Heredity **74** (1995) 274–285
©The Genetical Society of Great Britain

Received 18 May 1994

Population structure, stepwise mutations, heterozygote deficiency and their implications in DNA forensics

LI JIN & RANAJIT CHAKRABORTY*

Center for Demographic and Population Genetics, Graduate School of Biomedical Sciences, University of Texas at Houston Health Science Center, P.O. Box 20334, Houston, TX 77225, U.S.A.

In a substructured population the overall heterozygote deficiency can be predicted from the number of subpopulations (s), their time of divergence (t), and the nature of the mutations. At present the true mutational mechanisms at the hypervariable DNA loci are not known. However, the two existing mutation models (the infinite allele model (IAM) and the stepwise mutation model (SMM)) provide some guides to predictions from which the possible effect of population substructuring may be evaluated, assuming that the subpopulations do not exchange any genes among them during evolution. The theory predicts that the loci with larger mutation rate, and consequently showing greater heterozygosity within subpopulations, should exhibit a smaller proportional heterozygote deficiency (G_{ST}) and, hence, the effects of population substructuring should be minimal at the hypervariable DNA loci (an order of magnitude smaller than that at the blood group and protein loci).

Applications of this theory to data on six Variable Number of Tandem Repeat (VNTR) loci and five short tandem repeat (STR) loci in the major cosmopolitan populations of the USA show that while the VNTR loci often exhibit a large significant heterozygote deficiency, the STR loci do not show a similar tendency. This discordant finding may be ascribed to the limitations, coalescence and nondetectability of alleles associated with the restriction fragment length polymorphism (RFLP) analysis through which the VNTR loci are scored. Such limitations do not apply to the polymerase chain reaction (PCR) method, through which the STR loci are scored. The implications of these results are discussed in the context of the forensic use of DNA typing data.

Keywords: coefficient of gene diversity, DNA forensics, heterozygote deficiency, stepwise mutations, VNTR loci.

Introduction

The study of genetic structure of subdivided populations is at least 60 years old (Wright, 1931). Although the phenomenon of population substructuring is generally discussed in the context of evolutionary studies, its implications can be much broader (Chakraborty, 1993), as evidenced in the recent controversy in DNA forensics (Chakraborty & Kidd, 1991; Lewontin & Hartl, 1991). A population is said to be substructured when it consists of components (subpopulations) among which gene flow is somewhat restricted. There could be a complete absence of gene flow between subpopulations because of geographical, ecological and social, as well as biological barriers. Under complete isolation, gene differentiation between subpopulations proceeds with a speed governed by the mutation rate and effective sizes of subpopulations. Partial gene flow between subpopulations retards the process of genetic differentiation (Nei & Feldman, 1972; Chakraborty & Nei, 1974; Li, 1976a; Slatkin, 1985), so that eventually a steady-state will be reached with regard to inter-subpopulation genetic variation. The dynamics of this process depend on the nature of mutations and their selective differentials, as well as on the pattern of gene migration. These issues received a great deal of attention, theoretically (Morton, 1969; Karlin, 1982) as well as empirically (Jorde, 1980). The common theme is that under isolation of subpopulations (complete or partial) the allele frequency differences between subpopulations produce genotype frequencies in the total population that differ from what would be expected if

*Correspondence.

the individuals in the total population were to mix at random (i.e. no substructuring).

Such a deviation, quantified in terms of summary measures of genetic variation, has been termed the fixation index (Wright, 1943), kinship (Morton, 1969) or coefficient of gene differentiation (Nei, 1973). Unequivocal is the observation in all these studies that in the presence of substructuring heterozygosity in the total population at codominant loci would be reduced from what could be predicted from allele frequencies in the total population assuming Hardy–Weinberg genotypic proportions. Such a deficiency of heterozygosity, although a crude measure of the effects of population subdivision, when normalized (Nei, 1977), may allow examination of the strength of the effects of substructuring in terms of how many subpopulations exist within the total population, as well as their evolutionary time of separation from the common ancestry. Of course, the mutational pattern as well as the role of natural selection have to be postulated in such interference (Nei, 1975; Li, 1976a, Slatkin, 1985).

Because in the presence of substructuring the genotype frequencies in the total population cannot strictly be predicted from the overall allele frequencies using the Hardy–Weinberg rule, but the average magnitude of errors of such a prediction may be evaluated from the summary measures mentioned above, two types of enquiries may be made with regard to population structure. In one, we would have to identify each component subpopulation within the population, have representative samples drawn from each of them and evaluate the errors of prediction of genotype frequencies in the total population sample. In the other, a sample from the total population would be analysed from which the nature of substructuring has to be inferred from the errors of prediction observed.

Whereas the first approach gives a direct observation on the level of significance of allele frequency differences between subpopulations, it does not establish whether the subpopulations themselves are substructured. The problems of the second strategy are mainly statistical; namely, the power of detection of deviation from the Hardy–Weinberg predictions of genotype frequencies is notoriously weak (Ward & Sing, 1970) and, consequently, large sample sizes are required to detect realistic differences from the Hardy–Weinberg predictions (Ward & Sing, 1970; Chakraborty & Rao, 1972). Whereas these conclusions are valid for traditional serological and biochemical markers where the levels of polymorphisms are limited because of small numbers of alleles and consequent low levels of heterozygosity (Roychoudhury & Nei, 1988), the use of data on hypervariable loci (such as the Variable Number of Tandem Repeat loci, the VNTR loci) gives a somewhat different picture (Chakraborty & Zhong, 1993). Furthermore, empirical data on allele frequency distributions at hypervariable loci are not as extensive for diverse populations as yet and, therefore, at present inference regarding the extent of population structure at VNTR loci has to be inferred from the second type of studies mentioned above.

The purpose of this work is to provide a theoretical expectation of a measure of the effect of substructuring on genotype frequencies. Specifically, we extend our earlier work on this subject (Chakraborty & Jin, 1992) and derive expectations of the coefficient of gene diversity (G_{ST}; Nei, 1973) under the stepwise mutation model (Ohta & Kimura, 1973) for a given number of subpopulations and time of divergence from their common ancestry. As the proportional heterozygote deficiency is mathematically equivalent to G_{ST} (Nei, 1977), we examine whether the observed proportional heterozygote deficiency at the VNTR loci in two USA cosmopolitan populations (Budowle *et al.*, 1991), scored by the restriction fragment length polymorphism (RFLP) analysis, can be explained by substructuring within these populations alone. We find that substructuring alone cannot explain the observed proportional deficiencies of apparent heterozygotes in the RFLP data on VNTR loci. To confirm our conclusion that these result from the inherent limitations of the RFLP procedure (described in detail below), we also consider a parallel set of data on comparable populations for which the hypervariable loci are scored by a polymerase chain reaction (PCR)-based protocol, where the same technical limitations are not so critical. Analysis of these data shows that at hypervariable loci, where all genotypes are unequivocally identified, the effect of substructuring on G_{ST} is only trivial, in accordance with the expectation when mutation rates are high, and consequently, when the loci exhibit large numbers of alleles as well as high heterozygosity within subpopulations. These results have an indirect bearing on the statistical interpretation of DNA typing data used in forensics, because the predictions of this model, in conjunction with our previous work (Chakraborty & Jin, 1992), provide estimates of the effect of population substructuring that can be incorporated in the evaluation of the strength of DNA evidence in a forensic case study.

Theory

The proportional heterozygote deficiency as an indicator of population substructure

Consider a population with s subpopulations. Let x_{ik} be the frequency of the ith allele in the kth subpopulation at a locus. If H_k represents the proportion of heterozygote individuals in the kth subpopulation, the actual proportion of heterozygote individuals in the total population (H_S) is:

$$H_S = \sum_k N_k H_k / N, \tag{1}$$

where N_k is the size of the kth subpopulation and $N = \Sigma_k N_k$ is the size of the total population.

Let $\bar{x}_i = \Sigma_k N_k x_{ik} / N$ be the frequency of the ith allele in the total population. Under Hardy-Weinberg equilibrium (HWE), the proportion of heterozygote individuals (H_T) in the total population becomes:

$$H_T = 1 - \sum_i \bar{x}_i^2, \tag{2}$$

so that Nei's coefficient of gene diversity (G_{ST}; Nei, 1973) reduces to

$$G_{ST} = (H_T - H_S)/H_T, \tag{3}$$

the proportional heterozygote deficiency in the total population (Chakraborty & Jin, 1992).

Equation 3 dictates that as H_k within each subpopulation increases, so also would H_S, and consequently G_{ST} would reduce. In other words, for any given structure of subpopulations (i.e. when s is fixed), G_{ST} for hypervariable loci (with large H_k), in comparison to that for the traditional loci, should be smaller.

Nei (1973) defined G_{ST} as:

$$G_{ST} = \frac{D_{ST}}{H_S + D_{ST}}, \tag{4}$$

where $D_{ST} = \Sigma_k \Sigma_l D_{kl} / s^2$, in which D_{kl} is Nei's (1972) minimum genetic distance between the lth and kth subpopulations. Under any mutation-drift model, D_{ST} can be obtained by using:

$$D_{ST} = s(s-1) D_{XY} / s^2, \tag{5}$$

where D_{XY} is the mutation-drift (evolutionary) expectation of D_{kl}s, which is the same for any pair of subpopulations. Algebraically, D_{XY} may also be written as:

$$D_{XY} = 1 - H_S - J_{XY}, \tag{6}$$

where J_{XY} is the probability that two genes drawn randomly, one from each of two populations X and Y, are identical (Nei, 1977).

When G_{ST} is interpreted as a proportional heterozygote deficiency (eqn 3), negative values of G_{ST} are permissible (when $H_T < H_S$). But, in its alternative formation (eqn 4), $0 \leq G_{ST} \leq 1$, as D_{ST} as well as H_S are both bounded by 0 and 1. Of course, if the difference between H_T and H_S (eqns 1 and 2) were truly the result of population substructuring alone, H_T would be larger than H_S (because, in such an event, each subpopulation would be in HWE, so that $H_k = 1 - \Sigma_i x_{ik}^2$). As a consequence, G_{ST} of eqn 3 would always be positive. Therefore, negative observed values of G_{ST} from eqn 3, as well as its exceptionally large values, cannot be ascribed to population substructuring alone, as discussed later.

Mutation models for hypervariable loci and expectations of G_{ST}

For the analysis of traditional serological and biochemical polymorphism, two mutation models have been invoked to explain the maintenance of new allelic variations within and between populations. In the first, it is assumed that each mutation arising in the population is a new one, not previously seen in the population. This is called the infinite allele model (IAM), whose analytical properties are well-studied (Wright, 1949; Kimura & Crow, 1964; Ewens, 1972). For studying genetic variation at a molecular level where genetic alterations can be interpreted in terms of nucleotide substitutions, this model appears most appropriate. Ohta & Kimura (1973) showed that when genetic changes are noted by charge changes produced at a molecule through amino acid changes caused by nucleotide substitutions, mutational alterations can be studied by a stepwise mutation model (SMM). In this model, the allelic states may be represented by a ladder of integer values, so that each mutation may change the allelic state either in the forward or backward direction. Analysis of genetic data gathered through protein electrophoresis in natural populations demonstrated that the stepwise mutation model is an adequate description of mutational changes for the protein electrophoresis protocol of genetic studies (Weir *et al.*, 1976; Fuerst *et al.*, 1977; Chakraborty *et al.*, 1978).

The recently discovered hypervariable polymorphisms (Wyman & White, 1980; Jeffreys *et al.*, 1985), where the genetic variation is caused by copy number variation of short tandemly repeated DNA sequences, may not precisely fit these models of mutations. As of now, the exact molecular mechanism of copy number changes of tandemly repeated sequences is still speculative. Several mechanisms are postulated, such as strand slippage in DNA replication (Levinson & Gutman, 1987), and unequal chromosome exchange in mitosis (Jeffreys *et al.*, 1988; Wolff *et al.*, 1989). Whereas none of these mechanisms may strictly conform to either of the two existing mutation models, several authors have shown that the empirical data on genetic variation at the hypervariable loci often follows predictions of these models (Clarke, 1987; Jeffreys *et al.*, 1988; Flint *et al.*, 1989; Chakraborty *et al.*, 1991; Deka *et al.*, 1991; Edwards *et al.*, 1992). Two recent studies examined this question in further detail. When the hypervariable loci are grouped according to the length of their repeat units, e.g. microsatellites (1–2 base pair (bp) repeat unit), short tandem repeat (STR) loci (3–5 bp repeat unit) and minisatellites (9–70 bp repeat unit), Valdes *et al.* (1993) and Shriver *et al.* (1993) showed that the STR and microsatellite polymorphisms follow the predictions of the stepwise mutation model more consistently than the predictions of the infinite allele model. However, the minisatellite polymorphisms are better described by the infinite allele model (Clark *et al.*, 1989; Flint *et al.*, 1989; Chakraborty *et al.*, 1991). These results may still be only tentative and provisional but it appears that these two models (IAM and one step SMM) may provide plausible realistic magnitudes of genetic variation that may hold for most patterns of hypervariable polymorphisms.

Using the IAM of mutations, Chakraborty & Jin (1992) showed that G_{ST} can be written as a function of the number of subpopulation (s), the within-subpopulation gene diversity (H_S) and the divergence time among subpopulations (in generations t), given by:

$$G_{ST}=\frac{(1-s^{-1})(1-H_S)[1-e^{-H_S T/(1-H_S)}]}{H_S+(1-s^{-1})(1-H_S)[1-e^{-H_S T/(1-H_S)}]}, \tag{7}$$

where $T=t/2N_e$, and N_e is the effective subpopulation size. To determine the maximum expected effect of substructuring, the above result considers the s subpopulations (each of which is at Hardy-Weinberg equilibrium) to have diverged from their common ancestral population t generations ago, and to have remained in isolation from each other since then. Furthermore, we also assume that the effective sizes of the subpopulations are the same (N_e) as that of the ancestral population, and that within each subpopulation allele frequencies are at mutation–drift equilibrium.

Under the one-step stepwise mutation model, J_{XY}(of eqn 6) can be written as (Li, 1976b):

$$J_{XY}=e^{-MT}\sum_{i=-\infty}^{\infty}\left[\frac{1+M-(1-H_S)^{-1}}{M}\right]^{|i|}(1-H_S)I_i(MT), \tag{8}$$

where $M=[(1-H_S)^{-2}-1]/2$, $T=t/2N_e$, and $I_i(x)$ is a modified Bessel function of the first kind and defined as:

$$I_i(2x)=\sum_{s=0}^{\infty}x^{2s+i}/[s!(i+s)!]. \tag{9}$$

By using eqns 4, 5 and 6, we have:

$$G_{ST}=\frac{(s-1)(1-H_S-J_{XY})}{sH_S+(s-1)(1-H_S-J_{XY})}, \tag{10}$$

giving the predicted G_{ST} value under the stepwise mutation model. Substituting eqns 8 and 9 in eqn 10, we may then define G_{ST} under the one-step stepwise mutation model as a function of s, H_S and $T=(t/2N_e)$. Therefore, like IAM, for a given level of gene diversity (or heterozygosity) within subpopulations, the coefficient of gene diversity (G_{ST}) is specified by the number of subpopulations (s) and their time of divergence ($t/2N_e$) in units of $2N_e$ generations under the SMM.

Figure 1 shows the relationships of G_{ST} with the number of subpopulations (s) and the isolation time since divergence (T) where $H_S=90$ per cent. For any given number of subpopulations (s) and their divergence time (T), G_{ST} of IAM is always greater than that of SMM when the mutation rates for both models are the same. This is so because for a given time of divergence, under the IAM subpopulations would exhibit greater differentiation than that which would be predicted under the SMM for the same rate of mutation. However, for the same within-population diversity (H_S), the ratio of the mutation rates of the SMM (v_{SMM}) and of the IAM (v_{IAM}) under the assumption of mutation–drift equilibrium is $1+2N_e v_{IAM}$ which is always larger than 1. Interestingly enough, it can be shown numerically (e.g. Fig. 1) that the G_{ST} under the IAM is still larger than that under the SMM.

The numerical computations shown in Fig. 1 also illustrate other parallel properties of expected G_{ST} under the two mutation models. Under both models of mutation, G_{ST} very quickly reaches its plateau as a function of both s and $t/2N_e$. In fact, the number of subpopulations beyond 10 has hardly any effect on G_{ST} in either model, and even in terms of $t/2N_e$, the rate of approach to the asymptotic G_{ST} is much slower under the SMM in comparison to that under the IAM.

Equation 8 also shows that J_{XY} goes to zero when $t \to \infty$. In other words, the subpopulations eventually will share no allele when each of them accumulates a large number of new mutations. In this case, the asymptotic value of G_{ST} of the SMM is the same at that of the IAM, i.e.

$$G_{ST}=\frac{(s-1)(1-H_S)}{s-1+H_S} \tag{11}$$

In the light of eqn 11, the numerical calculations shown in Fig. 1 may appear deceptive as it seems as though the asymptote for G_{ST} under the SMM is lower than that of the IAM for fixed s and H_S. Actually, this is not the case; it simply reflects that the slower rate of

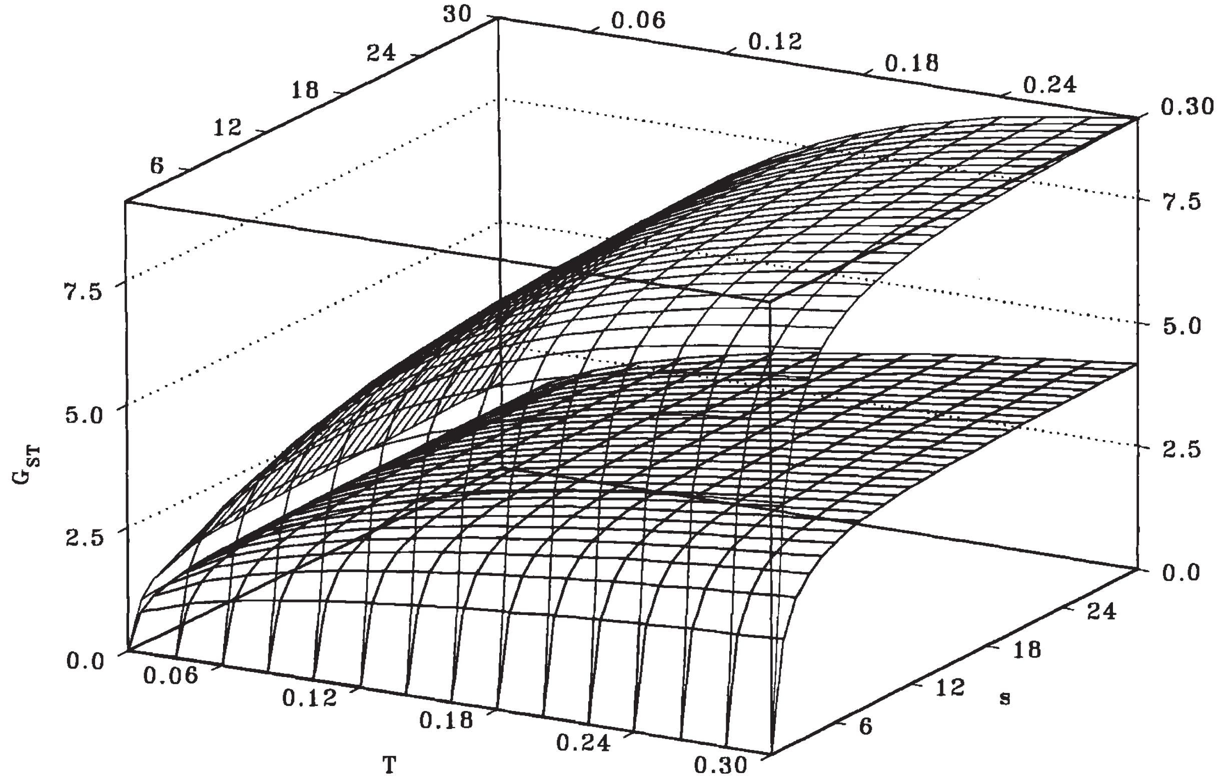

Fig. 1 Expected coefficient of gene diversity (G_{ST} as a percentage) in a subdivided population as a function of the number of subpopulations (s), and their time of divergence, $T(=t/2N_e)$. The higher surface is for the Infinite Allele Model (IAM), and the lower surface is for the one-step Stepwise Mutation Model (SMM) of mutations.

approach to an asymptotic G_{ST} value under the SMM in comparison to the IAM applies even when the within-subpopulation heterozygosity ($H=90$ per cent in Fig. 1) is kept fixed for both models, and consequently the mutation rate is made larger for the SMM (in comparison with the IAM).

In our previous work, we indicated that, for the IAM, G_{ST} should become smaller as H_S becomes larger (Chakraborty & Jin, 1992). This property also holds for the SMM. Figure 2 shows the relationship of G_{ST} with H_S for three sets of s and $t/2N_e$, with $t/2N_e$ ranging from 0.05 to 0.5 and s from 2 to 20. For both models G_{ST} decreases when H_S increases, indicating that a locus with a higher mutation rate generally exhibits a smaller coefficient of gene differentiation (G_{ST}) compared with the locus with a reduced mutation rate. A locus with a higher mutation rate can accumulate more new mutations and thus will exhibit a larger within-population heterozygosity (H_S) and between-population divergence (D_{ST}) than a locus with a smaller mutation rate for any given number of subpopulations and the divergence time. The numerical illustrations shown in Fig. 2, therefore, provide a theoretical support of Morton *et al.*'s (1993) finding that at hypervariable loci (where the average heterozygosity is of the order of 90 per cent, or higher) the kinship (mathematically equivalent to G_{ST}) is almost an order of magnitude smaller than at the traditional blood group and protein loci (where H_S is generally 40 per cent, or lower).

These analytical properties may also be described in terms of isolines of G_{ST} for various values of s and $t/2N_e$. Figure 3 shows some computations in this regard, where combinations of the time of divergence ($t/2N_e$) and the number of subpopulations (s) required to explain a given value of G_{ST} ($=5$ per cent) are plotted for the two mutation models (IAM and SMM) for $H_S=70$ per cent and 90 per cent. These computations again illustrate that a number of subpopulations larger than 10 has virtually no effect on the value of G_{ST}, and even when subpopulations exchange no genes among them, large G_{ST} (say $\geqslant 5$ per cent) would be unlikely to arise from population substructuring when each subpopulation exhibits high levels of heterozygosity (say, H_S $\geqslant 90$ per cent). Chakraborty & Jin (1992) also showed that a small amount of gene flow among subpopulations reduced the values of G_{ST} even more drastically. While the effect of migration was analyti-

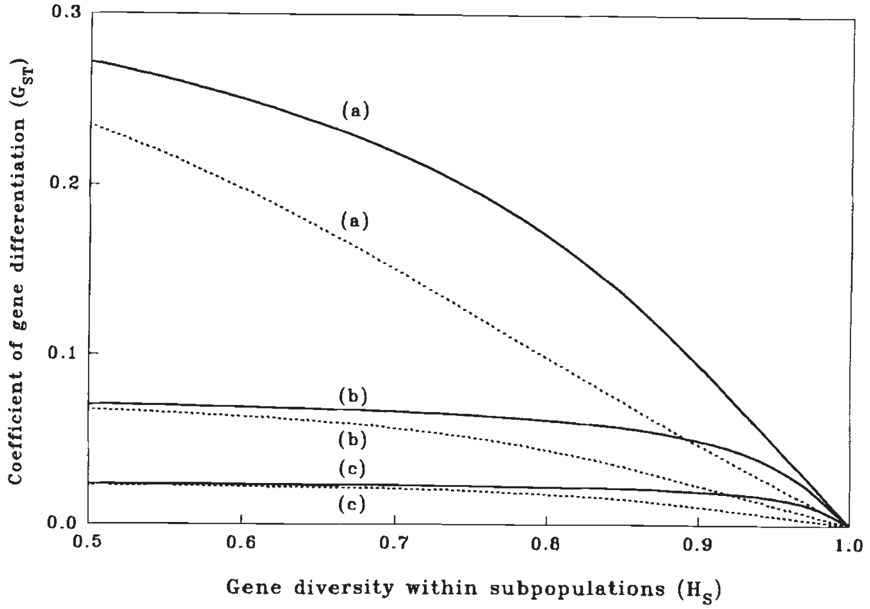

Fig. 2 Expected coefficient of gene diversity (G_{ST}) under the two mutation models (IAM: solid lines; SMM: dashed lines) in a subdivided population as functions of within-subpopulation heterozygosity (H_S) for three different situations of substructuring: (a) $s = 2$, $t/2N_e = 0.05$; (b) $s = 5$, $t/2N_e = 0.10$; and (c) $s = 10$, $t/2N_e = 0.50$.

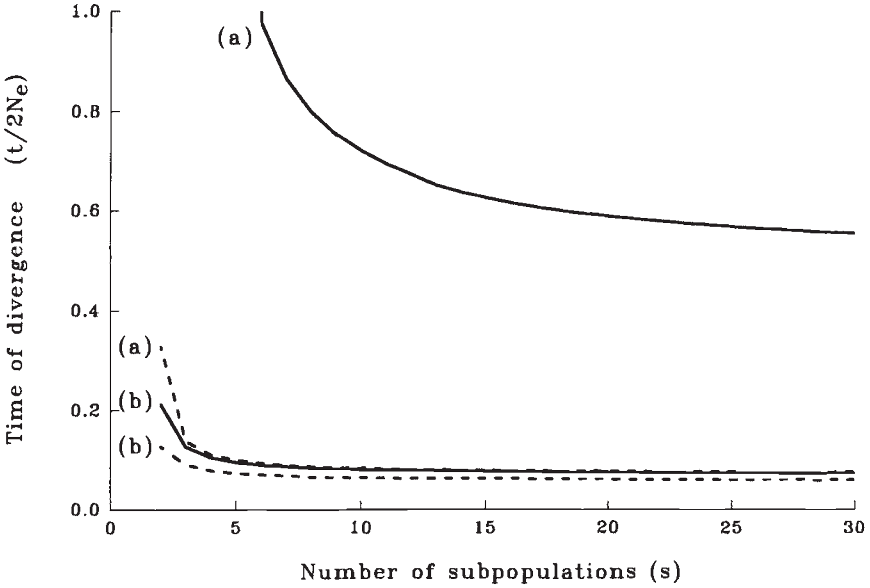

Fig. 3 Isolines of different substructuring situations (s and $t/2N_e$) resulting in $G_{ST} = 5$ per cent under the two mutation models (IAM: solid lines; SMM: dashed lines) for (a) $H_S = 70$ per cent and (b) $H_S = 90$ per cent.

cally studied for the IAM, comparable analytical results for expected G_{ST} in the presence of migration for the SMM are not yet available.

Data analysis

RFLP data

Firstly, we consider data on six Variable Number of Tandem Repeat (VNTR) loci (D1S7, D2S44, D4S139, D14S13, D16S85 and D17S79) as reported in Budowle *et al.* (1991), which were reanalysed in our previous work (Chakraborty & Jin, 1992). Budowle *et al.* (1991) observed significant deficiencies of apparent heterozygosities at two loci (D1S7 and D17S79) for the USA Caucasians and at five loci (all except D4S139) for the USA Black populations, in which all single-banded DNA patterns at each locus were treated as homozygotes. The proportional heterozygote deficiencies observed in these cases range between 3.07 per cent (D1S7 in Blacks) and 10.75 per cent (D17S79 in Caucasians) (see Table 1). Table 1 also includes the G_{ST} values under the two mutation models (IAM and SMM), assuming no gene migration among subpopulations within each racial group. As G_{ST} is virtually unaltered for $s \geq 10$ (see Fig. 1), the

Table 1 Observed and expected heterozygosities, and observed and expected G_{ST} without migration under the Infinite Allele Model (IAM) and Stepwise Mutation Model (SMM) at six VNTR loci in USA Caucasians and Blacks

					Expected G_{ST}^{a} ($t/2N_e = 0.1$)			
		Heterozygosity			$s=2$		$s=10$	
Locus	Sample size (N)	Observed[a]	Expected[a]	G_{ST}[a]	IAM	SMM	IAM	SMM
Caucasians								
D1S7	210	91.0*	94.2	3.50	3.06	1.32	5.37	2.34
D2S44	218	91.3	92.0	0.75	3.00	1.27	5.28	2.27
D4S139	144	87.5	90.1	2.88	3.47	1.84	6.08	3.26
D14S13	218	89.9	91.2	1.42	3.20	1.48	5.62	2.64
D16S85	210	90.0	90.8	0.83	3.19	1.47	5.60	2.61
D17S79	209	70.8**	79.3	10.75	4.25	3.59	7.40	6.28
Blacks								
D1S7	268	91.4*	94.3	3.07	2.98	1.26	5.24	2.24
D2S44	295	88.1**	93.4	5.66	3.41	1.74	5.97	3.10
D4S139	304	90.8	92.6	1.97	3.08	1.35	5.41	2.40
D14S13	258	90.7*	94.0	3.52	3.09	1.36	5.43	2.42
D16S85	212	77.8*	83.7	7.03	4.05	3.04	7.05	5.34
D17S79	281	80.8*	85.8	5.90	3.92	2.72	6.84	4.79

[a] Heterozygosity and G_{ST} values are expressed as percentages.
*Significant deficiency ($P \leqslant 0.05$).
**Significant deficiency ($P \leqslant 0.01$).

number of subpopulations (s) was chosen as 2 and 10 for these computations. We used $t/2N_e = 0.1$, which amounts to 20000–25000 years for $N_e = 5000$. This is a conservative estimate of the time of separation of the different ethnic populations within the Caucasians or Blacks (Nei & Roychoudhury, 1982; Chakraborty & Jin, 1992). In all seven population–locus combinations where significant heterozygote deficiencies are observed, the observed G_{ST} values are larger than their predictions under the SMM. Therefore, we reiterate our previous conclusion (Chakraborty & Jin, 1992), namely, if the observed proportional deficiencies of heterozygotes at these six VNTR loci are to be ascribed only to substructuring within USA Caucasians or Blacks, we have to assume that the different subpopulations never exchanged any mates among them and that their time of separation is much longer than that suggested by their ethnohistory. Neither of these postulates can be correct (Chakraborty & Jin, 1992), considering the substantial amount of gene migration within ethnic populations of USA Caucasians and Blacks, at least in this century.

PCR data

Edwards *et al.* (1992) reported the data at five short tandem repeat (STR) loci (THO1, RENA4, FABP, HPRTB and ARA) in four ethnic populations (Caucasians, Blacks, Hispanics and Asians) currently residing in Houston, Texas. These loci consist of tri- or tetranucleotide repeat units (Table 2), with respect to which copy number variations were scored to determine the allelic types. DNA samples were first amplified by the polymerase chain reaction (PCR) technique and then the alleles were separated and identified by acrylamide gel electrophoresis. Of 20 population–locus combinations, only two contrasts of observed and expected heterozygosities resulted in significant heterozygote deficiency (Table 2). Both are only marginally significant ($P \sim 0.05$). Therefore, although the populations in this study are comparable to the ones considered in Budowle *et al.* (1991), we do not find any evidence of a substantial degree of substructuring at these STR loci. This is particularly intriguing because in general these short tandem repeat (STR) loci show a lower level of observed heterozygosities (39 per cent, 64 per cent, 71 per cent, 75 per cent, 86 per cent) than the VNTR loci (77 per cent, 84 per cent, 90 per cent, 90 per cent, 90 per cent, 91 per cent) and, hence, should substructuring alone be the true cause of the observed heterozygote deficiencies, the effect should have been more pronounced at the STR loci. Furthermore, of the five STR loci, the RENA4 locus exhibits the lowest level of heterozygosity (37–41 per cent), yet for no population

Table 2 Observed and expected (under HWE) total number of heterozygotes for five STR loci in four populations

		Populations				
Locus	Statistics	Whites	Blacks	Hispanics	Asians	Pooled
HPRTB	N	134	90	46	30	300
(AGAT)n	Heterozygotes					
	Observed	95	63	33	22	213*
	Expected ± SE	103.2 ± 4.9	69.8 ± 4.0	33.1 ± 3.1	19.5 ± 2.6	229.2 ± 7.4
TH01	N	186	185	192	77	640
(AATG)n	Heterozygotes					
	Observed	150	149	135	47*	481*
	Expected ± SE	143.1 ± 5.7	141.0 ± 5.8	146.1 ± 5.9	55.3 ± 3.9	505.1 ± 10.3
RENA4	N	177	186	187	75	625
(ACAG)n	Heterozygotes					
	Observed	66	77	71	29	243
	Expected ± SE	64.6 ± 6.4	82.2 ± 6.8	72.1 ± 6.7	29.2 ± 4.2	249.2 ± 12.2
FABP	N	287	191	176	76	730
(AAT)n	Heterozygotes					
	Observed	176	159	91	42	468*
	Expected ± SE	186.0 ± 8.1	153.4 ± 5.5	100.2 ± 6.6	43.6 ± 4.3	503.3 ± 12.5
ARA	N	78	81	43	31	228
(AGC)n	Heterozygotes					
	Observed	63	67*	38	28	196*
	Expected ± SE	64.8 ± 2.7	73.7 ± 2.6	38.8 ± 2.0	27.7 ± 1.7	207.6 ± 4.3

*Significant deficiency ($P \leqslant 0.05$).

does this locus show any significant heterozygote deficiency. The only two significant heterozygote deficiencies occur for locus–population combinations where the observed levels of heterozygosity are not so small (61 per cent for THO1 in the Asians, and 83 per cent for ARA in Blacks). In contrast, if we consider the pooled data (i.e. create a known substructuring), four of the five loci exhibit significant heterozygote deficiencies (last column of Table 2). Therefore, we claim that the effects of substructuring within these four racial groups at the five STR loci is very small, if not negligible.

By combining Caucasians with Blacks, Caucasians with Asians and Blacks with Asians, we created three populations with known population substructure. The observed G_{ST}s and their corresponding expectations under both the IAM and the SMM are presented in Table 3. For times of divergence between subpopulations we have used 115 000 years, 55 000 years and 120 000 years for Caucasians and Blacks, Caucasians and Asians, and Blacks and Asians, respectively (Nei, 1975). Only three of the 15 locus–population combinations reveal significant heterozygote deficiency. However, contradicting the situation for the six VNTR loci, none of the observed G_{ST} exceeds the predictions of the IAM, and most of them (except two) are also smaller than the predictions of the SMM. Keeping in mind that a complete isolation of subpopulations was assumed in evaluating the expected G_{ST} values, the generally lower observed G_{ST} values in relation to their predicted values indicate that a certain amount of migration did occur in the past. In other words, these calculations indicate that for populations with known substructuring, the extent of heterozygote deficiencies in the total population can be predicted by observed values of G_{ST}.

Discussion and conclusions

The results shown above for the two sets of hypervariable loci are apparently discordant with each other. The six VNTR loci, scored by the RFLP procedure, show significant heterozygote deficiencies for the two major racial populations of the USA, whereas the five STR loci show results quite contrary to this. Translated into terms of G_{ST}, the VNTR loci exhibit higher levels of coefficient of gene differentiation than that at the STR loci. This is contrary to the theory of substructuring presented above. Should the observed heterozygote deficiencies noted at these VNTR loci be entirely the result of substructuring alone, we would expect higher levels of G_{ST} for the STR loci. Furthermore, for all locus-population combinations, the observed levels of G_{ST} should have been smaller than the predicted ones because we used levels of substructuring more severe than the reality by allowing a larger time of divergence, and by assuming no gene migration between subpopulations. These should have occurred for either of the two mutation models considered here.

With construction of known levels of subpopulations, the STR loci exhibit a tendency towards such predictions (Table 3). That leads us to conclude that the apparently discordant findings at the six VNTR loci result from the designation of homozygotes and heterozygotes from RFLP analysis of hypervariable loci. Devlin *et al.* (1990) and Chakraborty *et al.* (1992) have shown that there is an inherent limitation of genotype assignment from the Southern blot RFLP analysis of DNA typing. Alleles of nearly similar size may not always be distinguished from each other, and the alleles that are too small or too large may not be reliably sized in such a protocol. As a result, the single-banded individuals may not always be true homozygotes. Coalescence or nondetectability of alleles in the Southern blot RFLP analysis, therefore, might cause apparent heterozygote deficiency, mimicking a substructuring effect. Devlin *et al.* (1990) have shown that the coalescence phenomenon is inherent in allele sizing from Southern blot data. Similarly, the presence of nondetectable alleles is demonstrated by restriction digestion with alternative enzymes (e.g. *Pvu*II), whereby small *Hae*III derived fragments (that remained undetected in the original database) could be identified (see Budowle *et al.*, 1991 and Chakraborty & Jin, 1992 for citations). Furthermore, Jeffreys *et al.* (1991) present other data and cite examples of true 'nondetectable' alleles, and Fornage *et al.* (1992) present direct evidence of small-size alleles that are detectable by PCR, which would have remained undetected by a RFLP analysis.

Table 3 Observed and expected G_{ST} without migration under the Infinite Allele Model (IAM) and Stepwise Mutation Model (SMM) at five STR loci in three mixtures of Caucasian, Black and Asian population data from Texas, USA

		Observed[a]		Expected G_{ST}[a]	
Locus	Sample size	H_S	G_{ST}	SMM	IAM
Mixture of Caucasians and Blacks					
HPRTB	224	70.5*	9.13	8.75	13.52
THO1	371	80.6	—[b]	5.49	9.85
RENA4	363	39.4	2.72	17.41	19.35
FABP	478	70.1	5.08	8.89	13.63
ARA	154	84.4*	7.57	4.31	8.11
Mixture of Caucasians and Asians					
HPRTB	164	71.3	6.33	6.41	9.06
THO1	263	74.9	5.43	5.63	8.58
RENA4	252	37.7	—[b]	10.64	11.24
FABP	363	60.1	5.47	8.40	10.12
ARA	104	87.5	1.92	2.71	5.75
Mixture of Blacks and Asians					
HPRTB	120	70.8	7.28	8.77	13.65
THO1	262	74.8	3.45	7.44	12.29
RENA4	261	40.6	5.55	17.59	19.75
FABP	267	75.3	1.68	7.27	12.10
ARA	112	84.8*	7.27	4.24	7.96

[a] Heterozygosity (H_S) and G_{ST} values are expressed as percentages.

[b] Heterozygote excess was found in these cases (and hence, G_{ST} should be estimated as zero because negative G_{ST} cannot occur).

*Significant deficiency ($P \leqslant 0.05$).

If one entertains the possibility of 'nondetectable' alleles as a source of large apparent heterozygote deficiency at the six VNTR loci, it is reasonable to ask what frequency of 'nondetectable' alleles would explain these discordant findings. Chakraborty *et al.* (1992a) addressed this issue, and showed that even the largest value of G_{ST} (10.75 per cent at D17S79 in Caucasians) can be explained with about 5 per cent nondetectable alleles. Although for some PCR primers, nondetectability of alleles may occur from differential amplifications or nucleotide substitutions within the primer sequence (Callan *et al.*, 1993; Koorey *et al.*, 1993), no such evidence exists for the primers used in the STR survey data (Edwards *et al.*, 1992) of the present analysis. Therefore, we argue that the level of G_{ST} values obtained for the STR loci are realistic effects of substructuring within the major cosmopolitan populations of the USA.

Because at present RFLP data are the primary basis of forensic applications of DNA typing, one might ask, 'In the presence of such levels of heterozygote deficiencies how can one support the calculations of DNA profile frequencies using the traditional Hardy-Weinberg principles?' As argued in Chakraborty *et al.* (1992a,b), three levels of conservatism support the current computations. Firstly, allele frequencies used in DNA forensics are exaggerated as the binning of allele sizes (Budowle *et al.*, 1991) produces allele frequencies that are on an average two times as large as they should be (in reference to the match criteria used by the forensic laboratories; Budowle & Monson, 1992; Chakraborty *et al.*, 1993). Secondly, in the presence of 'nondetectable' alleles, the gene count estimates of allele frequencies are overestimates of the true allele frequencies in a population (Chakraborty *et al.*, 1992a). Thirdly, even if substructuring exists in a population, the Hardy–Weinberg prediction of genotype frequencies for heterozygotes provides an overestimate for the true frequency of heterozygotes (Chakraborty *et al.*, 1992b). The use of the modified estimate ($2p$ instead of p^2) of Budowle *et al.* (1991) guards against the possibility of under-reporting of the homozygote (single-banded) frequencies.

These arguments, we might note, apply to the current forensic computations of the unconditional probability of a specific DNA profile in a population. Some geneticists, however, argue that in a DNA forensic analysis one might also attempt to determine the probability with which an individual (say, x) has the specific profile, given that a defendant has the same profile. In general, this is *not* the one predicted by the unconditional profile frequency in the population. The answer to the above alternative question will vary according to the degree of shared ancestry of x with the defendant, which in turn may range over close relatives, similar ethnic background and the like. To what extent this extraneous information changes the conditional probabilities from the unconditional ones is discussed at length by Morton (1992) and Balding & Nichols (1994).

In summary, this work shows that even though at present the exact mutational model for analysing hypervariable loci is not known, the expected effect of substructuring can be postulated from the two existing mutation models (IAM and SMM) used in population genetic analysis. Taking into account the ethnohistory of human populations, both models predict that the effect of substructuring in the major cosmopolitan populations at hypervariable loci should be small, much smaller than the levels seen in empirical studies involving the traditional blood groups and protein loci. Use of G_{ST} (or F_{ST}) of levels equivalent to 5–10 per cent (Nichols & Balding, 1991) appears to be too large, as shown in the present analysis. This is so because gene migration among ethnic groups of racial populations has been documented to be extensive as well as long-standing, both from demographic (Kennedy, 1944) and molecular studies (Bowcock *et al.*, 1991).

We recognize, however, that in some parts of the world subpopulations may have effective population sizes (N_e) much smaller than the empirical value of 5000 we used in our numerical analysis. When forensic calculations have to consider a relevant population where isolation and mating practice necessitate a smaller value of N_e, the predicted G_{ST} values have to be changed accordingly. Equations 7, 8 and 10, along with empirical values of H_S in such populations, offer guides to the appropriate G_{ST} in such special cases. Until such studies are complete, we contend that when the true identity of individuals contributing a DNA sample is unknown, the frequency estimates for the combined multilocus profile should be presented for as many populations as possible that are relevant in a forensic case analysis.

Acknowledgements

This work was supported by US Public Health Service Research grants GM 41399 and GM 45861 from the National Institutes of Health, and grant NIJ-92-CX-K024 from the US National Institute of Justice. Comments from an anonymous reviewer were particularly helpful in improving the presentation.

References

BALDING, D. J. AND NICHOLS, R. A. 1994. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci. Int.*, **64**, 125–140.

BOWCOCK, A. M., KIDD, J. R., MOUNTAIN, J. L., HERBERT, J. M., CAROTENUTO, L., KIDD, K. K. AND CAVALLI-SFORZA, L. L. 1991. Drift, admixture and selection in human evolution: a study with DNA polymorphism. *Proc. Natl. Acad. Sci. U.S.A.*, **88**, 839–843.

BUDOWLE, B., GIUSTI, A. M., WAYE, J. S., BAECHTEL, F. S., FOURNEY, R. M., ADAMS, D. E., PRESLEY, L. A., DEADMEN, H. A. AND MONSON, K. L. 1991. Fixed-bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci for use in forensic comparisons. *Am. J. Hum. Genet.*, **48**, 841–855.

BUDOWLE, B. AND MONSON, K. L. 1992. Perspective on the fixed bin method and the floor approach/ceiling principle. *Proceedings from The Third International Symposium on Human Identification*, 1992, pp. 391–406.

CALLEN, D. F., THOMPSON, A. D., SHEN, Y., PHILLIPS, H. A., RICHARDS, R. I., MULLEY, J. C. AND SUTHERLAND, G. R. 1993. Incidence and origin of 'null' alleles in the (AC)n microsatellite markers. *Am. J. Hum. Genet.*, **52**, 922–927.

CHAKRABORTY, R. 1993. Analysis of genetic structure of populations: meaning, methods and implications. In: Majumder, P. P. (ed.) *Modern Human Genetics: A Centennial Tribute to J.B.S. Haldane*, pp. 189–206. Plenum, New York.

CHAKRABORTY, R., DE ANDRADE, M., DAIGER, S. P. AND BUDOWLE, B. 1992a. Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. *Ann. Hum. Genet.*, **56**, 45–57.

CHAKRABORTY, R., FORNAGE, M., GUEGUE, R. AND BOERWINKLE, E. 1991. Population genetics of hypervariable loci: analysis of PCR based VNTR polymorphism within a population. In: Burke, T., Dolf, G., Jeffreys, A. J. and Wolff, R. (eds) *DNA Fingerprinting: Approaches and Applications*, pp. 127–143. Birkhäuser Verlag, Basel.

CHAKRABORTY, R., FUERST, P. A. AND NEI, M. 1978. Statistical studies on protein polymorphism in natural populations. II. Gene differentiation between populations. *Genetics*, **88**, 367–390.

CHAKRABORTY, R. AND JIN, L. 1992. Heterozygote deficiency, population substructure and their implications in DNA fingerprinting. *Hum. Genet.*, **88**, 267–272.

CHAKRABORTY, R., JIN, L., ZHONG, Y., SRINIVASAN, M. R. AND BUDOWLE, B. 1993. On allele frequency computation from DNA typing data. *Int. J. Legal Med.*, **106**, 103–106.

CHAKRABORTY, R. AND KIDD, K. K. 1991. The utility of DNA typing in forensic work. *Science*, **254**, 1735–1739.

CHAKRABORTY, R. AND NEI, M. 1974. Dynamics of gene differentiation between incompletely isolated populations of unequal sizes. *Theor. Pop. Biol.*, **5**, 460–469.

CHAKRABORTY, R. AND RAO, D. C. 1972. Detection of the inbreeding coefficient from ABO blood-group data. *Am. J. Hum. Genet.*, **24**, 352–354.

CHAKRABORTY, R., SRINIVISAN, M., JIN, L. AND DE ANDRADE, M. 1992b. Effects of population subdivision and allele frequency differences on interpretation of DNA typing data from human identification. *Proceedings from The Third International Symposium on Human Identification*, 1992, pp. 205–222.

CHAKRABORTY, R. AND ZHONG, Y. 1993. Statistical power of an exact test of Hardy–Weinberg proportions of genotypic data on a multi-allelic locus. *Hum. Hered.*, **44**, 1–9.

CLARK, A. G. 1987. Neutrality tests of highly polymorphic restriction-fragment-length polymorphisms. *Am. J. Hum. Genet.*, **41**, 948–956.

DEKA, R., CHAKRABORTY, R. AND FERRELL, R. E. 1991. A population genetic study of six VNTR loci in three ethnically defined populations. *Genomics*, **11**, 83–92.

DEVLIN, B., RISCH, N. AND ROEDER, K. 1990. No excess of homozygosity at loci used for DNA fingerprinting. *Science*, **24**, 1416–1420.

EDWARDS, A., HAMMOND, H., JIN, L., CASKEY, C. T. AND CHAKRABORTY, R. 1992. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics*, **12**, 241–253.

EWENS, W. J. 1972. The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.*, **3**, 87–112.

FLINT, J., BOYCE, A. J., MARTINSON, J. J. AND CLEGG, J. B. 1989. Population bottlenecks in Polynesia revealed by minisatellites. *Hum. Genet.*, **83**, 257–263.

FORNAGE, M., CHAN, L., SIEST, G. AND BOERWINKLE, E. 1992. Allele frequency distribution of the $(TG)_n(AG)_m$ microsatellite in the apolipoprotein C-II gene. *Genomics*, **12**, 63–68.

FUERST, P. A., CHAKRABORTY, R. AND NEI, M. 1977. Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. *Genetics*, **86**, 455–483.

JEFFREYS, A. J., ROYLE, N. J., PATEL, I., ARMOUR, J. A. L., MACLEOD, A., COLLICK, A., GRAY, I. C., NEUMANN, R., GIBBS, M., GROSIER, M., HILL, M., SIGNER, E. AND MONCKTON, D. 1991. Principles and recent advances in human DNA fingerprinting. In: Burke, T., Dolf, G., Jeffreys, A. J. and Wolff, R. (eds) *DNA Fingerprinting: Approaches and Applications*, pp. 1–19. Birkhäuser Verlag, Basel.

JEFFREYS, A. J., ROYLE, N. J., WILSON, V., AND WONG, Z. 1988. Spontaneous mutation rates to new length alleles at tandem-repetitive loci in human DNA. *Nature*, **332**, 278–281.

JEFFREYS, A. J., WILSON, V. AND THEIN, S. L. 1985. Hypervariable 'minisatellite' regions in human DNA. *Nature*, **314**, 67–73.

JORDE, L. B. 1980. The genetic structure of subdivided human populations. In: Mielke, J. H. and Crawford, M. H. (eds) *Current Developments in Anthropological Genetics*, pp. 135–208. Plenum, New York.

KARLIN, S. 1982. Classifications of selection–migration structures and conditions for a protected polymorphism. *Evol. Biol.*, **14**, 61–204.

KENNEDY, R. J. R. 1944. Single or triple melting pot? Intermarriage trends in New Haven, 1870–1940. *Am. J. Sociol.*, **49**, 331–339.

KIMURA, M. AND CROW, J. F. 1964. The number of alleles that can be maintained in a finite population. *Genetics*, **49**, 725–738.

KOOREY, D. J., BISHOP, G. A. AND McCAUGHAN, G. W. 1993. Allele nonamplification: a source of confusion in linkage studies employing microsatellite polymorphisms. *Hum. Mol. Genet.*, **2**, 289–291.

LEVINSON, G. AND GUTMAN, G. A. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.*, **4**, 203–221.

LEWONTIN, R. C. AND HARTL, D. L. 1991. Population genetics in forensic DNA typing. *Science*, **254**, 1745–1750.

LI, W.-H. 1976a. Effect of migration on genetic distance. *Am. Nat.*, **110**, 841–847.

LI, W.-H. 1976b. Electrophoretic identity of proteins in a finite population and genetic distance between taxa. *Genet. Res.*, **43**, 45–52.

MORTON, N. E. 1969. Human population structure. *Ann. Rev. Genet.*, **3**, 53–73.

MORTON, N. E. 1992. Genetic structure of forensic populations. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 2556–2560.

MORTON, N. E., COLLINS, A. AND BALAZS, I. 1993. Kinship bioassay on hypervariable loci in Blacks and Caucasians. *Proc. Natl. Acad. Sci. U.S.A .*, **90**, 1892–1896.

NATIONAL RESEARCH COUNCIL. 1992. *DNA Technology in Forensic Science*. National Academy Press, Washington DC.

NEI, M. 1973. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U.S.A.*, **70**, 3321–3323.

NEI, M. 1975. *Molecular Population Genetics and Evolution*. North-Holland, Amsterdam.

NEI, M. 1977. *F*-statistics and analysis of gene diversity in subdivided populations. *Ann. Hum. Genet.*, **41**, 225–233.

NEI, M. AND FELDMAN, M. W. 1972. Identity of genes by descent within and between populations under mutation and migration pressures. *Theor. Pop. Biol.*, **3**, 460–465.

NEI, M. AND ROYCHOUDHURY, A. K. 1982. Genetic relationship and evolution of human races. *Evol. Biol.*, **14**, 1–59.

NICHOLS, R. A. AND BALDING, D. J. 1991. Effects of population structure on DNA fingerprint analysis in forensic science. *Heredity*, **66**, 297–302.

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

ROYCHOUDHURY, A. K. AND NEI, M. 1988. *Human Polymorphic Genes*. Oxford University Press, New York.

SHRIVER, M. D., JIN, L., CHAKRABORTY, R. AND BOERWINKLE, E. 1993. VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. *Genetics*, **134**, 983–993.

SLATKIN, M. 1985. Rare alleles as indicators of gene flow. *Evolution*, **39**, 53–65.

VALDES, A. M., SLATKIN, M. AND FREIMER, N. B. 1993. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics*, **133**, 737–749.

WARD, R. H. AND SING, C. F. 1970. A consideration of the power of the chi-square test to detect inbreeding effects in natural populations. *Am. Nat.*, **104**, 355–365.

WEIR, B. S., BROWN, A. H. D. AND MARSHALL, D. R. 1976. Testing for selective neutrality of electrophoretically detectable protein polymorphisms. *Genetics*, **84**, 639–659.

WOLFF, R. K., PLAETKE, R., JEFFREYS, A. J. AND WHITE, R. 1989. Unequal crossing over between homologous chromosomes is not the major mechanism involved in the generation of new alleles at VNTR loci. *Genomics*, **5**, 382–384.

WRIGHT, S. 1931. Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

WRIGHT, S. 1943. Isolation by distance. *Genetics*, **28**, 114–138.

WRIGHT, S. 1949. Genetics of populations. *Encyclopedia Britannica*, 14th edn, **10**, 111–112.

WYMAN, A. R. AND WHITE, R. 1980. A highly polymorphic locus in human DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **77**, 6754–6758.