Am. J. Hum. Genet. 51:534–548, 1992

Ethnic Differentiation at VNTR Loci, with Special Reference to Forensic Applications

B. Devlin* and Neil Risch*,†

Departments of *Epidemiology and Public Health and †Human Genetics, Yale University, New Haven

Summary

Allele-rich VNTR loci provide valuable information for forensic inference. Interpretation of this information is complicated by measurement error, which renders discrete alleles difficult to distinguish. Two methods have been used to circumvent this difficulty—i.e., binning methods and direct evaluation of allele frequencies, the latter achieved by modeling the data as a mixture distribution. We use this modeling approach to estimate the allele frequency distributions for two loci—D17S79 and D2S44—for black, Caucasian, and Hispanic samples from the Lifecodes and FBI data bases. The data bases are differentiated by the restriction enzyme used: *Pst*I (Lifecodes) and *Hae*III (FBI). Our results show that alleles common in one ethnic group are almost always common in all ethnic groups, and likewise for rare alleles; this pattern holds for both loci. Gene diversity, or heterozygosity, measured as one minus the sum of the squared allele frequencies, is greater for D2S44 than for D17S79, in both data bases. The average gene diversity across ethnic groups when *Pst*I (*Hae*III) is used is .918 (.918) for D17S79 and is .985 (.983) for D2S44. The variance in gene diversity among ethnic groups is greater for D17S79 than for D2S44. The number of alleles, like the gene diversity, is greater for D2S44 than for D17S79. The mean numbers of alleles across ethnic groups, estimated from the *Pst*I (*Hae*III) data, are 40.25 (41.5) for D17S79 and 104 (103) for D2S44. The number of alleles is correlated with sample size. We use the estimated allele frequency distributions for each ethnic group to explore the effects of unwittingly mixing populations and thereby violating independence assumptions. We show that, even in extreme cases of mixture, the estimated genotype probabilities are good estimates of the true probabilities, contradicting recent claims. Because the binning methods currently used for forensic inference show even less differentiation among ethnic groups, we conclude that mixture has little or no impact on the use of VNTR loci for forensics.

Introduction

Allele-rich VNTR loci should provide a wealth of information for human genetics. Because of the limited resolving power of current laboratory methods, however, estimating allele frequencies can be challenging. Even when the alleles themselves are defined by length and are therefore discrete, measurement error is often large enough to make the exact fragment sizes, and therefore the alleles, unresolvable (but, for an example of a discrete VNTR allele distribution, see Budowle et al. 1991*a*). Instead of a discrete allele distribution, the data follow a mixture distribution, being a mixture of the discrete allele distribution and the continuous measurement error distributions (Devlin et al. 1991*a*). One approach to evaluating the information of these data is to categorize the measured fragment lengths into discrete groups and then to treat these groups (bins) as alleles (e.g., see Balazs et al. 1989; Budowle et al. 1991*b*). On the other hand, using relatively few assumptions, Devlin et al. (1991*a*) have shown that allele frequencies are estimable for some VNTR loci by modeling the data as a mixture distribution. The modeling approach, then, allows us to evaluate the underlying discrete distribution.

Received January 7, 1992; revision received March 16, 1992.

Address for correspondence and reprints: B. Devlin, Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, P.O. Box 3333, New Haven, CT 06510.

0002-9297/92/5103-0010$02.00

A common use of VNTR loci is for criminal inference, especially the evaluation of evidentiary samples (Budowle et al. 1991*b*). VNTR profiles may vindicate a suspect when the suspect and evidentiary VNTR profiles do not match, or they may place the suspect at the scene of the crime, with large probability, when they do match. Even a two-locus VNTR profile is very informative because these profiles will rarely match if the samples are obtained from two different individuals, and the evidence is very persuasive when the profiles are obtained from the same individual (Devlin et al. 1992). Nonetheless, the use of VNTR profiles has been criticized by some researchers, who argue that the loci are not likely to be in Hardy-Weinberg equilibrium (HWE) and linkage equilibrium (LE) in the forensic data bases (Lander 1989, 1991*a*, 1991*b*; Cohen 1990; Cohen et al. 1991; Green and Lander 1991; Lewontin and Hartl 1991). They argue that the ethnic categories or populations used for the data base, especially Hispanics and blacks, are actually mixtures of diverse subgroups. Such mixture would induce a Wahlund effect, exhibiting an excess of single-band phenotypes in the data base. Moreover, an excess of single-band phenotypes is frequently found in these data bases. Consequently, the critics conclude that the data bases must be far from equilibrium and that multiplication to obtain genotype probabilities is invalid.

Nevertheless, Chakraborty and Jin (1992) show theoretically that the diversity among subgroups cannot be sufficient to cause the substantial excess of single-band phenotypes observed in the data. In this regard, the excess of single-band phenotypes in the Lifecodes data base can be explained simply by the coalescence of heterozygotes of similar-sized fragments into single-band phenotypes (Devlin et al. 1990, 1991*b*). For the FBI data base, null alleles apparently produce most of the excess of single-band phenotypes (Devlin and Risch 1992; Weir 1992*a*). Moreover, Risch and Devlin (1992) and Weir (1992*a*, 1992*b*) have shown that there is no violation of linkage-equilibrium assumptions in either the Lifecodes data base or the FBI data base—telling findings because population mixture causes disequilibrium both within loci and across loci. Hence the failure to detect violations of independence suggests that the data bases are not composed of populations having vastly different allele frequencies.

In this paper we have two objectives: (1) to examine the allele frequency differences (henceforth simply "differentiation") at VNTR loci between and sometimes within the major ethnic groups of Caucasians, blacks and Hispanics and (2) to determine the relative error in genotype frequency estimates that is created by the mixture of differentiated subpopulations. From the Lifecodes data base, we previously estimated the Caucasian allele frequencies for two VNTR loci, D17S79 and D2S44 (Devlin et al. 1991*a*). Analyzing both the Lifecodes data base and the FBI data base, we test whether the ethnic groups are significantly differentiated at the D17S79 and D2S44 loci. Using the estimated allele frequency distributions, we then examine the effect that mixing these populations has on the error in genotype probability estimates.

Methods

Allele Frequency Estimates

The ethnic classification and the sample sizes for each data base are reported in table 1. The FBI categorizes Hispanics into two groups, those from the southeastern United States (mainly of Cuban ancestory) and those from the southwestern United States (predominantly of mixed Spanish and Amerindian descent [Reed 1974; Chakraborty et al. 1986; Cerda-Flores et al. 1992]). Lifecodes usually does not subdivide the Hispanic group. They have provided us, however, with data subdivided by whether the identified Hispanic's parents were born in Cuba or whether the identified Hispanic resides in California or Texas (A portion of these data were provided by Gene Screen, Inc.). We have combined the Texas and California data to make these data similar in structure to the FBI data (i.e., southeastern and southwestern Hispanics).

Table 1

Ethnic Classifications and Sample Sizes, by Data Base and Locus

Data Base and Ethnic Group	No. of Individuals (no. of alleles) Sampled	
	For D17S79	For D2S44
*Pst*I:		
Black	1,007 (2,014)	1,010 (2,020)
Caucasian	1,399 (2,798)	1,529 (3,058)
Southeastern Hispanic	129 (258)	129 (258)
Southwestern Hispanic	571 (1,142)	471 (942)
*Hae*III:		
Black	549 (1,098)	475 (950)
Caucasian	795 (1,590)	790 (1,580)
Southeastern Hispanic	314 (628)	300 (600)
Southwestern Hispanic	293 (586)	284 (568)

For summaries of the molecular techniques employed by Lifecodes and the FBI, see the papers by Balazs et al. (1989) and Budowle et al. (1991*b*), respectively. Both groups presently use D17S79 and D2S44 for forensic inference. Lifecodes uses the restriction enzyme *Pst*I, while the FBI uses *Hae*III. The sizes of the alleles, for these loci, are larger for *Pst*I than they are for *Hae*III. In addition, for D2S44, *Pst*I reveals flanking-region polymorphism not present in fragments cut with *Hae*III. This polymorphism complicates comparisons, across data bases, for this locus. This is not true for D17S79, for which the alleles differ, across data bases, only by a constant length.

The statistical methods we use to estimate allele frequencies have been explained by Devlin et al. (1991*a*). These methods are outlined below only for a simple VNTR locus, one with no polymorphism in the flanking region and composed of one type of repeat; those interested in more details should consult the original paper. For a simple VNTR locus, such as D17S79 cut with *Hae*III or *Pst*I, we assume that the size of allele, a_r, having r repeats of size ρ and a flanking size u, is given by $a_r = u + r\rho$. However, we do not observe a_r. What we observe is a_r measured with random error ε, or, more precisely, the random variable $X = u + r\rho + \varepsilon$; ε is distributed approximately normally with a mean of zero and variance $\sigma^2 = (ca_r)^2$. The variance of the measurement error increases with increasing fragment size. To estimate the vector of allele frequencies, we first use the method of maximum likelihood to find the most likely vector of allele frequencies for the observed data X. Because we estimate a large number of parameters, which are related, we use an empirical Bayes method to effect a local smoothing. The amount of smoothing depends on σ^2—and thus on fragment size; consequently the maximum-likelihood estimate of the allele frequency distribution, from *Pst*I data, will be smoothed substantially more than that estimated from the *Hae*III data.

Three quantities are needed to estimate allele frequencies for simple VNTRs such as D17S79 and D2S44 cut with *Hae*III: (1) the repeat size, (2) the relationship between fragment size and measurement error, and (3) the probability that a pair of fragments similar but not identical in size will blur together (coalesce) so that they appear as a single band. The consensus repeat size for D17S79 is 38 bp, and for D2S44, it is 31 bp (I. Balazs, personal communication). From repeated measurements of the same allele, the SD of measurement error of *Pst*I-derived alleles is estimated to be $(.00575)\,a_r$; from repeated measurements of the same allele from the *Hae*III-derived alleles, we estimate the SD of the error to be $(.00625)a_r$ (see Budowle et al. 1991*b*, table 2). For *Hae*III-derived alleles smaller than 1.5 kb, we assumed a constant SD of $1.5\ (.00625) = .00938$. The methods of estimating coalescence and the pertinent estimates for the *Pst*I data have been reported by Devlin et al. (1990). In brief, for a particular mean fragment-pair size, Devlin et al. (1990) estimated the probability of coalescence of a pair of fragments as a logistic function of the absolute difference in their lengths. The parameters of this function were obtained by fitting a logistic model to the observed versus expected numbers of heterozygotes found in adjacent intervals of fragment-pair length differences. We performed an identical analysis for the *Hae*III data.

Our analysis of the *Hae*III data for the D17S79 and D2S44 loci indicates that two fragments are unlikely to coalesce when the absolute value of their difference is more than two repeats. The probability of coalescence, p, is approximated by a logistic function of the absolute value of fragment-pair difference X (in kb): $p = 1 - [\exp(-14.154 + 267.44X)]/[1 + \exp(-14.154 + 267.44X)]$. Therefore, for a pair of D2S44 alleles differing by 1, 2, and 3 repeats, $p =$ .997, .081, and 0, respectively.

For VNTRs having variation in the flanking region, such as D2S44 cut with *Pst*I, we also need to know the number and approximate size of the flanking regions. This information can be obtained by also cutting an individual's DNA with a frequently cutting restriction enzyme, such as *Hae*III (Devlin et al. 1991*a*). Even with this information, the allele distribution may not be estimable, as we have discussed previously; for D2S44, it is estimable after certain assumptions elaborated by Devlin et al. (1991*a*) have been made. We made the same assumptions for our new analyses of the *Pst*I data. For all of the ethnic groups, there are two major flanking-region sizes when D2S44 is excised by using *Pst*I. In the black population, several additional uncommon polymorphisms in the flanking region have been detected, while, for Caucasians and Hispanics, fewer rare polymorphisms have been detected (I. Balazs, personal communication).

To test for significant differentiation between a pair of populations or ethnic groups, we use the Hellinger distance between allele frequency distributions. When $\pi_i^{(1)}$ ($\pi_i^{(2)}$) is the frequency of allele i in population 1 (2), the Hellinger distance H is given by

$$H = \sqrt{\sum_{i=1}^{r}\left(\sqrt{\pi_i^{(1)}} - \sqrt{\pi_i^{(2)}}\right)^2}.$$

This statistic has desirable properties for measuring the distance between distributions: the square root of the allele frequencies is a variance-stabilizing transformation (e.g., see Rice 1988), and the behavior of the statistic is similar to that of the familiar χ^2 statistic (e.g., see Rice 1988).

We use the method of bootstrapping to determine whether each pair of distributions is significantly different (Efron and Tibshirani 1991). The bootstrap involves combining the data from both populations and then randomly resampling from this data set to form two new data sets of sample sizes identical to those of the original data sets. The allele frequency distributions are then calculated for the new data sets. The Hellinger distance between these two distributions is then calculated and recorded. This operation is performed 500 times. If the original Hellinger distance exceeds all but 5% of the bootstrap-derived distances, the two distributions are significantly different at the $\alpha = .05$ level.

Hellinger distances, while statistically appealing, are not readily interpretable. Therefore, rather than report these distances, we report a statistic based upon the probability of randomly drawing matching alleles. Specifically, we report the ratio of match probabilities, in which the numerator is the probability of randomly drawing matching alleles from each of the two populations being compared and the denominator is the geometric average of the two within-population match probabilities. The formula for this ratio S is

$$S = \frac{\sum \pi_i^{(1)}\pi_i^{(2)}}{\sqrt{\sum(\pi_i^{(1)})^2 \sum(\pi_i^{(2)})^2}}.$$

This formula is identical to that of Nei's genetic similarity (Nei 1987), which, after log transformation, is frequently used to estimate the time since two populations diverged. Beyond allele matching, however, we make no interpretation of this quantity, as the assumptions underlying the estimation of divergence times are violated by these data. Thus, we use this statistic only as a measure of similarity between two allele frequency distributions.

The large number of alleles at these loci will have a strong effect on similarity estimates; in particular, the estimates will be smaller than they should be (biased downward). To reveal the effect of the large number of alleles on the bias, we perform a series of simulations. For each simulation, we generate two data sets from the same allele frequency distribution. Then, from these data, we calculate the new allele distributions and their similarity. Theoretically the similarity should be 1, because the data are derived from the same distribution. Hence the deviation from 1 is a measure of the bias of the estimate. Bias will be a function of the sample size, and therefore we vary the sample size between 100 and 1,000 alleles sampled. For each sample size and allele frequency distribution, we perform 30 simulations and present the average of the 30 similarity estimates. The data are generated having either measurement error consistent with that of the *Pst*I data or no measurement error at all; that is to say, for the latter, the alleles are completely classifiable by size. The bias will also be a function of gene diversity, and so we use three estimated allele distributions to generate the data: the allele distribution derived from the black and Caucasian data bases for the D17S79 locus (*Pst*I alleles) and that derived from the Caucasian data base for the D2S44 locus (*Hae*III alleles).

Effects of Mixture

To examine the effect of mixture, we considered three cases: (1) mixture based on the estimated allele frequencies of the *Pst*I-excised D17S79 locus for the southeastern and southwestern Hispanic populations; (2) mixture based on the estimated allele frequencies for *Hae*III-excised D17S79 and D2S44 loci for the two Hispanic populations; and (3) mixture based on the estimated allele frequencies of the *Hae*III-excised D17S79 and D2S44 loci for black and Caucasian populations. The latter is included for heuristic purposes, as it represents an extreme case of mixture.

Let $\pi_i^{(k)}$ be the frequency of allele a_i in subpopulation k, and let $\bar{\pi}_i$ be its average frequency over all subpopulations. Then $\bar{\pi}_i = \Sigma_k\, w_k\, \pi_i^{(k)}$, where k indexes the population, and w_k is the proportional representation of subpopulation k. Then, when $\hat{\gamma}_{ij}$ is written as the estimated genotype probability obtained by ignoring the substructure, and when γ_{ij} is written as the genotype frequency, with the substructure taken into account (the true value), these values are calculated as

$$\hat{\gamma}_{ij} = \begin{cases} 2\bar{\pi}_i\bar{\pi}_j, & \text{for } i \neq j \\ \bar{\pi}_i^2, & \text{for } i = j \end{cases}$$

and

$$\gamma_{ij} = \begin{cases} \sum_{k=1}^{m} 2w_k\pi_i^{(k)}\pi_j^{(k)}, & \text{for } i \neq j \\ \sum_{k=1}^{m} w_k(\pi_i^{(k)})^2, & \text{for } i = j \end{cases}$$

where m is the number of subpopulations. Estimated multilocus genotype probabilities are calculated as the product of $\hat{\gamma}_{ij}$ over loci. True multilocus genotype probabilities are calculated as the product of the single-locus genotype probabilities *within* a subpopulation, then multiplying by w_k, and summing over k. Hence HWE and LE are assumed for subpopulations. For each case, we used equal weights ($w_k = .5$), which induces the greatest amount of disequilibrium.

To evaluate the effect of mixture on forensic inference, we calculate the ratio $R = \gamma/\hat{\gamma}$ for all $[r(r+1)]/2$ possible genotypes (where r is the number of alleles observed in the data). If $\hat{\gamma}$ underestimates the true probability of the genotype, then $R > 1$; otherwise, $R \leq 1$. R's much greater than 1 are problematic for forensic inference because the true probability of observing the genotype in the population is much greater than the estimate under HWE. Of course, the magnitude of the R must be evaluated relative to the value of the true and estimated probabilities: it would be problematic if $R = 100$ when $\gamma = 10^{-3}$, but it would be substantially less so when $\gamma = 10^{-12}$, as both the true and estimated probabilities are extremely small. Consequently the R's will be reported relative to the true frequencies. Values close to 1 or <1 are of no concern; in the latter case, the estimate is advantageous to the suspect.

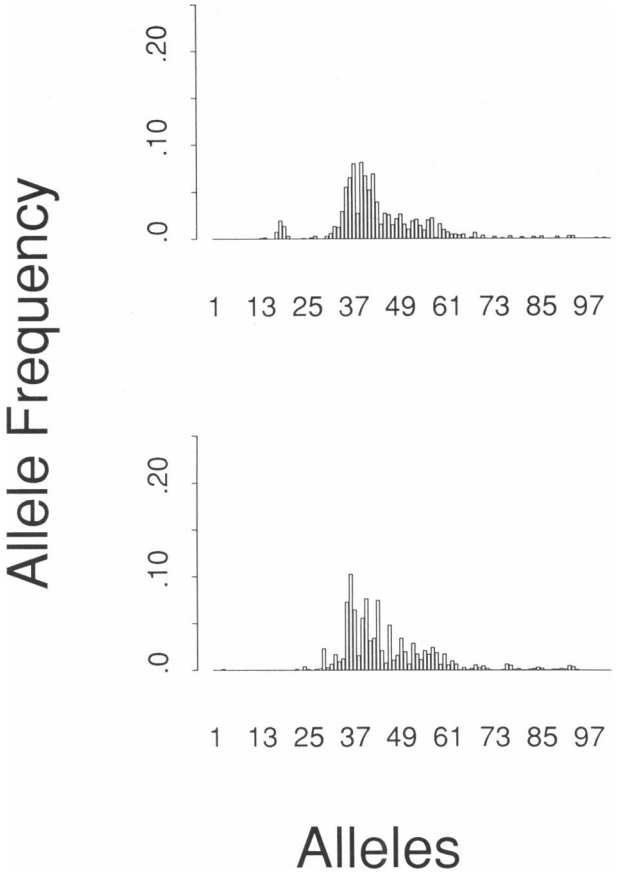


Figure 1 Allele frequency estimates of the D17S79 locus, from *Pst*I-derived data (*top*) and *Hae*III-derived data (*bottom*), from the black population. Alleles are ordered by size, from smallest to largest. The numbering of alleles does not convey the number of repeats.

Results

Allele Frequency Estimates

Figures 1–4 present the D17S79 allele frequency estimates for the black, Caucasian, southeastern Hispanic, and southwestern Hispanic populations, respectively, where the uppermost figure is the estimate from the *Pst*I data and the lowermost figure is the estimate from the *Hae*III data. We note that the leading allele of the *Hae*III distribution is actually a dummy class of small alleles that cannot be measured accurately. There is little difference between the estimates across the data bases, as one would expect, since they are samples of the same population. However, as noted previously, the enzymes used to obtain the original data are different, which impacts on the accuracy of the allele frequency estimates. In particular, for a given sample size, estimates from the *Hae*III data will be more accurate. This difference in accuracy is reflected by the smoother nature of the *Pst*I-derived allele frequency distributions compared with the *Hae*III-derived allele frequency distributions.

The similarities S between pairs of distributions are reported in table 2. The similarities are most reliable for the *Hae*III data (because of small measurement error) and for the black-Caucasian comparison of the *Pst*I data (because of large sample sizes). In any discussion of these similarities, we are referring to these relatively accurate values. The most similar populations are the Caucasian and southeastern Hispanic, and the least similar populations are the black and Caucasian. There are significant differences between all of the ethnic groups for these distributions, even for the Cauca-

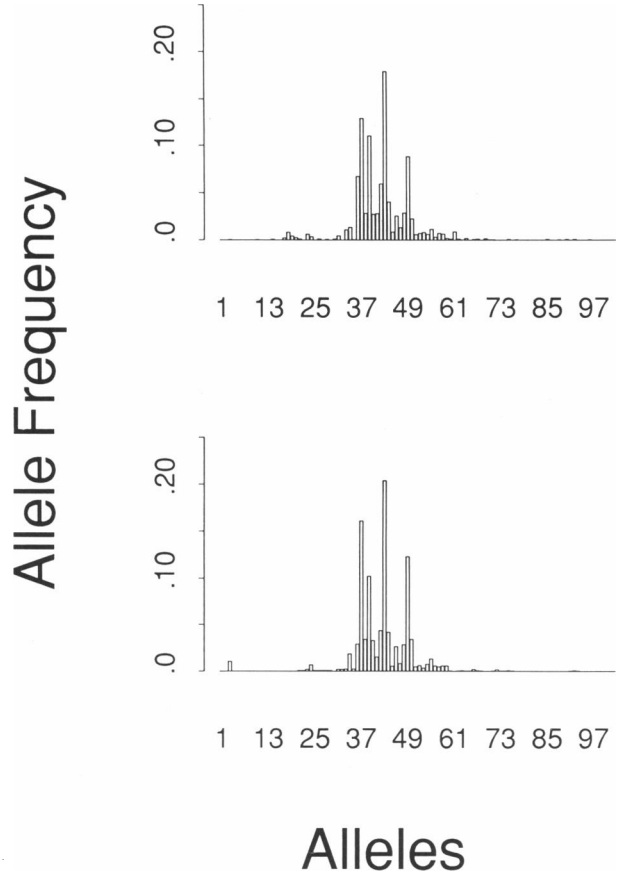


Figure 2 Allele frequency estimates of the D17S79 locus, from *Pst*I-derived data (*top*) and *Hae*III-derived data (*bottom*), from the Caucasian population. Alleles are ordered by size, from smallest to largest. The numbering of alleles does not convey the number of repeats.

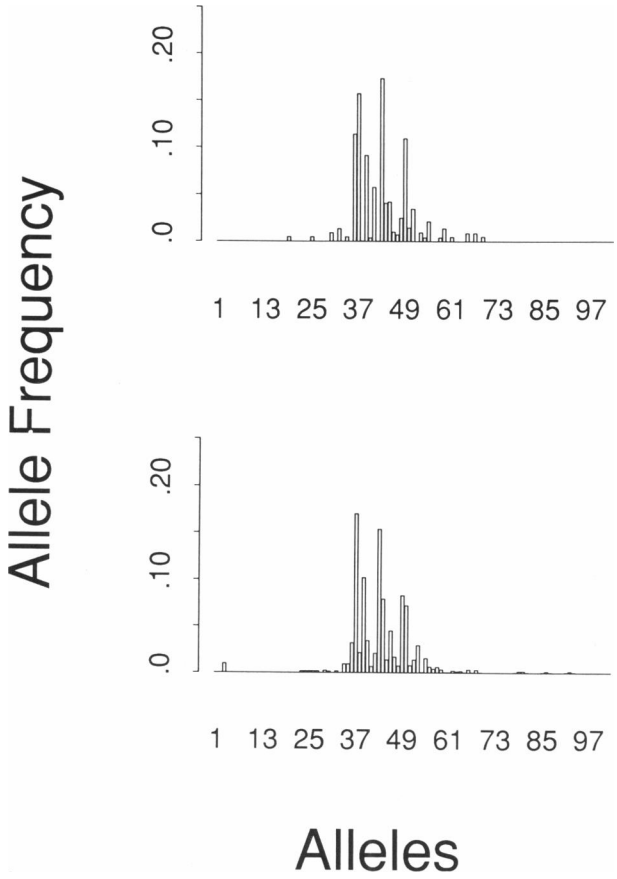
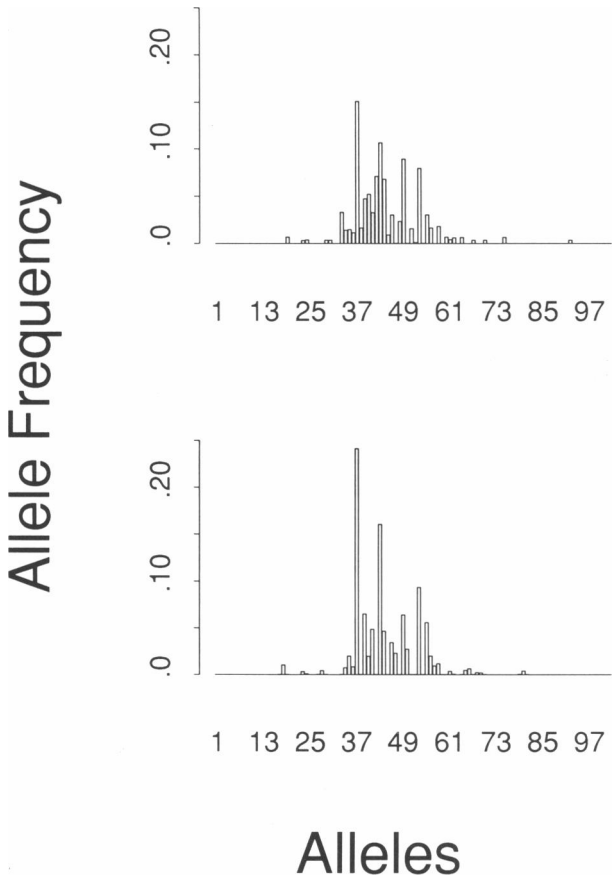


Figure 3 Allele frequency estimates of the D17S79 locus, from *Pst*I-derived data (*top*) and *Hae*III-derived data (*bottom*), from the southeastern Hispanic population. Alleles are ordered by size, from smallest to largest. The numbering of alleles does not convey the number of repeats.

sian and southeastern Hispanic populations ($P < .05$). Even though there are significant differences between each pair of populations, the allele distributions are similar in that most alleles are infrequent. In addition, relative to the general infrequency of alleles, alleles that occur frequently in one population occur frequently in another, and likewise for rare alleles. The gene diversity (or heterozygosity), computed as $1 - \Sigma_i \pi_i^2$, was .958 (.954), .906 (.896), .902 (.912), and .905 (.908) for the *Pst*I (*Hae*III) data from the black, Caucasian, southeastern Hispanic, and southwestern Hispanic populations, respectively. One minus the gene diversity equals the probability of drawing matching alleles, by chance, from the population. Thus these gene-diversity estimates allow the reader to gain more insight into the similarity values in table 2. For instance, the between-population match probability for *Pst*I (*Hae*III) black and Caucasian data is .046 (.048), while the geometric average of the within-population match probability is .060 (.069). Clearly the match probabilities are small, regardless of whether alleles are selected from the same or different populations, although the ratio of the match probabilities is substantially different than 1.

One particularly striking difference between the allele frequency estimates for the *Pst*I and the *Hae*III data bases can be seen most clearly in figure 1, the distributions for the black population samples. There are numerous small alleles in the sample when *Pst*I is used that are not present in the sample when *Hae*III is used. The absence of small alleles in the *Hae*III data is evidence of null alleles when that enzyme is used. In



fact, we estimate the frequency of null alleles to be about 4.4%, by counting (Devlin and Risch 1992). The presence of null alleles leads to a large excess of single-band phenotypes for blacks at D17S79 (Devlin and Risch 1992), as Budowle et al. (1991*b*) predicted. This phenomenon is less important for the Caucasian and Hispanic populations, which have fewer alleles with a small number of repeats (figs. 2–4). The presence of null alleles affects the similarity of distributions within and between ethnic groups. For instance, the *Pst*I and *Hae*III data for Caucasians yield allele frequency distributions having similarity of .98, while the same comparison for blacks yields a distance of .90. When adjustment is made for the presence of the null alleles, the similarity increases to .92. Hence the similarities reported in table 2 are altered by the presence of null alleles, especially for comparisons with the black population.

Figures 5–8 present the D2S44 allele frequency estimates for the black, Caucasian, southeastern His-

Figure 4 Allele frequency estimates of the D17S79 locus from *Pst*I-derived data (*top*) and *Hae*III-derived data (*bottom*) from the southwestern Hispanic population. Alleles are ordered by size, from smallest to largest. The numbering of alleles does not convey the number of repeats.

Table 2

Ratios of Match Probabilities, by Restriction Enzyme, Ethnic Group, and Locus

	Ratio of Match Probability for					
	*Pst*I			*Hae*III		
	Caucasian	Southeastern Hispanic	Southwestern Hispanic	Caucasian	Southeastern Hispanic	Southwestern Hispanic
For D17S79 locus:						
Black	.77	.69	.70	.70	.72	.70
Caucasian		.93	.87		.95	.87
Southeastern Hispanic			.89			.90
For D2S44 locus:						
Black	.79	.54	.67	.74	.74	.76
Caucasian		.70	.88		.89	.78
Southeastern Hispanic			.60			.81
Over both loci:						
Black	.78	.62	.69	.72	.73	.73
Caucasian		.82	.88		.92	.83
Southeastern Hispanic			.72			.86

Note.—The numerator of the ratio is the probability of randomly drawing a pair of matching alleles when each member of the pair is drawn from a different population. The denominator is the geometric average match probability when the pair of alleles are drawn from the same population. In the text, we call the ratios "similarities," as their formulation is identical to Nei's approach to calculating genetic similarities.

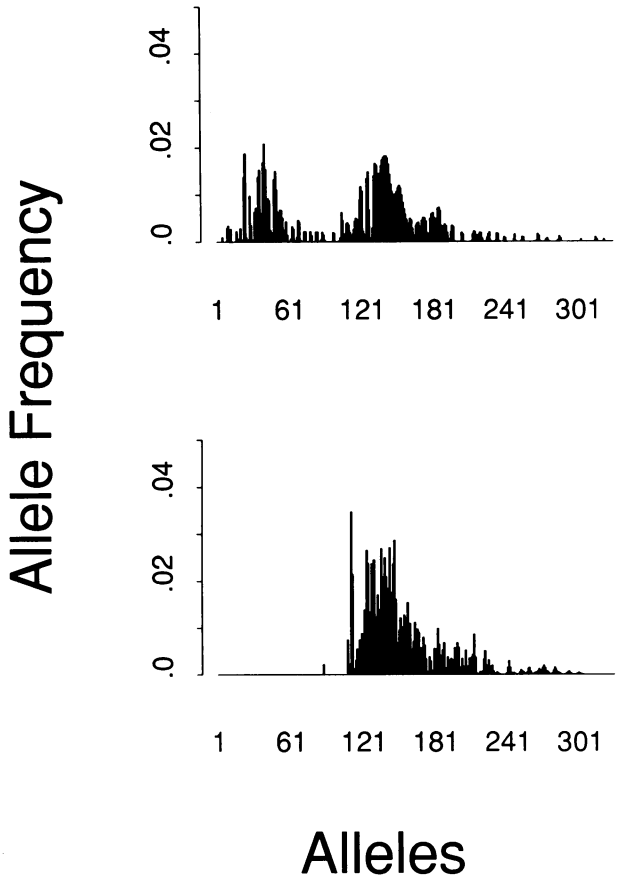


Figure 5 Allele frequency estimates of the D2S44 locus, from *Pst*I-derived data (*top*) and *Hae*III-derived data (*bottom*), from the black population. Alleles are ordered by size, from smallest to largest. *Pst*I reveals polymorphism in the flanking region for this locus; to compare the distributions, match the right three-quarters of the top graph's distribution to the bottom graph's distribution. The numbering of alleles does not convey the number of repeats.

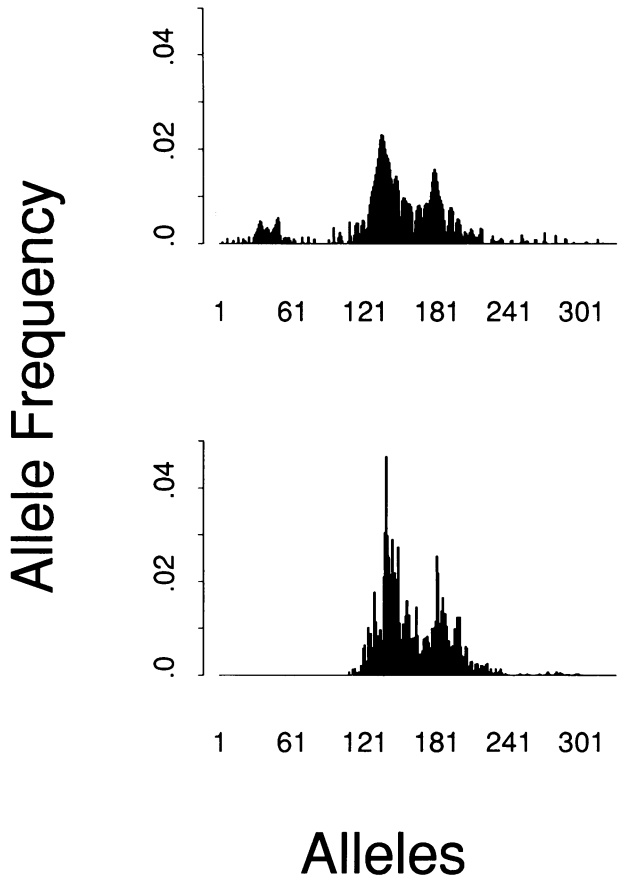


Figure 6 Allele frequency estimates of the D2S44 locus, from *Pst*I-derived data (*top*) and *Hae*III-derived data (*bottom*), from the Caucasian population. Alleles are ordered by size, from smallest to largest. *Pst*I reveals polymorphism in the flanking region for this locus; to compare the distributions, match the right three-quarters of the top graph's distribution to the bottom graph's distribution. The numbering of alleles does not convey the number of repeats.

panic, and southwestern Hispanic populations, respectively, in the format described above. For this locus, *Pst*I reveals polymorphism in the flanking region while *Hae*III does not. Therefore the distributions for different restriction enzymes are not strictly comparable, although a qualitative comparison can be made by contrasting the rightmost three-quarters of the *Pst*I-derived allele distributions with the *Hae*III-derived allele distributions. We have added space to the left side of the *Hae*III distribution, to facilitate visual comparisons. Qualitatively the distributions are very similar for each ethnic group, even for the southeastern Hispanic population, for which the number of allele frequencies to estimate from the *Pst*I data is quite large relative to the sample size. Again, the ratio of match probabilities (table 2) indicates that the allele frequency distributions for Caucasians and southeastern Hispanics are most similar, while those for blacks and Caucasians are the least similar. There are significant differences ($P \leqslant .05$) between the ethnic groups when the *Hae*III-derived allele distributions are compared, and likewise for the black-Caucasian comparison when the *Pst*I data are used. We did not evaluate whether the Hispanic allele frequency distributions from the *Pst*I data are significantly differentiated, because the sample sizes are too small to obtain adequate power for this test. (We are estimating 329 parameters.) Again, relative to the general infrequency of all alleles, high-frequency alleles in one population have high frequency in others, and likewise for rare alleles.

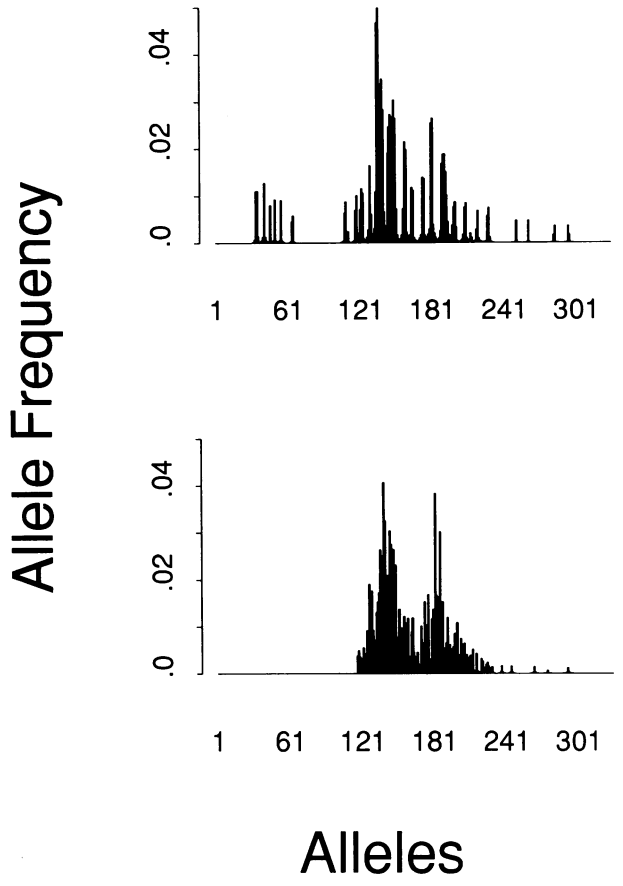


Figure 7 Allele frequency estimates of the D2S44 locus, from *Pst*I-derived data (*top*) and *Hae*III-derived data (*bottom*), from the southeastern Hispanic population. Alleles are ordered by size, from smallest to largest. *Pst*I reveals polymorphism in the flanking region for this locus; to compare the distributions, match the right three-quarters of the top graph's distribution to the bottom graph's distribution. The numbering of alleles does not convey the number of repeats.

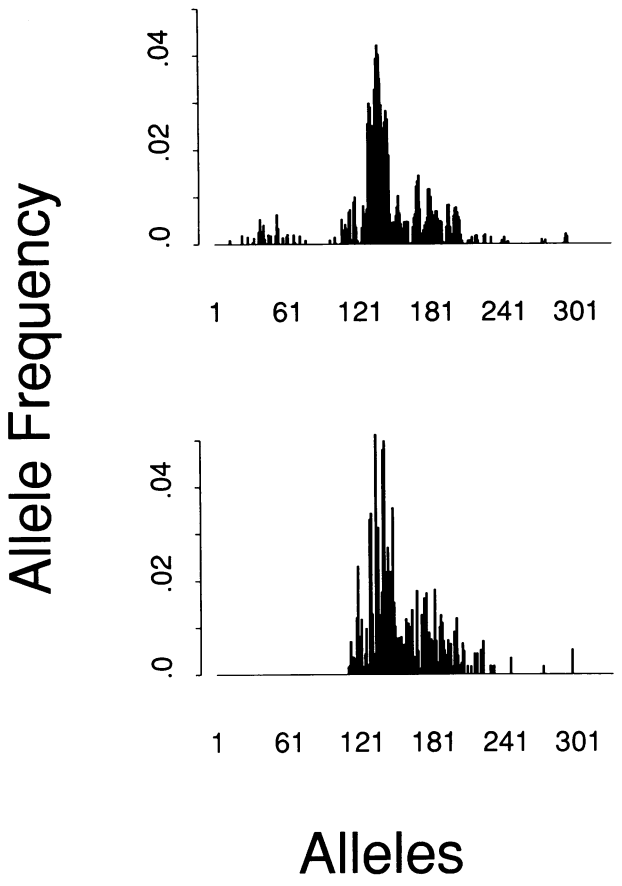


Figure 8 Allele frequency estimates of the D2S44 locus, from *Pst*I-derived data (*top*) and *Hae*III-derived data (*bottom*), from the southwestern Hispanic population. Alleles are ordered by size, from smallest to largest. *Pst*I reveals polymorphism in the flanking region for this locus; to compare the distributions, match the right three-quarters of the graph's distribution to the bottom graph's distribution. The numbering of alleles does not convey the number of repeats.

The gene diversity was .990 (.985), .990 (.985), .978 (.983), and .981 (.980) for *Pst*I (*Hae*III) data from black, Caucasian, southeastern Hispanic, and southwestern Hispanic populations, respectively. Hence the probability of drawing matching alleles is <.025 in any of the populations assessed. Again it is interesting to examine the differences between match probabilities when alleles are drawn from the same population versus when they are drawn from different populations. When comparing blacks and Caucasians, the similarity of the distributions for *Pst*I (*Hae*III) was .79 (.74; table 2), suggesting substantial differences between the populations. This ratio is composed of a between-population match probability of .008 (.011) and an average within-population match probability of .010 (.015).

The gene diversity for D2S44 is substantially greater than that for D17S79, whether the alleles at these loci are excised with *Pst*I or *Hae*III. On the other hand, variance (among populations) in gene diversity is greater for D17S79. In terms of similarity between populations, Caucasians and southeastern Hispanics are most similar, and Caucasians and blacks are the least similar (table 2).

The differences between ethnic groups is undoubtedly exaggerated by sampling variance. The effect of sampling variance can be substantial because, even though the number of alleles sampled is usually large

(table 1), the number of possible alleles is also quite large. From the *Pst*I data from blacks, we estimated that, for the D17S79 locus, there are 56 alleles having a frequency greater than zero, versus 52 alleles for the *Hae*III data; 53 versus 47 for Caucasians; 23 versus 33 for southeastern Hispanics; and 42 versus 34 for southwestern Hispanics (allele numbers are derived from the maximum-likelihood estimates of the allele frequencies). For the D2S44 locus, the estimates are much larger. From the *Pst*I data from blacks, we estimate that there are 167 alleles having a frequency greater than zero, versus 110 for the *Hae*III data; 172 versus 113 for Caucasians; 48 versus 96 for southeastern Hispanics; and 109 versus 93 for southwestern Hispanics (again, allele numbers are derived from the maximum-likelihood estimates of the allele frequencies). The effect of *Pst*I, which reveals D2S44 flanking-region polymorphism, is obvious in the increase in the number of alleles for the black and Caucasian samples.

The smaller number of alleles observed in the *Pst*I data from Hispanics versus the number observed when *Hae*III is used can be accounted for, in part, by the larger fragments obtained when *Pst*I is used. The larger fragments limit resolution, particularly when the sample sizes are relatively small. In addition, for the *Pst*I D2S44 data, we are estimating the frequency of 329 alleles. Therefore, when measurement error and small sample sizes (256 alleles sampled for southeastern Hispanics) are taken into consideration, large variance in the allele distribution estimate is predictable, with numerous allele frequencies overestimated and many others estimated to be zero (Devlin et al. 1991*a*).

The effect of sample size on the number of alleles observed is also apparent. As we would expect from sampling theory, the number of alleles observed is correlated with sample size. For D17S79, the correlation is .86 (P = .003); for D2S44, the partial correlation, when the restriction enzyme is taken into account, is .94 (P = .0004). These results indicate that the populations are likely to contain alleles that have not been observed in the samples, though, for the black and Caucasian samples, most alleles of any notable frequency have been observed. For instance, if we took from the black population another sample of *Pst*I-excised alleles that was the same size as those in table 1, the Turing estimates of the number of new alleles that would be observed could be 2 and 11 for D17S79 and D2S44, respectively. (The Turing estimate is simply the number of singletons observed in the original sample [Good 1953]). The presence of rare unobserved alleles induces a minor bias in the allele frequency estimates, though, for practical applications, the bias is easily adjusted (Smouse and Chakraborty 1986).

Predictably, the bias in similarity decreases with increasing sample size (fig. 9). More interesting, however, is that it is also a function of the number of alleles at the locus and of the gene diversity, as well as of measurement error. For instance, generating data from the estimated D17S79 distributions of the black and Caucasian populations (figs. 1 and 2, respectively) leads to very different similarity estimates (middle and

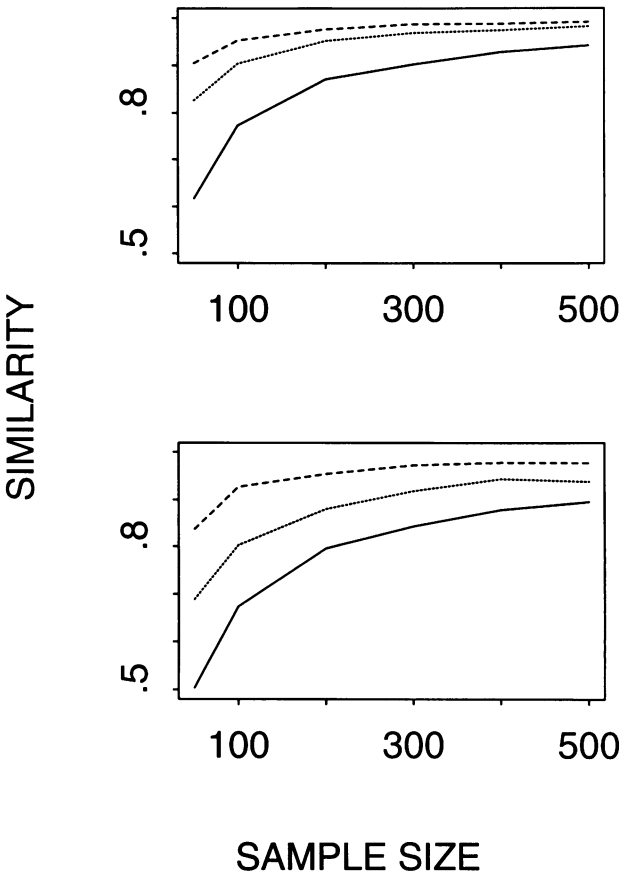


Figure 9 Simulation results estimating the bias of the similarity statistic, as a function of sample size (number of alleles sampled, represented on the horizontal axis), gene diversity (different lines on the figures), and measurement error (top graph—no error; and bottom graph—measurement error of the *Pst*I data). For simulation methods, see the text. The theoretical similarity is 1.0; therefore the deviation from 1.0 is an estimate of the bias. The lines on the figure correspond to allele frequency distributions having gene diversities of .91, .95, and .99, in descending order.

top lines, respectively, of fig. 9). Recall that the number of alleles at the D17S79 locus is essentially the same for each population but that the allele diversity is larger for the black population (.96 vs. .91). Greater allele diversity results in substantially greater bias in the similarity statistic because of the greater variability in the estimated allele frequency distributions. When the data are generated using the allele distribution of the bottom graph of figure 6, which has an even greater number of alleles (113) and allele diversity (.99), far greater variability in the estimated allele frequency distributions is induced, and thus similarity is severely biased. The similarity is also severely biased by measurement error (compare the top and bottom graphs of fig. 9).

Effects of Mixture

The effect of mixture is assessed by assuming that the estimated allele frequency distributions obtained in the preceding, Allele Frequency Estimates subsection are the true distributions and by mixing these theoretical distributions. Three cases are examined: (1) mixture based on the estimated allele frequencies of the *Pst*I-excised D17S79 alleles from the two Hispanic populations, (2) mixture based on the estimated allele frequencies for *Hae*III-excised D17S79 and D2S44 alleles from the two Hispanic populations, and (3) mixture based on the estimated allele frequencies of the *Hae*III-excised D17S79 and D2S44 alleles from the black and Caucasian populations. The effect of mixture is summarized by the ratio of the true genotype frequency divided by the estimated genotype frequency under HWE, as described in Methods.

For case 1, mixture has little impact on the total frequency of homozygotes: the expected frequency, under HWE, is 9.8% whereas the true value is 10.9%. None of the ratios R for case 1 is >2 (table 3). The largest R, ≈2.0, is achieved when the true genotype probability is .062. This translates into a true frequency of 6/100, versus the estimate of 3/100. Thus, for this single-locus case, mixture results in, at most, a twofold deviation from the true values. Moreover, the small sample sizes for the Hispanic populations exaggerate the differences among subpopulations (fig. 9).

For case 2, mixture also had little impact on the total frequency of homozygotes, for either the D17S79 locus (9.03% vs. 8.58%) or the D2S44 locus (1.96% vs. 1.72%). For this two-locus example, the largest R, 8.0, occurs when the true genotype probability is 4.8×10^{-11} (table 3). Again, the majority (93.30%) of the R values are <2, and larger R values cooccur with very small probabilities (table 3).

Case 3 represents mixture of ethnic groups, an extreme form of mixture because the variation among subpopulations of an ethnic group should be less than the variation among ethnic groups. Even with this extreme mixture, the increase, in total homozygotes,

Table 3

Effects of Mixture

Ratio	Minimum($-\log_{10}\gamma$)[a]	Mean($-\log_{10}\gamma$)[a]	No of Genotypes	Total Probability
Southeastern Hispanic and southwestern Hispanic mixure for D17S79 cut with *Pst*I:				
$0 < R < 2$	1.21	4.20	1,139	1.0
Southeastern Hispanic and southwestern Hispanic mixure for D17S79 and D2S44 cut with *Hae*III:				
$0 < R < 2$	3.69	8.07	5,434,486	.933
$2 < R < 4$	4.14	8.41	1,040,643	.065
$4 < R < 6$	4.96	8.86	198,881	.002
$6 < R < 8$	6.34	9.50	57,903	.000
Black and Caucasian mixure for D17S79 and D2S44 cut with *Hae*III:				
$0 < R < 2$	4.05	8.32	11,933,009	.811
$2 < R < 4$	4.09	8.47	2,868,745	.171
$4 < R < 6$	4.41	8.80	689,629	.016
$6 < R < 8$	5.53	9.39	233,642	.001

[a] Values are transformed from the true genotype probabilities (γ_{ij} in the text) and are reported by intervals of the ratio statistic R, which is the ratio of the true to the estimated genotype probability. The transformation was $-\log_{10}(\gamma_{ij})$. The minimum values correspond to the largest genotype probability.

over that expected under HWE is small: 7.58% versus 6.26% for the D17S79 locus and 1.57% versus 1.34% for the D2S44 locus. The maximum R is again 8.0 and occurs when the true probability is 4.3×10^{-12}. As we would expect, the variance of the R distribution is greater for this extreme case of mixture, i.e., mixing two ethnic groups, than it is for case 2, i.e., mixing subpopulations (table 3). Nevertheless, no R is very large. Most R's (81.10%), however, are <2, and larger R's cooccur, as before, with small probabilities.

Discussion

If subpopulations of an ethnic group have drastically different allele frequencies, as has been argued by some authors (Lander 1989, 1991*b;* Cohen 1990; Cohen et al. 1991; Green and Lander 1991), approximate genotype probabilities calculated under HWE and LE might be too error prone to be useful for forensic inference. For instance, Lander (1991*a*) claims that "genetic drift can have much greater proportional effects on allele frequencies" for VNTR loci, which have large mutation rates. This statement might be true if an infinite-alleles-type model and no gene flow are assumed (for the infinite-alleles model, each mutation is assumed to yield a novel allele); such a model, however, is quite unrealistic for VNTR loci, where alleles of 15 repeats or 27 repeats could both mutate to an allele of 21 repeats (Jeffreys et al. 1988). Moreover, if these loci have unequal mutation rates among alleles, this process affects the rate of subpopulation differentiation (Cockerham 1984). For instance, relatively large alleles may be more likely to mutate to smaller allele sizes than to even larger allele sizes, while smaller alleles may be more likely to mutate to intermediate allele sizes than to even smaller alleles. In this case, differentiation of populations is substantially decreased over that expected when alleles mutate at equal rates (Cockerham 1984). Finally, it ignores gene flow, which homogenizes populations even when opposed by selection (e.g., see Slatkin 1973). Gene flow is common among human populations (Nei and Roychoudhury 1982), making extreme divergence almost impossible.

The empirical results of the Allele Frequency Estimates subsection also belie the notion that ethnic groups are highly differentiated. Although almost all alleles are infrequent, relatively common alleles in the black population are relatively common in the Caucasian population, and they are also relatively common in the Hispanic populations; likewise, rare alleles in the black population are rare in the Caucasian population, and they are also rare in the Hispanic populations. The largest differences in allele distributions are between those of the black and Caucasian populations, and even those populations have remarkably similar distributions. For instance, for D2S44 alleles excised with *Hae*III, the average probability of randomly drawing matching alleles from within these populations is .015, while the probability of randomly drawing matching alleles when one is drawn from each population is .011. Because ethnic groups do not differ very much, subgroups of the same ethnic group cannot be very different, as there must be substantially greater gene flow among the subgroups. Nei and Roychoudhury (1982) and many other researchers have found similar results for traditional genetic markers.

Our simulations (fig. 9) show that the allele distributions of the ethnic groups are actually more similar than they appear to be in table 2, because the similarity statistic is biased downward (toward dissimilarity)—and, in some cases, substantially so. The bias increases with increasing allele diversity and decreases with increasing sample size. This pattern of bias makes interpretation of the similarity statistics complicated. Nevertheless, some very interesting features emerge from these statistics, particularly when they are considered in light of the bias results. For instance, when blacks and Caucasians are compared with respect to either *Pst*I alleles or *Hae*III alleles, their similarity is greater for the D2S44 locus than it is for the D17S79 locus. Because the sample sizes are essentially the same for both loci (or less for D2S44), the bias should be greatest for the D2S44 locus, which has a far greater number of alleles and greater allele diversity. Hence we can conclude that the true D2S44 allele distributions for the black and Caucasian populations are even more similar than they appear to be in table 2. This result might be counterintuitive to some researchers, who expect, on the basis of the infinite-alleles model, that loci having larger allele diversity will show the greatest dissimilarity among and within ethnic groups (Lander 1991*b*; Lewontin and Hartl 1991). The result is less surprising under other models of allele generation.

Another contrast of substantial interest for forensics is the comparison of *Pst*I allele distributions and *Hae*III allele distributions for the same ethnic group. This is especially true for the Caucasian data bases, which are derived from distinctly different sources: mothers and putative fathers from paternity cases, for the *Pst*I data, and law enforcement officers, for the

*Hae*III data. The *Hae*III data base (from the FBI) has been strongly criticized as not being a representative sample of the Caucasian ethnic group (Lander 1991*b*; Lewontin and Hartl 1991). This criticism is difficult to rectify with our results, as the distributions from distinctly different sources are strikingly similar (.98). Performing simulations to determine the expected bias in the similarity statistic, we estimate the bias to be .01. (Simulations were performed as described previously, with sample sizes for Caucasians given in table 1 and with measurement error typical of the particular data base; the simulations did not take into account null alleles.) Considering the expected bias, the similarity value provides compelling evidence that the data bases used to derive the allele distributions could be samples from the same population. Thus it appears that the FBI's data base is representative of the Caucasian population. Results for the Hispanic ethnic groups, for the D17S79 locus, again suggest that the data bases are derived from nearly identical populations, when both the similarities and expected biases are considered (for approximate biases, see fig. 9). For the allele frequency distributions derived from the black population, there may be small differences in the populations from which the data bases are derived, as the expected bias is .05.

Lewontin (1972) analyzed blood group and protein loci, partitioning the allele frequency variation into three components: between individuals within the same population, between populations, and between ethnic groups. His results showed that most of the variation is attributable to variation between individuals within populations (85.4%), with substantially less being attributable to the other two sources (8.3% for variation between populations and 6.3% for variation between ethnic groups). Treating the southeastern Hispanic and southwestern Hispanic samples as populations and using the *Hae*III-derived data, we find that almost all of the genetic variation in the Hispanics is attributable to differences between individuals and that very little is attributable to differences between the populations: 97.4% versus 2.6%, respectively, for D17S79 and 97.1% versus 2.9%, respectively, for D2S44. Over the black, Caucasian, and Hispanic ethnic groups, the between-group variation is larger for D17S79 (10.4%) than for D2S44 (3.3%), as we would expect given the allele frequency distributions (figs. 1–8). From the genetic data he examined, Lewontin (1972) concluded that about 6.3% of the diversity is attributable to variation among ethnic groups; for the VNTR data presented here, the estimate would be 6.9%.

Because Lewontin's (1972) results show that most human diversity is attributable to differences among individuals within a population, we find it curious that Lewontin and Hartl (1991) emphasize that there is, on average, more variation between populations than between ethnic groups. Lander (1991*b*) echos these sentiments when he argues that subpopulations are more important than ethnic categories. Empirical data on genetic distances, when traditional loci from human populations are used, do not support these claims (Nei and Roychoudhury 1982; Smouse et al. 1982; Cerda-Flores et al. 1992), and Lewontin's results can be interpreted as artifacts of inappropriate analytic methods (Smouse et al. 1982; Devlin et al., submitted). The claim is also unsupported by the simple analysis that we performed; clearly, the variance attributable to Hispanic subpopulations is far less than that attributable to ethnic groups, even when Lewontin's methods are used. Moreover, the Hispanic allele distributions are more similar to each other than they are to those of other ethnic groups, except the Caucasians and southeastern Hispanics; recall that the Hispanic data bases have the smallest sample sizes, which presumably exaggerates differences in the allele distributions (fig. 9). Finally, from the forensic perspective, the probability of matching genotypes is very small, regardless of which population is used to calculate the probability, suggesting that differences among subpopulations are of questionable importance.

Of course, there is some differentiation of allele frequencies among the ethnic groups. Why does this differentiation have so little impact on genotype probabilities? Because comparing allele frequency distributions is misleading, causing an exaggerated impression of the amount of error incurred. If one falsely assumes that a mixed population is a single homogeneous population, then the error incurred is a function of the variances and covariances of allele frequencies (e.g., see Li 1969)—and is not a function of the ratio of the extremes of the frequencies, as has been claimed by others (Lewontin and Hartl 1991). Hence mixture of subpopulations results in only modest error, unless the variation about the mean is large. In the United States, variation among ethnic groups appears to be relatively small and, therefore, insufficient to cause substantial deviations of the estimated genotype probabilities relative to the true probabilities.

A point that has been raised repeatedly regarding

errors due to mixture is that the errors *might* compound over loci when multilocus genotype probabilities are considered. This claim is certainly true, but it ignores the fact that the multilocus genotype probability is also a product. Single-locus genotype probabilities are orders of magnitude less than one, whereas the error induced by mixing is not orders of magnitude greater than 1. Therefore, as the number of loci constituting the multilocus genotype increases, the probability of observing each genotype decreases much faster than the error rate increases, as is obvious from the results shown in table 3.

Because the courts often do not use allele frequencies per se for forensic inference, but rather use alleles grouped into bins (e.g., see Balazs et al. 1989; Budowle et al. 1991*b*), the effect of mixture for the binning methods will be much less than the values we report here. This is because the binning methods act to "smooth" the allele frequency distributions, reducing the differences among populations. The effect of binning groups of alleles carries over to genotypes, where many genotypes are combined into what we call a "binotype," or a two-bin phenotype. The binotype distribution is substantially smoothed relative to the genotype distribution. Hence it is not surprising that Weir (1992*a*, 1992*b*) has shown there is a very strong correlation between multilocus binotype probabilities for all pairs of ethnic groups, both when the fixed-bin method (used by the FBI) is used and when the floating-bin method (used by Lifecodes) is used. In fact, his results suggest that it would make little difference, to forensic inference, which ethnic group was used when binotype probabilities are calculated. His results complement ours, being a reflection of the fact that allele distributions are not very different for different ethnic groups. Moreover, his results reiterate our point above: if ethnic groups differ very little in binotype probabilities, then subpopulations should differ even less, making mixture an unimportant influence on forensic inference.

Acknowledgments

Tables or files of allele frequency estimates for each restriction enzyme, locus, and ethnic group are available on written request. Tables will be sent by surface mail, and files will be sent via electronic mail. We thank Lifecodes and the FBI for supplying the data, and we thank Ivan Balazs, Bruce Budowle, John Hartmann, Kathryn Roeder, and Peter Smouse for discussion and/or comments on a previous draft of the manuscript. We also thank Ivan Balazs for allowing us to use his unpublished data. This work was supported by NIH grants HG00648 and CA45052 to N.R.

References

Balazs I, Baird M, Clyne M, Meade E (1989) Human population genetic studies of five hypervariable DNA loci. Am J Hum Genet 44:182–190

Budowle B, Chakraborty R, Giusti AM, Eisenberg AJ, Allen RC (1991*a*) Analysis of the VNTR locus D1S80 by the PCR followed by high-resolution PAGE. Am J Hum Genet 48:137–144

Budowle B, Giusti AM, Waye JS, Baechtel FS, Fourney RM, Adams DE, Presley LA, et al (1991*b*) 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

Cerda-Flores RM, Kshatriya GK, Bertin TK, Hewett-Emmett D, Hannis CL, Chakraborty R (1992) Gene diversity and estimation of genetic admixture among Mexican-Americans of Starr County, Texas. Ann Hum Biol 19:347–360

Chakraborty R, Ferrel RE, Stern MP, Haffner SM, Hazuda HP, Rosenthall M (1986) Relationship of prevalence of non-insulin dependent diabetes mellitus with Amerindian admixture in Mexican-Americans of San Antonio, Texas. Genet Epidemiol 3:435–454

Chakraborty R, Jin L (1992) Heterozygote deficiency, population substructure, and their implications for DNA fingerprinting. Hum Genet 88:267–272

Cockerham CC (1984) Drift and mutation with a finite number of allelic states. Proc Natl Acad Sci USA 81:530–534

Cohen JE (1990) DNA fingerprinting for forensic identification: potential effects on data interpretation of subpopulation heterogeneity and band number variability. Am J Hum Genet 46:358–368

Cohen JE, Lynch M, Taylor CE (1991) Forensic DNA tests and Hardy-Weinberg equilibrium. Science 243:1037–1038

Devlin B, Risch N (1992) A note on Hardy-Weinberg equilibrium of VNTR data by using the Federal Bureau of Investigation's fixed-bin method. Am J Hum Genet 51: 549–553

Devlin B, Risch N, Roeder K (1990) No excess of homozygosity at DNA fingerprint loci. Science 249:1416–1420

——— (1991*a*) Estimation of allele frequencies for VNTR loci. Am J Hum Genet 48:662–676

——— (1991*b*) Forensic DNA tests and Hardy-Weinberg equilibrium. Science 243:1039–1041

———. Forensic inference from DNA fingerprints (1992) J Am Stat Assoc 87:337–350

———. Statistical evaluation of DNA fingerprinting: a critique of the NRC's report (submitted)

Efron B, Tibshirani R (1991) Statistical data analysis in the computer age. Science 253:390–394

Good IJ (1953) The population frequencies of species and the estimation of population parameters. Biometrika 43: 45–63

Green P, Lander E (1991) Forensic DNA tests and Hardy-Weinberg equilibrium. Science 243:1038–1039

Jeffries AJ, Royle NJ, Wilson V, Wong Z (1988) Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. Nature 332: 278–281

Lander E (1989) DNA fingerprinting on trial. Nature 339: 501–505

——— (1991*a*) Lander reply. Am J Hum Genet 49:899–903

——— (1991*b*) Research on DNA typing catching up with courtroom application. Am J Hum Genet 48:819–823

Lewontin RC (1972) The apportionment of human diversity. Evol Biol 6:381–398

Lewontin RC, Hartl D (1991) Population genetics in forensic DNA typing. Science 254:1745–1750

Li CC (1969) Population subdivision with respect to multiple alleles. Ann Hum Genet 33:23–29

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nei M, Roychoudhury AK (1982) Genetic relationship and evolution of human races. Evol Biol 14:1–59

Reed TE (1974) Ethnic classification of Mexican-Americans. Science 185:283

Rice JA (1988) Mathematical statistics and data analysis. Wadsworth & Brook/Cole, Pacific Grove, CA

Risch N, Devlin B (1992) On the probability of matching DNA fingerprints. Science 255:717–720

Slatkin M (1973) Gene flow and selection in a cline. Genetics 75:733–756

Smouse PE, Chakraborty R (1986) The use of restriction fragment length polymorphisms in paternity analysis. Am J Hum Genet 38:918–939

Smouse PE, Speilman RS, Park MH (1982) Multiple-locus allocation of individuals to groups as a function of the genetic variation within and the differences among populations. Am Nat 119:445–463

Weir B (1992*a*) Independence of VNTR alleles defined as fixed bins. Genetics 130:873–887

——— (1992*b*) Independence of VNTR alleles defined as floating bins. Am J Hum Genet 51 (in press)