \frac{2(M+1)}{M+2}\left[b_{12}(0)+b_{21}(0) - \frac{6b_{11}(0)}{M+2}\right] \times e^{-(3\nu+1/2N)t} + \frac{4(M+1)^2}{(M+2)^2} b_{11}(0)\, e^{-2\nu t}. \quad (12)$$

When the ancestral population before the divergence is at equilibrium (i.e. $t = 0$), we have $b_{11}(0) = a_2(0), b_{12}(0) = b_{21}(0) = a_3(0), b_{22}(0) = a_4(0)$. Therefore,

$$E[n_b(t)] = \frac{2M}{(M+1)(M+2)^2(M+3)} e^{-(4\nu+1/N)t} + \frac{4(M+1)}{(M+2)^2} e^{-2\nu t}. \quad (13)$$

(v) *Genetic distance*

Equation (13) indicates that as t tends to ∞, the drift expectation of $n_b(t)$ approaches zero. In contrast, at $t = 0$, $E[n_b(t)] = E[n_w(\infty)]$ when the initial population before splitting is at mutation-drift equilibrium. The genetic distance based on the number of shared alleles (bands) between individuals, therefore, can be defined as

$$D(t) = \frac{n_{w1} + n_{w2}}{2} - n_b(t), \quad (14)$$

where n_{w1} and n_{w2} are the number of shared alleles (bands) within population 1 and population 2, respectively, and each of them can be estimated as the average number of shared bands of all pairwise comparisons of DNA fingerprints of the individuals in that population. Since under the assumption that the initial population is at mutation-drift equilibrium before split, $E[(n_{w1} + n_{w2})/2] = E[n_w(\infty)]$ is a constant, the expectation of $D(t)$ can be written as

$$E[D(t)] = E[n_w(\infty)] - E[n_b(t)] = \frac{2M}{(M+1)(M+2)^2(M+3)}[1 - e^{-(M+1)T}] + \frac{4(M+1)}{(M+2)^2}[1 - e^{-MT}], \quad (15)$$

where $M = 4N\nu$, and $T = t/2N$.

This makes the properties of the distance measure based on the number of shared bands (alleles) similar to that of Nei's minimum genetic distance, since Nei's minimum genetic distance has the drift expectation

$$E[D_m(t)] = J_X(\infty)[1 - e^{-MT}], \quad (16)$$

where $J_X(\infty)$ is the probability of gene identity within an equilibrium population (Nei, 1987). Because of the feature that equations (15) and (16) both reach an asymptote depending upon the within-population genetic diversity, the asymptote being $E[n_w(\infty)]$ for $E[D(t)]$, and $J_X(\infty)$ for $D_m(t)$, we may also define an index of genetic dissimilarity based on the number of shared bands (alleles) by

$$D_i(t) = 1 - \frac{2n_b(t)}{n_{w1} + n_{w2}}, \quad (17)$$

whose drift expectations is approximately given by

$$E[D_i(t)] \approx 1 - \frac{E[n_b(t)]}{E[n_w(\infty)]} = \frac{M}{(2M^2+6M+3)(M+2)}[1 - e^{-(M+1)T}] + \frac{2(M+1)^2(M+3)}{(M+2)(2M^2+6M+3)}[1 - e^{-MT}], \quad (18)$$

by using equations (11) and (13).

Figure 1 shows some numerical computations on the expected distance between populations as functions of their time of divergence, for different levels of heterozygosity values within populations. The solid lines in these diagrams are the relationships for the expectations of $D(t)$ and $D_i(t)$ given by equations (15) and (18), while the dashed lines are for the expectation of Nei's minimum distance by equation (16) (compared with $D(t)$) and Nei's standard distance (compared with $D_i(t)$).

These computations indicate several important features of the proposed distance functions (equations (14) and (17)). First, both measures of genetic distances are not completely proportional to the time of divergence. However, the deviation from linearity with the time of divergence starts approximately at a point of time when Nei's distance statistics also fall off from the linear time-dependence. Second, the proportionality with time of divergence holds for a time period that depends on the degree of heterozygosity (H). When H is larger, the linearity holds for a shorter time of divergence. Third, for $H \leqslant 90\%$ (as in the case of many STR and VNTR loci), $D(t)$ and $D_i(t)$ appears to hold the linear relationship quite adequately for $t \leqslant N$ generations.

When M is large (say, $M \geqslant 2$), the second term of equation (18) is the dominant component of drift expectation of $D_i(t)$. In the context of hypervariable loci, since most SLPs and polymorphic loci in MLPs show levels of heterozygosity (H) 70% or above (Nakamura *et al.* 1987; Wong *et al.* 1987; Armour *et al.* 1990; Edwards *et al.* 1992), and hence $M = H/(1-H) \geqslant 2$, we may approximate the drift expectation of $D_i(t)$ by

$$E[D_i(t)] \approx \frac{2(M+1)^2(M+3)}{(M+2)(2M^2+6M+3)}[1 - e^{-MT}], \quad (19)$$

for the study of population differentiation with fingerprint data. Figure 2 shows the effect of such approximations for different levels of heterozygosity, while the dotted lines are using the approximation of equation (19). These computations indicate that while equation (19) underestimates the expected genetic distance, even for H at the level of 50%, the approximation is fairly accurate for times of divergence of the order of $3N$.

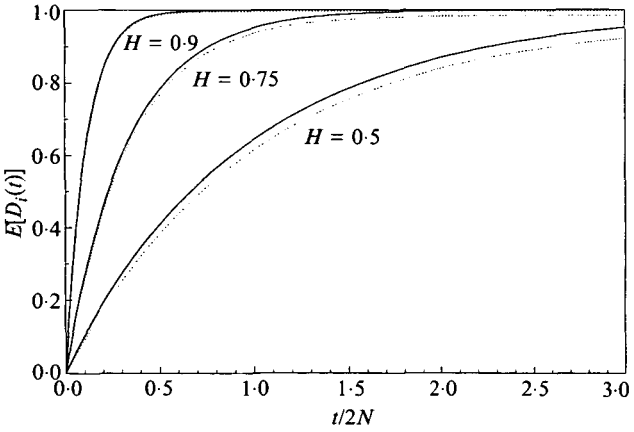

Fig. 2. Effect of approximation (equation (19)) on the expected genetic distance, $E[D_t(t)]$, based on band (allele) sharing data from DNA fingerprinting patterns for different levels of average heterozygosity (H) and times of divergence ($t/2N$) measured in units of $2N$ generations. The solid lines are exact expectations (equation (18)) while the dotted lines represent their respective approximate values (equation (19)).

(vi) *Genetic distance from DNA fingerprint data*

The DNA fingerprint generated by a single MLP or a group of SLPs is a combination of patterns of many VNTR loci. Viewing MLP as a collection of several SLPs where the co-migration of alleles at different loci are neglected, our previous definition of genetic distance based on the number of shared bands within and between population (equation (14)) holds for multiple loci. This is so because the number of shared bands both for within and between populations are additive across loci, which makes our genetic distance applicable to DNA fingerprint data.

On the other hand, $D_t(t)$ is not additive across loci. This may compromise the applicability of $D_t(t)$ for DNA fingerprint data. However, it can be shown that as long as the divergence time (t) is relatively small, the linearity between $D_t(t)$ and t still holds.

Table 1. *Comparison of D_t with Nei's distances by using STR data*

	D_t	D_s	D_m
5-loci†			
C–B	0·15265	0·15482	0·15272
C–A	0·11603	0·10989	0·10479
B–A	0·13815	0·14224	0·14779
7-loci†			
C–B	0·15027	0·14957	0·14622
C–A	0·09765	0·11355	0·10844
B–A	0·14299	0·14139	0·14579

D_s: Nei's standard distance.
D_m: Nei's minimum distance/average homozygosity.
C, Caucasians; B, Blacks; A, Asians.
† See text for a listing of the loci.

3. Numerical results

Recently, Edwards *et al.* (1991) described several short tandem repeat (STR) loci, each of which demonstrates considerable degrees of polymorphism within populations. The population genetic characteristics of five of these loci were previously described (Edwards *et al.* 1992) in Caucasians (200 individuals), American Blacks (200 individuals), and Asians (80 individuals) currently residing in Houston, Texas. Two more STR loci have now been typed recently for the same individuals from the populations mentioned above.

Using 7-locus (TH01, RENA4 FARB, HPRTB, ARA, CD4, and PLA2A1) genotype data we computed the pairwise numbers of shared bands within and between populations and then the numerical values of the genetic distance based on the measure of similarity index (D_t) (see Table 1). For comparison, the allele frequencies from each locus are also used to compute Nei's minimum and standard genetic distances, using the estimation procedure suggested by Nei (1978). Since two of the seven loci are X-linked (HPRTB, and ARA), we first computed D_t for the five autosomal loci from all individuals. We also computed the distances for all seven loci by using female individuals only. Note that for the estimation of Nei's minimum distance, a standardization was carried out by dividing Nei's minimum distance with the average homozygosity of two populations compared in order to make Nei's minimum distance range from 0 to 1 so that it would be comparable with other distance measurements.

The computations in Table 1 show that even though the measures of genetic distance based on allele sharing data consider only a summary measure of genotype data (number of bands shared between individuals), such data summarization does not compromise the evaluation of evolutionary distances between populations, since the computed distance values are virtually identical to the ones obtained by Nei's method of estimation of genetic distances.

4. Discussion and conclusions

The statistic $D_t(t)$ has a similarity in appearance with indices that have been proposed to study restriction fragment length polymorphism (RFLP) data (Nei & Li, 1979; Lynch, 1990, 1991), but we might note that these concepts are somewhat different, since our proposed statistic (equation (14)) is based on pairwise comparison of individuals, at the level of within as well as between populations, so that genetic dissimilarity between random samples of individuals within and between populations are being contrasted here, in the spirit of the formulation of Nei's distance indices (Nei, 1972). On the contrary, Nei & Li (1979), and Lynch (1990, 1991) were attempting to normalize band (allele) sharing between individuals in terms of

numbers of bands (alleles) present in the individuals under comparison. Therefore, conceptually, our proposed statistic is different from the ones suggested before in the context of analysing DNA fingerprint data.

Throughout our derivations, we equated bands to alleles, while in principle for DNA fingerprint data, co-migration of fragment sizes resulting from alleles at different loci remains a viable possibility. Furthermore, the underlying minisatellite loci reflecting the DNA fingerprinting pattern may be linked, so that Li & Nei's (1975) theory may not be strictly applicable. While a rigorous study of these possibilities is impossible without hard data on the extent of co-migration and linkage relationships between the underlying loci, applicability of our theory may be intuitively justified. First, by a position by position analysis of fragment sizes, Krawczak & Bockel (1992), and Bockel *et al.* (1992) showed that the problem of co-migration may be examined by postulating 'position-specific genetic factors' (F), an unobservable variable. When F takes values larger than zero, an individual's DNA fingerprint pattern would exhibit the presence of a band in that position. The relative frequencies of presence of bands at specific positions (x, in the terminology of Bockel *et al.* 1992), in turn, predicts how large the values of F can be in any DNA fingerprint database in population surveys. In general, values of x do not exceed 0·15 to 0·25, so that the probability of $F \geqslant 2$ is less than 0·035, under the assumption that F is distributed as a Poisson variate. This suggests that co-migration of alleles at different loci is not a very common phenomenon. Furthermore, there is no evidence that alleles of specific frequencies would be more likely to co-migrate than the ones that form distinctly different bands. Therefore, as long as the co-migrating alleles represent a random sample of all alleles (a reasonable assumption to work with), our theory of expected genetic distance should apply without any systematic bias in evolutionary predictions. Of course, the variance of genetic distance (particularly, the stochastic component, or the intra-locus component) would be under-predicted by neglecting the effect of co-migration. Since we have not derived the variance of distance measures in the present work, we conclude that the effect of co-migration is not critical for the conclusions reached in this work.

Similarly, the problem of linkage of underlying loci cannot be rigorously examined without a full dissection of all loci underlying a MLP used. Such data are lacking. Nevertheless, attempts to locate dispersal of hypervariable minisatellite loci in the genome indicate that they are located on chromosomal bands that are identifiable from *in situ* hybridization of metaphase chromosomes (Royle *et al.* 1988; Zischler *et al.* 1989). Therefore, in terms of physical distances such loci are generally located far apart from each other, so that the effect of linkage disequilibrium between bands (alleles) can be neglected. This is so, because they are also separated by large recombination distances, so that for all practical purposes they may be assumed independent. We might add that even when the allele frequencies are dependent, the expected genetic distance should not be affected, since expectation of sums of powers of allele frequencies is not altered even when the alleles at different loci are dependent.

In summary, this work shows that DNA fingerprint data allow a calibration of genetic divergence between populations or taxa that are evolutionarily close enough, because even divergence between them would be reflected in their fingerprint profiles caused by the high rate of mutation. Linearity with time of divergence holds for taxa that are separated from each other up to N generations, as long as the average heterozygosity approximates 90%, as is the case of most minisatellite loci. Evolutionary studies involving such hypervariable loci are important for other reasons as well. It is now well-established that genomes of several organisms have such interspersed hypervariable regions which involve genetic alterations due to copy number variation of tandemly repeated short sequences. For example, the minisatellite core sequences (such as 33·6 and 33·15) appear to have been conserved over evolution in plants, mammals, apes, and human (see citations in Kelly *et al.* 1989). Most contemporary methods (e.g. RFLP markers, RAPD markers, or DNA sequence analysis) of studying genetic relationships between taxa rely on genetic alterations caused by nucleotide substitutions. In comparison, polymorphisms in DNA fingerprinting are caused by mechanisms different from them so that the value of DNA fingerprint patterns for comparative taxonomic analysis cannot be denigrated. The theory presented here demonstrates that neither the effects of co-migration, linkage and incomplete resolution, nor the unknown number of loci compromise such inference substantially. Of course, further studies are needed to examine the sampling properties of such summary measures of genetic divergence, from which the limiting features of DNA fingerprint protocols may be empirically established.

This work was supported by US Public Health Service Research Grants GM-41399 and GM-45861 from the National Institutes of Health, and grant 92-IJ-CX-K024 from the National Institute of Justice. The opinions, of course, are those of authors and do not constitute an endorsement of the granting agencies.

References

Armour, J. A., Povey, S., Jeremiah, S. & Jeffreys, A. J. (1990). Systematic cloning of human minisatellites from ordered array charomid libraries. *Genomics* **8**, 501–512.

Bockel, B., Nürnberg, P. & Krawczak, M. (1992). Likelihoods of multilocus DNA fingerprints in extended families. *American Journal of Human Genetics* **51**, 554–561.

Chakraborty, R. & Rao, C. R. (1991). Measurement of

genetic variation for evolutionary studies. In *Handbook of Statistics*, Vol. 8 (ed. C. R. Rao and R. Chakraborty). Amsterdam/London/New York/Tokyo: North-Holland.

Chakraborty, R., Fornage, M., Gueguen, R. & Boerwinkle, E. (1991). Population genetics of hypervariable loci: analysis of PCR based VNTR polymorphism within a population. In *DNA Fingerprinting: Approaches and Applications* (ed. T. Burke, G. Dolf, A. J. Jeffreys and R. Wolff). Basel/Boston/Berlin: Birkhäuser Verlag.

Chakraborty, R., Deka, R., Jin, L. & Ferrell, R. E. (1992). Allele sharing at six VNTR loci and genetic distances among three ethnically defined human populations. *American Journal of Human Biology* **4**, 387–397.

Chakraborty, R. & Jin, L. (1993). Determination of relatedness between individuals by DNA fingerprinting. *Human Biology* **65**, 875–895.

Clark, A. G. (1987). Neutrality tests of highly polymorphic restriction fragment length polymorphisms. *American Journal of Human Genetics* **41**, 948–956.

Edwards, A., Civitello, A., Hammond, H. A. & Caskey, C. T. (1991). DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *American Journal of Human Genetics* **49**, 746–756.

Edwards, A., Hammond, H. A., Jin, L., Caskey, C. T. & Chakraborty, R. (1992). Genetic variation at five trimeric and tetrameric random repeat loci in four human population groups. *Genomics* **12**, 241–253.

Flint, J., Boyce, A. J., Martinson, J. J. & Clegg, J. B. (1989). Population bottlenecks in Polynesia revealed by minisatellite. *Human Genetics* **83**, 257–263.

Gilbert, D. A., Lehman, N., O'Brien, S. J. & Wayne, R. K. (1990). Genetic fingerprinting reflects population differentiation in the California Channel Island fox. *Nature* **344**, 764–767.

Gilbert, D. A., Packer, C., Pusey, A. E., Stephens, J. C. & O'Brien, S. J. (1991). Analytical DNA fingerprinting in lions: Parentage, genetic diversity, and kinship. *Journal of Heredity* **82**, 378–386.

Jeffreys, A. J., Wilson, V. & Thein, S. L. (1985*a*). Hypervariable 'minisatellite' regions in human DNA. *Nature* **314**, 67–73.

Jeffreys, A. J., Wilson, V. & Thein, S. L. (1985*b*). Individual-specific 'fingerprints' of human DNA. *Nature* **316**, 76–79.

Jeffreys, A. J., Brookfield, J. F. Y. & Semeonoff, R. (1985*c*). Positive identification of an immigration test-case using human DNA fingerprints. *Nature* **317**, 818–819.

Jeffreys, A. J., Royle, N. J., Wilson, V. & Wong, Z. (1988). Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* **332**, 278–281.

Jin, L. & Chakraborty, R. (1993). A bias-corrected estimate of heterozygosity for single-probe multilocus DNA fingerprints. *Molecular Biology and Evolution* **10**, 1112–1114.

Kelly, R., Bulfield, G., Collick, A., Gibbs, M. & Jeffreys, A. J, (1989). Characterization of a highly unstable mouse minisatellite locus: Evidence for somatic mutation during early development. *Genomics* **5**, 844–856.

Kimura, M. (1968). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetical Research* **11**, 247–269.

Kimura, M. & Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738.

Krawczak, M. & Bockel, B. (1992). A genetic factor model for the statistical analysis of multilocus DNA fingerprints. *Electrophoresis* **13**, 10–17.

Li, C. C., Weeks, D. E. & Chakraborty, A. (1993). Similarity of DNA fingerprints due to chance and relatedness. *Human Heredity* **43**, 45–52.

Li, W.-H. & Nei, M. (1975). Drift variances of heterozygosity and genetic distance in transient states. *Genetical Research* **25**, 229–248.

Lynch, M. (1988). Estimation of relatedness by DNA fingerprinting. *Molecular Biology and Evolution* **5**, 584–599.

Lynch, M. (1990). The similarity index and DNA fingerprinting. *Molecular Biology and Evolution* **7**, 478–484.

Lynch, M. (1991). Analysis of population genetic structure by DNA fingerprinting. In *DNA Fingerprinting: Approaches and Applications* (ed. T. Burke, G. Dolf, A. J. Jeffreys and R. Wolff). Basel/Boston/Berlin: Birkhäuser Verlag.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E. & White, R. (1987). Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235**, 1616–1622.

Nei, M. (1972). Genetic distance between populations. *American Naturalist* **106**, 283–292.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**, 583–590.

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Nei, M. & Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of National Academy of Sciences USA* **76**, 5269–5273.

Royle, N. J., Clarkson, R. E., Wong, Z. & Jeffreys, A. J. (1988). Clustering of hypervariable minisatellites in the proterminal regions of human autosomes. *Genomics* **3**, 352–360.

Saitou, N. & Nei, M. (1986). The number of nucleotides required to determine the branching order of three species with special reference to the human–chimpanzee–gorilla divergence. *Journal of Molecular Evolution* **24**, 189–204.

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

Stephens, J. C., Gilbert, D. A., Yuhki, N. & O'Brien, S. J. (1992). Estimation of heterozygosity for single-probe multilocus DNA fingerprints. *Molecular Biology and Evolution* **9**, 729–743.

Wong, Z., Wilson, V., Jeffreys, A. J. & Thein, S. L. (1986). Cloning a selected fragment from a human DNA 'fingerprint': Isolation of an extremely polymorphic minisatellite. *Nucleic Acids Research* **14**, 4605–4616.

Wong, Z., Wilson, V., Patel, I., Povey, S. & Jeffreys, A. J. (1987). Characterization of a panel of highly variable minisatellites cloned from human DNA. *Annual of Human Genetics* **51**, 269–288.

Wright, S. (1949). Genetics of populations. *Encyclopedia Britannica*, 14th ed. **10**, 111–112.

Yuhki, N. & O'Brien, S. J. (1990). DNA variation of the mammalian major histocompatibility complex reflects genomic diversity and population history. *Proceedings of National Academy of Sciences USA* **87**, 836–840.

Zischler, H., Nanda, I., Schäfer, R., Schmid, M. & Epplen, J. T. (1989). Digoxigenated oligonucleotide probes specific for simple repeats in DNA fingerprinting and hybridization in situ. *Human Genetics* **82**, 227–233.

Appendix

The expectations of $a_m(t)$ and $b_{mn}(t)$ in equations (3) and (6) can be written in the form of the solutions of equation (8) under the infinite-allele model. Here $\sum$ stands for the summation over all alleles.

$$\begin{aligned} E[a_2(t)] &= E[\textstyle\sum x_i^2(t)] \\ &= \textstyle\sum \mu_{20,00}^{(t)} \\ &= \left[a_2(0) - \frac{1}{M+1}\right] e^{-(2\nu+1/2N)t} + \frac{1}{M+1}. \end{aligned} \tag{A 1}$$

$$\begin{aligned} E[a_3(t)] &= E[\textstyle\sum x_i^3(t)] \\ &= \textstyle\sum \mu_{30,00}^{(t)} \\ &= \left[\frac{4}{(M+2)(M+4)} - \frac{6a_2(0)}{M+4} + a_3(0)\right] e^{-(3\nu+3/2N)t} \\ &\quad + \frac{6}{M+4}\left[a_2(0) - \frac{1}{M+1}\right] e^{-(2\nu+1/2N)t} \\ &\quad + \frac{2}{(M+1)(M+2)}. \end{aligned} \tag{A 2}$$

$$\begin{aligned} E[a_4(t)] &= E[\textstyle\sum x_i^4(t)] \\ &= \textstyle\sum \mu_{40,00}^{(t)} \\ &= \left[a_4(0) - \frac{12a_3(0)}{M+6} + \frac{36a_2(0)}{(M+5)(M+6)} \right. \\ &\quad \left. - \frac{18}{(M+3)(M+5)(M+6)}\right] e^{-(4\nu+3/N)t} \\ &\quad + \frac{12}{M+6}\left[\frac{4}{(M+2)(M+4)} - \frac{6a_2(0)}{M+4} + a_3(0)\right] \\ &\qquad \times e^{-(3\nu+3/2n)t} \\ &\quad + \frac{36}{(M+4)(M+5)}\left[a_2(0) - \frac{1}{M+1}\right] \\ &\qquad \times e^{-(2\nu+1/2N)t} \\ &\quad + \frac{2}{(M+1)(M+2)(M+3)}. \end{aligned} \tag{A 3}$$

$$\begin{aligned} E[b_{11}(t)] &= E[\textstyle\sum x_i(t)\, y_i(t)] \\ &= \textstyle\sum \mu_{10,10}^{(t)} \\ &= b_{11}(0)\, e^{-2\nu t} \end{aligned} \tag{A 4}$$

$$\begin{aligned} E[b_{12}(t)] &= E[\textstyle\sum x_i(t)\, y_i^2(t)] \\ &= \textstyle\sum \mu_{10,20}^{(t)} \\ &= \left[b_{12}(0) - \frac{2b_{11}(0)}{M+1}\right] e^{-(3\nu+1/2N)t} + \frac{2b_{11}(0)}{M+2} e^{-2\nu t}. \end{aligned} \tag{A 5}$$

$$\begin{aligned} E[b_{21}(t)] &= E[\textstyle\sum x_i^2(t)\, y_i(t)] \\ &= \textstyle\sum \mu_{20,10}^{(t)} \\ &= \left[b_{21}(0) - \frac{2b_{11}(0)}{M+1}\right] e^{-(3\nu+1/2N)t} + \frac{2b_{11}(0)}{M+2} e^{-2\nu t}. \end{aligned} \tag{A 6}$$

$$\begin{aligned} E[b_{22}(t)] &= E[\textstyle\sum x_i^2(t)\, y_i^2(t)] \\ &= \textstyle\sum \mu_{20,20}^{(t)} \\ &= \left[b_{22}(0) - \frac{2b_{12}(0)}{M+2} - \frac{2b_{21}(0)}{M+2} + \frac{4b_{11}(0)}{(M+2)^2}\right] \\ &\qquad \times e^{-(4\nu+1/N)t} \\ &\quad + \left[\frac{2b_{12}(0)}{M+2} + \frac{2b_{21}(0)}{M+2} - \frac{8b_{11}(0)}{(M+2)^2}\right] e^{-(3\nu+1/2N)t} \\ &\quad + \frac{4b_{11}(0)}{(M+2)^2} e^{-2\nu t}. \end{aligned} \tag{A 7}$$