Am. J. Hum. Genet. 52:60–70, 1993

Evaluation of Standard Error and Confidence Interval of Estimated Multilocus Genotype Probabilities, and Their Implications in DNA Forensics

Ranajit Chakraborty, M. R. Srinivasan, and Stephen P. Daiger

Genetics Centers, Graduate School of Biomedical Sciences, University of Texas Health Science Center, Houston

Summary

Multilocus genotype probabilities, estimated using the assumption of independent association of alleles within and across loci, are subject to sampling fluctuation, since allele frequencies used in such computations are derived from samples drawn from a population. We derive exact sampling variances of estimated genotype probabilities and provide simple approximation of sampling variances. Computer simulations conducted using real DNA typing data indicate that, while the sampling distribution of estimated genotype probabilities is not symmetric around the point estimate, the confidence interval of estimated (single-locus or multilocus) genotype probabilities can be obtained from the sampling of a logarithmic transformation of the estimated values. This, in turn, allows an examination of heterogeneity of estimators derived from data on different reference populations. Applications of this theory to DNA typing data at VNTR loci suggest that use of different reference population data may yield significantly different estimates. However, significant differences generally occur with rare (less than 1 in 40,000) genotype probabilities. Conservative estimates of five-locus DNA profile probabilities are always less than 1 in 1 million in an individual from the United States, irrespective of the racial/ethnic origin.

Introduction

With more than 4,000 discovered genetic polymorphic loci available in the human genome (Solomon and Rawlings 1991), it has become possible to define multilocus genotypes for any specific individual for any given subset of these loci. Since many of these loci exhibit a large number of segregating alleles, the number of possible mutilocus genotypes can be very large. For example, since a typical VNTR locus can easily exhibit 20 or more segregating alleles (Odelberg et al. 1989), four such loci will produce $[(20 \times 21)/2]^4 \simeq 2$ billion possible genotypes or more. In many applications of multilocus genotype data, it is important to know the relative frequencies of such genotypes in a population. Therefore, it is relevant to consider how to estimate the multilocus genotype probabilities and to assign bounds of errors of their estimates. Intuitively, the simplest and most direct method would be to predict genotype (single-locus or multilocus) probabilities from their relative frequencies observed in a sample. This is feasible—and is found reliable—for single-locus genotypes at the traditional blood-group and protein loci (Mourant et al. 1976; Tills et al. 1983). But, because of large numbers of segregating alleles at VNTR loci, and because of the fact that the exact number of possible alleles at such loci may be unknown (Chakraborty and Daiger 1991; Devlin et al. 1991), by necessity, it is apparent that alternative methods must be employed to estimate multilocus genotype probabilities for VNTR loci (Chakraborty 1992). However, this challenge is not unique to VNTR loci.

Genotype probabilities at HLA loci generally are computed from HLA-haplotype frequencies (Albert et al. 1984), and this analogy also holds for the immunoglobin Gm genotype computations (Steinberg and

Received June 18, 1992; revision received August 24, 1992.

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

0002-9297/93/5201-0007$02.00

Cook 1981). In other words, the theory of predicting (single-locus or multilocus) genotype probabilities on the basis of allele frequency data has been employed based on Hardy-Weinberg expectations (HWE) for random union of alleles and on the chain-multiplication rule to combine individual locus-specific genotype probabilities (Li 1976). Of course, to justify this method it is necessary to establish random association of alleles within as well as across loci in multilocus data observed in a sample, which has been a subject of various studies (Chakraborty and Kidd 1991; Risch and Devlin 1992; Weir 1992; Chakraborty et al., in press-*a*). Although it is well known that deviations from the product (i.e., HWE) and chain-multiplication (gametic-phase disequilibria) rules may result from population mixture and/or substructuring present in a population (Li 1976), in the context of analysis of VNTR data gathered by the RFLP typing, it is shown that both of these assumptions are adequate. For example, when artifacts (such as incomplete resolution of similar-size RFLP alleles) of the electrophoresis technique were taken into account, Devlin et al. (1990) showed no significant excess of homozygosity in VNTR data in the U.S. populations. Use of intraclass correlations of DNA fragment lengths within loci revealed that the HWE assumption is reasonable (Weir 1992; Chakraborty et al., in press-*a*), and no significant gametic-phase disequilibrium was noted in the analysis by Risch and Devlin (1992). Furthermore, Chakraborty and Jin (1992) demonstrated that the initial observation of heterozygote deficiencies (equivalent to excess homozygosity) noted by Budowle et al. (1991*a*) cannot be explained by population substructuring alone. Presence of nondetectable alleles and/or incomplete resolution of similar-size alleles is a more appropriate factor that explains such observations (Chakraborty et al. 1992*a*). Finally, elsewhere (Chakraborty et al., in press-*b*) we show that the assumptions of alleles within and across loci lead to conservative (biased upward) estimates of multilocus genotype probabilities, with a simple modification of the estimation of homozygous genotype probabilities, as indicated later in the presentation.

Since, even in the above approach, the estimated genotype probabilities depend on allele frequency estimates, which are in turn subject to sampling fluctuation, the purpose of this work is to evaluate the standard errors of estimates of multilocus genotype probabilities based on allele frequency data. We show the derivation of the exact sampling variance for any general multilocus genotype. We further argue that, since multilocus genotype probabilities are generally small, their sampling distributions are not symmetric around the point estimate. Therefore, we discuss an approach of confidence interval estimation of multilocus genotype probability, using the sampling properties of a logarithmic transformation of such probabilities. The reliability of this methodology is checked by computer simulation. Using this theory, we finally describe a test of heterogeneity of different estimates of the same multilocus genotype probability, obtained from different reference populations. This allows a statistical interpretation of comparison of two or more estimates, each of which is small but which may differ from each other by severalfold. Finally, we discuss the appropriateness of such analyses in the context of VNTR fragment size data, because of the quasi-continuous nature of VNTR fragment size measurements in all population samples.

Relative Frequencies of Multilocus Genotypes, and Their Standard Errors

Consider L autosomal codominant loci at each of which there are multiple alleles segregating in a population. For the ℓth locus ($\ell = 1, 2, \ldots, L$), under the assumption of Hardy-Weinberg equilibrium, the genotype probability P_l takes one of the two following forms:

$$P_l = \begin{cases} p_l^2 & \text{for homozygotes} \\ 2p_lq_l & \text{for heterozygotes,} \end{cases} \quad (1)$$

where p_l and q_l are the respective allele frequencies in the population. Under the assumption of random association of alleles across loci (gametic phase equilibrium; Li 1976; Weir 1990), for any set of L loci, the multilocus genotype frequency is given by

$$P_{\text{com}} = \prod_{l=1}^{L} P_l, \quad (2)$$

which assumes that the relevant alleles at different loci aggregate at random to form multilocus genotypes in individuals of the population.

As defined above, equations (1) and (2) are applicable to the entire population, and in practice they must be estimated. This is done from a sample of individuals drawn from the population, from which counts of alleles in a sample of $n/2$ individuals (i.e., n alleles) the estimate of p_l (or q_l) becomes

$$\hat{p}_l = \frac{n_l}{n}, \tag{3}$$

which, substituted in equation (1), gives the estimate, $\hat{P}_l$, of any single-locus genotype probability. Through equation (2), this in turn generates the estimated multilocus probability, say, $\hat{P}_{com}$.

Direct application of the sampling theory of the multinomial distributions (Johnson and Kotz 1969) gives the sampling variance of $\hat{P}_l$ (for the critical steps of the derivation, see the Appendix), given by

$$V(\hat{P}_l) = \begin{cases} [4p_l^3(1-p_l)/n] + [2p_l^2(1-p_l)(3-5p_l)/n^2] + \\ \quad [p_l(1-p_l)(1-6p_l+6p_l^2)/n^3] \text{ for homozygotes} \\ 4p_lq_l[[(p_l+q_l-4p_lq_l)/n] + [(1-3p_l-3q_l+10p_lq_l)/n^2] - \\ \quad [(1-2p_l-2q_l+6p_lq_l)/n^3] \text{ for heterozygotes}. \end{cases} \tag{4}$$

These are the expressions for exact variance of $\hat{P}_l$ for any single-locus genotype probability. Note that when the genotype probability is estimated from allele frequencies (by using eq. [1]), the estimates are not unbiased. For example,

$$E(\hat{P}_l) = \begin{cases} p_l^2 + [p_l(1-p_l)/n] & \text{for homozygotes} \\ 2p_lq_l(n-1)/n & \text{for heterozygotes}, \end{cases} \tag{5}$$

so that for each homozygote, $\hat{P}_l$ is an overestimate of the true probability, while for heterozygotes, $\hat{P}_l$ is an underestimate. The bias, however, is negligible when the sample size (n) is large, say, when 100 or more individuals are sampled.

There is another reason for which an unbiased estimator of P_l may not be preferred. For example, by eq. (5) the unbiased estimator of p_l becomes

$$\hat{P}_l' = \begin{cases} (n\hat{P}_l - \hat{p}_l)/(n-1) & \text{for homozytotes} \\ n\hat{P}_l/(n-1) & \text{for heterozygotes}, \end{cases} \tag{5a}$$

and hence the unbiased estimate of a homozygous genotype becomes zero if the allele is found to have a single copy in the sample (i.e., $\hat{p}_l = 1/n$ and $\hat{P}_l = 1/n^2$).

The expressions of equation (4) are fairly complicated, and simpler approximations are available that ignore terms of order n^{-2} or lower. Using the approximation

$$V[f(x)] \simeq \left[\frac{df(x)^2}{dx}\right] V(x) \tag{6}$$

for any continuous function $f(x)$, of a random variable x, we have

$$\begin{aligned} V(\hat{p}_l^2) &\simeq 4p_l^2 V(\hat{p}_l) \\ &= \frac{4p_l^3(1-p_l)}{n}, \end{aligned} \tag{7a}$$

and

$$\begin{aligned} V(2\hat{p}_l\hat{q}_l) &\simeq 4\left[\left(\frac{\partial p_lq_l}{\partial p_l}\right)^2 V(\hat{q}_l) + \left(\frac{\partial p_lq_l}{\partial q_l}\right)^2 V(\hat{p}_l) \right. \\ &\quad \left. + 2\left(\frac{\partial p_lq_l}{\partial p_l}\right)\left(\frac{\partial p_lq_l}{\partial q_l}\right)\mathrm{Cov}(\hat{p}_l\hat{q}_l)\right] \\ &= 4\left[q_l^2\frac{p_l(1-p_l)}{n} + p_l^2\frac{q_l(1-q_l)}{n} \right. \\ &\quad \left. - 2p_lq_l\frac{p_lq_l}{n}\right] \\ &= 4p_lq_l\left[\frac{p_l+q_l-4p_lq_l}{n}\right], \end{aligned} \tag{7b}$$

which are the first terms of equation (4) for homozygotes and heterozygotes, respectively.

Once $V(\hat{P}_l)$ is obtained as described above, and since the estimated $\hat{P}_l$'s for different loci are independent, from Goodman (1960, 1992) we have

$$\begin{aligned} V(\hat{P}_{com}) &\simeq P_{com}^2\left[\sum_{l=1}^{L} G_l + \sum\sum_{l_1<l_2} G_{l_1}G_{l_3} \right. \\ &\quad + \sum\sum\sum_{l_1<l_2<l_3} G_{l_1}G_{l_2}G_{l_2} + \ldots \\ &\quad \left. + \prod_{l=1}^{L} G_l\right] \\ &= P_{com}^2\left[\prod_{l=1}^{L}(1+G_l) - 1\right], \end{aligned} \tag{8}$$

where $G_l = V(\hat{P}_l)/P_l^2$, giving the sampling variance of the estimated probability of the multilocus genotype. In equation (8) one can use either the exact expressions for $V(\hat{P}_l)$, as given in equation (4), or their approximations, equations (7a ~ 7b); the approximations are generally fairly accurate, which will be shown below. Further approximations may also be done without compromising the accuracy of the estimated variance. For example, in equation (8), since each G_l value is

generally small, the variance of $\hat{P}_{com}$ can be approximated, keeping only the first summation. Since equation (8) is not tedious to compute, we recommend that this later approximation is unnecessary, and hence in this work

$$V(\hat{P}_{com}) \simeq P_{com}^2\left[\prod_{l=1}^{L}(1 + V(\hat{P}_l)/P_l^2) - 1\right] \qquad (8a)$$

will be called "approximate variance of $\hat{P}_{com}$," in which $V(\hat{P}_l)$ is computed from equations (7a ~ 7b).

Confidence Interval of Multilocus Probability

While equations (4), (7), and (8) provide the sampling errors of estimated probabilities of any multilocus genotype, these may not provide an accurate confidence interval of $\hat{P}_{com}$, because $\hat{P}_{com}$ is generally small and its sampling distribution is not symmetric around the point estimate. In contrast, noting that equation (2) yields

$$\ln(\hat{P}_{com}) = \sum_{l=1}^{L}\ln(\hat{P}_l), \qquad (9)$$

we can use the approximation of equation (6) to get

$$V[\ln(\hat{P}_{com})] \simeq \sum_{l=1}^{L}\frac{V(\hat{P}_l)}{\hat{P}_l^2} = s_{com}^2, \qquad (10)$$

say, which gives a $100\alpha\%$ confidence interval for $\ln(\hat{P}_{com})$ defined by

$$\ln(\hat{P}_{com}) \pm Z_\alpha s_{com}, \qquad (11)$$

where Z_α is the two-sided $100\alpha\%$ value of a standard normal distribution. The lower and upper limit, U and L, respectively, of the interval represented by equation (11) can be transformed to get the confidence interval of $\hat{P}_{com}$, given by (e^L, e^U).

Alternatively, since $\hat{P}_{com}$ is obtained from a set of estimated probabilities $\hat{p}_l$ and $\hat{q}_l$, and their associated n for each locus, the empirical confidence interval of $\ln(\hat{P}_{com})$ can be generated by simulating multinomial distribution (up to three classes for each locus) to determine L and U, which can be used to obtain the confidence interval of $\hat{P}_{com}$, (e^L, e^U). In the following section we show that the approximation given in equation (11) is fairly accurate for the purpose of generating confidence intervals of $\hat{P}_{com}$.

Applications of the Theory and Accuracy of the Approximations

Bias of Estimated $\hat{P}_{com}$ and Accuracy of Approximate Variance of $\hat{P}_{com}$

In table 1 we present six examples of the application of the theory described in the previous sections. For the present, assume that there are six genotype data, giving genotype profiles based on single-locus (case 1) through six-loci (case 6) data in which the columns under the heading "x_{1l}" and "x_{2l}" may be regarded as the allelic designations of the genotypes. Also presented in this table are the frequencies of the alleles for each locus scored (for each genotype) in three population samples. In this section we will consider the allele frequency data for the first population sample, to illustrate the theoretical calculations described before; and in a later section we will consider the remaining allele frequency data, to examine the heterogeneity of genotype probability estimates based on these different reference population data.

In table 2 are shown the estimated genotype probabilities (based on the allele frequency data from the Caucasian population sample; table 1) in which the first set of estimated values (biased) are from equations (1) and (2) and in which the second ones (unbiased) are from equation (5a), and these are multiplied values over loci. The last three columns of table 2 represent the standard errors of the first set of estimates; the estimated values are computed using equations (7a ~ 7b), the exact values are from equations (4) and (8), and the empirical ones are from simulations of multinomial distributions (2,000 replications) with allele frequency and sample size data for the Caucasian sample, shown in table 1. For each case, the approximations appear accurate even when only the first terms of the respective equations (e.g., eqq. [1] and [7]) are kept. This is true irrespective of the allele frequencies. The bias of estimation also is small, as seen from comparison of the estimated and exact predictions. While the empirical values (averages from simulation of multinominal distributions with 2,000 replications) of the genotype probabilities generally agree with the estimates (exact or approximate), the empirical standard errors of estimates are always smaller than the estimated standard errors (from our analytical equations). This is expected, since even in 2,000 replica-

Table 1

Six Examples of DNA Typing in Forensic Works from Three Populations

Case and Locus	Band Size		Frequencies (sample size) of Binned Alleles					
			Caucasians		Blacks		Hispanics	
	x_{1l}	x_{2l}	p_l	q_l	p_l	q_l	p_l	q_l
1:								
D2S44	1,605	1,605	.124	.124 (1,584)	.093	.093 (950)	.112	.112 (1,032)
2:								
D14S13	1,787	1,360	.144	.053 (1,502)	.073	.068 (1,048)	.107	.055 (988)
D17S79	1,301	1,301	.224	.224 (1,552)	.256	.256 (1,100)	.268	.268 (1,042)
3:								
D2S44	3,486	1,819	.075	.083 (1,584)	.039	.093 (950)	.062	.100 (1,032)
D14S13	2,089	1,793	.081	.144 (1,502)	.086	.073 (1,048)	.067	.107 (988)
D17S79	1,741	1,486	.199	.198 (1,552)	.107	.195 (1,100)	.183	.166 (1,042)
4:								
D1S7	4,150	2,333	.062	.029 (1,190)	.074	.035 (718)	.090	.044 (1,042)
D2S44	1,915	1,799	.083	.107 (1,584)	.084	.093 (950)	.100	.105 (1,032)
D4S139	7,618	6,100	.131	.108 (1,188)	.109	.103 (896)	.175	.137 (1,044)
D17S79	1,772	1,551	.199	.263 (1,552)	.107	.195 (1,100)	.183	.255 (1,042)
5:								
D1S7	4,150	2,333	.067	.029 (1,190)	.074	.029 (718)	.090	.019 (1,042)
D2S44	2,772	2,357	.040	.038 (1,584)	.028	.076 (950)	.045	.040 (1,032)
D4S139	8,537	5,986	.095	.108 (1,188)	.109	.084 (896)	.130	.137 (1,044)
D14S13	1,826	1,573	.144	.228 (1,502)	.073	.078 (1,048)	.107	.218 (988)
D17S79	1,566	1,566	.263	.263 (1,552)	.195	.195 (1,100)	.255	.255 (1,042)
6:								
D1S7	6,520	4,500	.079	.067 (1,190)	.064	.050 (718)	.081	.081 (1,042)
D2S44	1,780	1,601	.107	.124 (1,584)	.093	.093 (950)	.105	.112 (1,032)
D4S139	14,302	6,630	.102	.191 (1,188)	.038	.103 (896)	.071	.167 (1,044)
D10S28	2,205	2,205	.059	.059 (858)	.076	.076 (576)	.093	.093 (880)
D14S13	3,453	2,124	.030	.081 (1,502)	.060	.086 (1,048)	.048	.067 (988)
D17S79	1,302	1,302	.224	.224 (1,552)	.256	.256 (1,100)	.268	.268 (1,042)

tions of simulations, the rare genotypes will not be represented, and hence it will reduce the observed sampling variance of the estimated genotype probability when relative counts of genotypes are used. This, indeed, shows that the direct count (relative-occurrence) method is not a good practice for estimation of small genotype probabilities. Other inadequacies of the direct-count method are also discussed by Morton (1992). Furthermore, these numerical results indicate that the use of analytical standard errors (with approx-

Table 2

Estimated Probabilities of DNA Profiles, and Their Standard Errors

Case	Estimate of Probability of DNA Sample ($\hat{P}_{com}$)		Standard Error (of $\hat{P}_{com}$)		
	By eq. (1) ~ (3)	By eq. (5a)	Exact	Estimate	Empirical
1 Combined	1.53×10^{-2}	1.54×10^{-2}	2.06×10^{-3}	2.05×10^{-3}	1.44×10^{-3}
2 Combined	7.65×10^{-4}	7.66×10^{-4}	1.17×10^{-4}	1.17×10^{-4}	1.07×10^{-4}
3 Combined	2.28×10^{-5}	2.29×10^{-5}	3.81×10^{-6}	3.80×10^{-6}	2.99×10^{-6}
4 Combined	1.88×10^{-7}	1.89×10^{-7}	4.80×10^{-8}	4.79×10^{-8}	4.21×10^{-8}
5 Combined	1.10×10^{-9}	1.10×10^{-9}	3.37×10^{-10}	3.37×10^{-10}	1.90×10^{-10}
6 Combined	9.46×10^{-12}	9.29×10^{-12}	3.61×10^{-12}	3.58×10^{-12}	3.02×10^{-12}

imation; eqq. [7a ~ 7b] substituted in eq. [8]) would provide a confidence interval that would be conservative (wider than the empirical).

Confidence Interval Estimation

Table 3 provides the confidence interval estimates for the six cases described in table 1. The estimated confidence intervals are from normal approximation of the sampling distribution of $\ln(\hat{P}_{com})$ shown in equation (11), while the empirical ones are from 2,000 replications of multinominal sampling (for each locus) using parameter values corresponding to the Caucasian sample data of table 1. Two observations can be made from these computations. First, the empirical confidence intervals, for $\ln(\hat{P}_{com})$ and $\hat{P}_{com}$, are always narrower than the estimated ones. This occurs because the rare genotypes are not represented in the simulations, which artificially tightens the spread of the observed relative genotype frequencies, making the resultant confidence interval narrower than expected. Again, this illustrates the limitation of the direct-count method of estimation of small genotype probabilities. Second, when the confidence bounds for $\hat{P}_{com}$ are compared with their respective estimates (shown in table 2), we find that the sampling distribution of $\hat{P}_{com}$ is asymmetric around the point estimates, while the symmetry of empirical confidence intervals of $\ln(\hat{P}_{com})$ around its point estimates ensures the normal approximation used (eq. [11]) in obtaining the estimated confidence intervals of the logarithmic transformation of genotype probabilities.

Test for Heterogeneity of Estimated Probabilities

Once the probability of a specific multilocus genotype is estimated, often a question arises whether such an estimate obtained using data sampled from another reference population would be significantly different. In general, empirical estimates of specific genotypes, derived from various alternative reference population data, have been shown to differ, but not in any meaningful way (i.e., no multilocus genotype that is rare in a population becomes common in another population; e.g., see Chakraborty and Kidd 1991; Weir 1992). However, the theory described in the previous sections can be used to formally test whether the estimated multilocus genotype probabilities based on different sets of allele frequency data are significantly different. This can be done using Rao's heterogeneity χ^2 test criterion (Rao 1973), applied on $\ln\hat{P}_{com}$ estimates, since the sampling distribution of this statistic approximates a normal distribution, as shown earlier. Suppose that $t_i = \ln[\hat{P}_{com}(i)]$ denotes the logarithm of the estimate of a multilocus DNA profile from the ith set of allele frequency data $i = 1, 2, \ldots, r$ and that

Table 3

Confidence Interval Estimates of Multilocus DNA Profile Frequencies

	$\ln(\hat{P}_{com})$			$\hat{P}_{com}$	
	Upper 95%	Lower 95%	Point Estimate	Upper 95%	Lower 95%
1:					
Estimated	−4.44	−3.91	−4.17	1.18×10^{-2}	1.99×10^{-2}
Empirical.....	−4.36	−4.00	−4.19	1.27×10^{-2}	1.82×10^{-2}
2:					
Estimated	−7.47	−6.87	−7.17	5.67×10^{-4}	1.03×10^{-3}
Empirical.....	−7.26	−6.94	−7.12	7.03×10^{-4}	9.61×10^{-4}
3:					
Estimated	−11.01	−10.36	−10.68	1.65×10^{-5}	3.17×10^{-5}
Empirical.....	−10.84	−10.49	−10.68	1.95×10^{-5}	2.79×10^{-5}
4:					
Estimated	−15.98	−14.98	−15.48	1.15×10^{-7}	3.11×10^{-7}
Empirical.....	−15.73	−15.22	−15.45	1.48×10^{-7}	2.46×10^{-7}
5:					
Estimated	−21.23	−20.03	−20.63	6.05×10^{-10}	2.00×10^{-9}
Empirical.....	−20.69	−20.36	−20.53	1.03×10^{-9}	1.44×10^{-9}
6:					
Estimated	−26.16	−24.65	−25.40	4.36×10^{-12}	1.98×10^{-11}
Empirical.....	−25.77	−25.15	−25.40	6.40×10^{-12}	1.19×10^{-11}

s_i^2 is the variance of this estimate (obtained from eq. [10]). The different sets of allele frequencies could be from different populations and/or could be data from different laboratories. The heterogeneity of these estimates is tested by the criterion

$$\chi^2_{r-1} = \sum_{i=1}^{r} t_i^2/s_i^2 - \left(\sum_{i=1}^{r} t_i/s_i^2\right)^2 / \left(\sum_{i=1}^{r} 1/s_i^2\right), \quad (12)$$

which is asymptotically distributed as a χ^2 with $(r - 1)$ df. If this test criterion suggests that the various estimates are not heterogeneous, the pooled estimate is given by

$$\bar{t} = \left(\sum_{i=1}^{r} t_i/s_i^2\right) / \left(\sum_{i=1}^{r} 1/s_i^2\right), \quad (13a)$$

whose sampling variance is

$$s^2 = \sum_{i=1}^{e} (1/s_i^2)^{-1}. \quad (13b)$$

The pooled estimate of P_{com} based on all available allele frequency data would then be given by $e^{\bar{t}}$, while its confidence interval can be obtained by using $\bar{t}$ and s^2 in equation (11) and translating the upper and lower limits, as done earlier for a single set of allele frequency data. Table 4 shows the application of this test procedure for the six sets of genotype data shown in table 1. Alternative estimates of these genotype probabilities were obtained by using the allele frequency data from the three samples (Caucasian, black, and Hispanic) given in table 1). While these examples are used only for illustrative purposes, the results are instructive in several respects. First, it is true that use of different reference population data might yield different estimates of the same DNA profile probability, but such differences are statistically significant (as detected by the heterogeneity χ^2 statistic) only when the estimates are small (in these examples, below 1 in 40,000). Second, in all cases shown in table 4, the heterogeneity χ^2 statistic is contributed mainly by the low estimate obtained from the sample of U.S. blacks. This is consistent with the evolutionary history of human populations. The U.S. blacks, 75% or more of whose gene pool is of African origin (Reed 1969; Chakraborty 1986; Chakraborty et al. 1992*b*), probably represent the oldest major racial group of the world (Nei and Roychoudhury 1982; Bowcock et al. 1991), and because of past as well as recent admixture they also have the highest gene diversity within them. This is true not only with respect to a greater number of alleles (Mohrenweiser et al. 1987), but also with respect to the frequencies of specific alleles within blacks, which are generally lower compared with those in other ra-

Table 4

Test of Heterogeneity of $\hat{P}_{com}$ Estimates from Different Samples

Case and Estimate	Caucasians	Blacks	Hispanics	χ^2 (2df)
1:				
$\log\hat{P}_{com}$	−4.17 ± .13	−4.74 ± .20	−4.37 ± .17	5.62
$\hat{P}_{com}$	1/65	1/114	1/79	
2:				
$\log\hat{P}_{com}$	−7.17 ± .15	−7.34 ± .18	−7.07 ± .19	1.03
$\hat{P}_{com}$	1/1,304	1/1,534	1/1,181	
3:				
$\log\hat{P}_{com}$	−10.69 ± .17	−12.48 ± .25	−11.44 ± .22	35.83**
$\hat{P}_{com}$	1/43,777	1/263,870	1/92,849	
4:				
$\log\hat{P}_{com}$	−15.48 ± .25	−16.40 ± .31	−14.11 ± .24	36.09**
$\hat{P}_{com}$	$1/(5.3 \times 10^6)$	$1/(13.2 \times 10^6)$	$1/(1.3 \times 10^6)$	
5:				
$\log\hat{P}_{com}$	−20.63 ± .31	−22.66 ± .41	−20.43 ± .36	20.44**
$\hat{P}_{com}$	$1/(9.1 \times 10^8)$	$1/(6.9 \times 10^9)$	$1/(7.5 \times 10^8)$	
6:				
$\log\hat{P}_{com}$	−25.38 ± .38	−26.39 ± .46	−24.25 ± .37	13.63**
$\hat{P}_{com}$	$1/(1.1 \times 10^{11})$	$1/(2.9 \times 10^{11})$	$1/(3.3 \times 10^{10})$	

** $P < .001$.

cial groups. These result in smaller genotype probabilities (at both the single-locus and multilocus levels), consistent with the findings in table 4. Third, the significant differences of multilocus genotype probability estimates for the different racial groups may have also emerged from the fact that we used rebinned fragment size distributions (Budowle et al. 1991*b*) in which the binned class limits are different for the three racial groups, and hence the different probability estimates do not always correspond to identical fragment size.

Applications to Forensic DNA Typing Data

Although in both theory and application we specifically used the concept of discrete alleles and genotypes, the principles also apply to forensic DNA typing data. In current forensic applications of DNA typing, DNA samples are digested with a restriction enzyme and are hybridized with a single-locus VNTR probe (e.g., MS1, YNH24, TBQ7, CMM101, V1; see Nakamura et al. 1987; Budowle et al. 1991*a*, 1991*b*). These generally give a single-band or two-banded pattern due to copy-number variation of the core repeat sequences at the VNTR locus. Sizes (approximate numbers of base pairs) of the relevant bands are determined by automated algorithms and by then comparing them with size standards run concurrently in Southern gel electrophoresis. Technical limitations of band sizing from Southern gels cannot eliminate measurement errors (Budowle et al. 1991*a*; Evett 1991; Berry et al. 1992), but, in general, measurement errors are small (about 2.5% or smaller) with respect to VNTR fragment sizes. Instead of definition of an allele by its actual fragment size, two alternative approaches of "binning" fragment sizes in population data have been suggested—i.e., "fixed" or "floating" bins (Budowle et al. 1991*a*, 1991*b*; Foreman 1991)—whereby fragment sizes are grouped in reference to the extent of measurement error for size range of fragment lengths (either by fixed class limits [Budowle et al. 1991*a*] or by allowing a certain width around the specific size observed for a given fragment length [Balazs et al. 1989; Foreman 1991]). The population data on fragment lengths would then correspond to categorized classes, analogous to allele frequency data in the case of the fixed-bin approach. In the case of floating bins, there is no uniform set of categories for all fragment sizes. But, since for any given individual we observe up to two different fragment sizes, the population data will be grouped in up to three categories specifically constructed for the given observation. This definition of binned alleles is no different from the traditional definition of alleles that are technology based (Morris et al. 1989). Therefore, the logic of the theory applies to DNA typing data as well, and measurement errors of fragment lengths are embedded in the definition of binned alleles.

We also should note that, for single-band DNA patterns at a locus, we used the probability $p\hat{?}$, assuming that the single-band patterns reflect true homozygosity. It is known that this may not always be true, since single-band patterns sometimes may result from the inability to detect another allele, which is either too small or too large to be scored in the Southern gel protocol (Budowle et al. 1991*a*; Chakraborty et al. 1992*a*). To guard against this, Budowle et al. (1991*a*) suggested a further conservative approach, representing the probability of a single band pattern by $2p_l$. The variance for such a probability estimate is even simpler, since $V(\hat{P}_l) = 4p_l(1 - p_l)/n$, which can be used instead of the formula given in the first expression of equation (4) (or eq. [7a]). By analogy, the theory described here also can be used to determine the standard error or confidence interval of phenotype probability estimates. The expression for $V(\hat{P}_l)$ needs to be changed suitably with the definition of the phenotypes. As mentioned before, this simple change in estimating a homozygous genotype probability, when used in conjunction with the chain-multiplication rule, gives an overestimate of any multilocus genotype probability, even if the population is not strictly a random mating one (Chakraborty et al., in press-*b*).

At this stage, we go back to the example of six genotypes shown in table 1. They are, indeed, six actual DNA profiles, and the second column of table 1 indicates the VNTR fragment sizes (represented by x_{1l} and x_{2l} for the two bands observed at the lth locus) for the corresponding VNTR loci. The allele frequencies in three samples (shown in the last three sets of columns of table 1) are the "rebinned allele" frequencies for the pooled data on Caucasians, blacks, and Hispanics, collected from different geographic locations in the United States by the Federal Bureau of Investigation (FBI) Forensic Sciences Research Unit and published by Budowle et al. (1991*b*).

To illustrate the general nature of the confidence intervals for DNA profile frequencies for all individuals in the FBI data base, we computed the estimate of each person's observed DNA profiles, grouping them by their recorded race/ethnicity and using the race-specific rebinned allele frequencies (Budowle et al. 1991*b*). The fragment size data were provided by Dr.

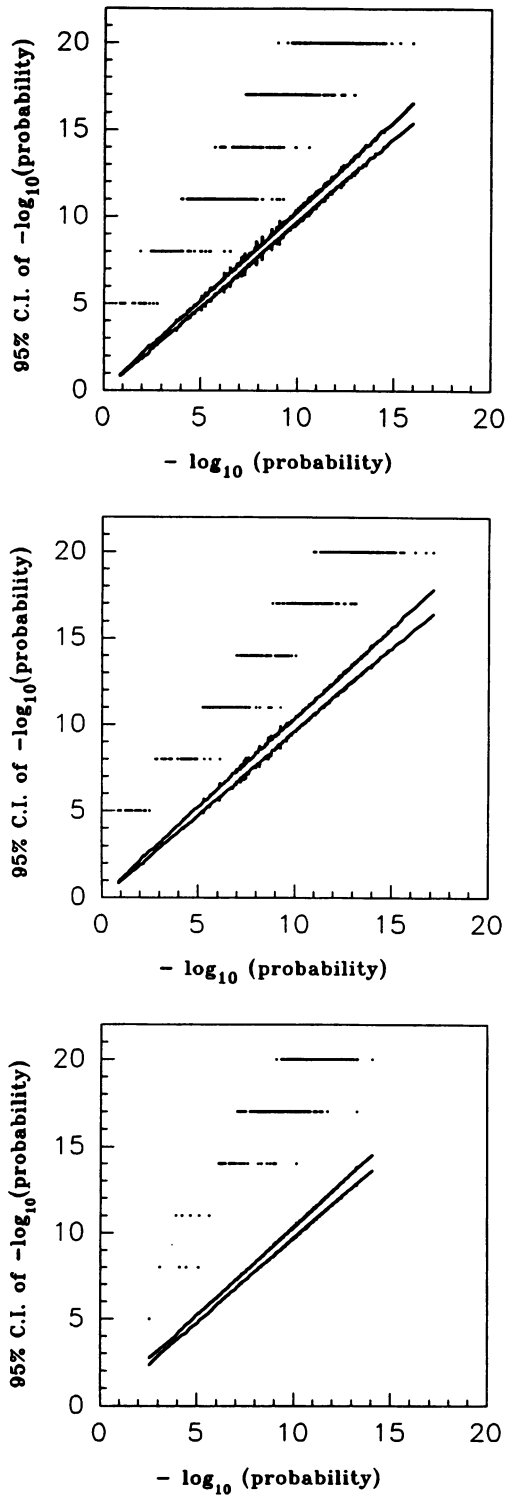

Figure 1 95% confidence interval (95% C.I.) estimates for DNA profile probabilities observed in the FBI data base (Budowle et al. 1991*b*), plotted against the estimates. All values are in a logarithmic (base 10) scale, so that the probabilities and confidence limits are to be interpreted as $1/x$, x in log(base10) scale. The individuals are differentiated on the basis of their racial/ethnic origin: *A*, 840 Caucasians; *B*, 583 blacks; and *C*, 552 Hispanics. Also shown in the graph as horizontal points are the ranges of estimated DNA profile probabilities for individuals having one-locus to six-locus data in the data base.

Bruce Budowle (FBI Academy) for this analysis. Figure 1 shows the confidence interval estimates plotted against the point estimates for each individual's DNA profile. Specifically, for each individual's DNA profile (based on all loci scored for the individual) in the data base, the upper and lower 95% confidence interval limits of $\hat{P}_{com}$ are plotted against $\hat{P}_{com}$, generating the lines of each panel of figure 1. The wide range of estimates is expected, since the data consist of variable numbers of loci scored for different individuals (from one locus up to six loci). In each panel, therefore, we indicate the range of point estimates of DNA profile frequencies, shown by horizontal points. All calculations shown in this figure are in logarithmic (base 10) scale.

At this stage we note that the numerical values plotted in figure 1 are based on the approximations used in deriving equations (10) and (11). Although we verified the accuracies of these approximations by using the six examples shown in table 1, these examples truly represent six individual's DNA profiles that are close to the averages of one-locus to six-loci profiles, shown in figure 1. The limited number of illustrations of the adequacy of approximations should not be a concern, since they encompass almost the entire range of multilocus genotype probabilities in the data base (see fig. 1). In fact, since each individual's DNA profile in the data base is unique, almost 2,000 (size of the data base) separate simulations (each with 2,000 replicates) would have been necessary to check the accuracy of the numerical values plotted in this figure. Nevertheless, data presented in table 3 indicate that the plotted confidence limits are wider than the empirical ones, because, even with 2,000 replications of multinomial distributions, rare genotypes may not be captured.

Data shown in figure 1 illustrate three points. First, it is not generally true that every single-locus DNA profile would be more frequent than any two-locus profile, since the horizontal points are overlapping for all ethnic groups. But, in general, the profile frequency is a decreasing function of the number of loci scored. This supports Crow's view (see Roberts 1991) that, with more loci scored, the specific DNA profile frequency would reach a diminishing value, irrespective

either of the racial origin of the individual or of the specific DNA profile observed. Second, it is true that the confidence interval (as judged by the ratio of the upper and lower limit) gets wider as the profile frequency becomes rarer (or, x gets larger). This is an inherent sampling property, and it explains how a rare DNA profile can have sampling fluctuation enough to make the confidence limits (upper and lower) different by more than one order of magnitude. Third, in spite of the relatively large standard errors of estimates of rare genotype profile frequencies, this figure supports Risch and Devlin's (1992) assertion that, irrespective of the racial/ethnic origin of individuals, most five-locus DNA profiles have a frequency no larger than 1 in 1 million, in fixed-bin data.

Acknowledgments

This work was supported by research grants NIH-IR01-GM41399 and NIJ-90-CX-0038. We thank Dr. B. Budowle and his laboratory staff for providing the data for analysis reported in this paper. Comments and suggestions of Drs. B. Budowle, E. Boerwinkle, and W. J. Schull and two anonymous reviewers are greatly appreciated. The opinions expressed in this document are those of authors, and they do not necessarily represent endorsement of the granting agencies supporting this research.

Appendix

Derivation of Equation (4)

If in a sample of n alleles ($n/2$ individuals), n_{1l} and n_{2l} of them reside in bins B_{1l} and B_{2l}, and the expectations of these frequencies are

$$E(n_{1l}) = np_l \text{ and } E(n_{2l}) = nq_l, \tag{A1}$$

where p_l and q_l are the true probabilities that DNA fragments are in bins B_{1l} and B_{2l} in the population. Furthermore, using the general theory of multinomial sampling (Johnson and Kotz 1969), we have

$$E(n_{1l}^2) = n(n-1)p_l^2 + np_l \tag{A2}$$

and

$$\begin{aligned} E(n_{1l}^4) &= E[n_{1l}(n_{1l}-1)(n_{1l}-2)(n_{il}-3) \\ &+ 6n_{1l}(n_{1l}-1)(n_{1l}-2) \\ &+ 7n_{1l}(n_{1l}-1)+n_{1l}] \\ &= n(n-1)(n-2)(n-3)p_l^4 \\ &+ 6n(n-1)(n-2)p_l^3 \\ &+ 7n(n-1)p_l^2 + np_l . \end{aligned} \tag{A3}$$

Therefore,

$$\begin{aligned} V(\hat{p}_{1l}^2) &= V(n_{1l}^2)/n^4 \\ &= n(n-1)(n-2)(n-3)p_l^4 \\ &+ 6n(n-1)(n-2)p_l^3 \\ &+ 7n(n-1)p_l^2 + np_l/n^4 \\ &- [(n(n-1)p_l^2 + np_l)/n^2]^2 . \end{aligned} \tag{A4}$$

Algebraic simplification of equation [A4] leads to the expression given in equation (4) for homozygotes.

For heterozygotes, we have to use $\hat{P}_l = 2\hat{p}_l\hat{q}_l = 2n_{1l}n_{2l}/n^2$. Again, as given by Johnson and Kotz (1969),

$$E(n_{1l}n_{2l}) = n(n-1)p_lq_l \tag{A5}$$

and

$$\begin{aligned} V(n_{1l}n_{2l}) &= E(n_{1l}^2n_{2l}^2) - E^2(n_{1l}n_{2l}) \\ &= E[n_{1l}(n_{1l}-1)n_{2l}(n_{2l}-1) \\ &+ n_{1l}(n_{1l}-1)n_{2l} + n_{1l}n_{2l}(n_{2l}-1) \\ &+ n_{1l}n_{2l}] - E^2(n_{1l}n_{2l}) \\ &= n(n-1)(n-2)(n-3)p_l^2q_l^2 \\ &+ n(n-1)(n-2)p_l^2q_l \\ &+ n(n-1)(n-2)p_lq_l^2 + n(n-1)p_lq_l \\ &- n^2(n-1)^2p_l^2q_l^2 . \end{aligned} \tag{A6}$$

Dividing equation (A6) by n^4 and rearranging the terms, we get the variance of $\hat{P}_l$ for heterozygotes that is given in the second part of equation (4).

References

Albert ED, Baur MP, Mayr WR (1984) Histo-compatibility testing. Springer, New York

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

Berry DA, Evett IW, Pinchin R (1992) Statistical inference in crime investigations using deoxyribonucleic acid profiling. Appl Stat 41:499–531

Bowcock AM, Kidd JR, Mountain JL, Herbert JM, Carotenuto L, Kidd KK, Cavalli-Sforza LL (1991) Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Proc Natl Acad Sci USA 88:839–843

Budowle B, Giusti AM, Waye JS, Baechtel FS, Fourney RM, Adams DE, Presley LA, et al (1991*a*) 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, Monson KL, Anoe KS, Baechtel DL, Bergman DL, Buel E, Campbell PA, et al (1991*b*) A preliminary

report on binned general population data on six VNTR loci in caucasians, blacks and hispanics from the United States. Crime Lab Dig 18:9–26

Chakraborty R (1986) Gene admixture in human populations: models and predictions. Yearbook Phys Anthropol 29:1–43

——— (1992) Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. Hum Biol 56:141–159

Chakraborty R, Daiger SP (1991) Polymorphism at VNTR loci suggest homogeneity of the white population of Utah. Hum Biol 63:571–587

Chakraborty R, de Andrade M, Daiger SP, Budowle B (1992*a*) Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. Ann Hum Genet 56:45–57

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

Chakraborty R, Kamboh MI, Nwankwo M, Ferrell RE (1992*b*) Caucasian genes in American blacks: new data. Am J Hum Genet 50:145–155

Chakraborty R, Kidd KK (1991) The utility of DNA typing in forensic work. Science 254:1735–1739

Chakraborty R, Srinivasan MR, de Andrade M. Estimation of intraclass and interclass correlations of allele sizes establishes random association of alleles within and between loci in DNA typing data. Genetics (in press-*a*)

Chakraborty R, Srinivasan MR, Jin L, de Andrade M. Effects of population subdivision and allele frequency differences on interpretation of DNA typing data for human identification. In: Proceedings of the Third International Symposium on Human Identification. Promega, Madison (in press-*b*)

Devlin B, Risch N, Roeder K (1990) No excess of homozygosity at loci used for DNA fingerprinting. Science 24: 1416–1420

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

Evett IW (1991) Trivial error. Nature 354:114

Foreman L (1991) The role of DNA in courtroom: issues and concerns in the analysis of VNTRs for forensic investigations. Am J Hum Genet Suppl 49:64

Goodman LA (1960) On the exact variance of products. J Am Stat Assoc 55:708–713

——— (1962) On the exact product of k random variables. J Am Stat Assoc 57:54–60

Johnson NL, Kotz S (1969) Discrete distributions. Houghton-Mifflin, Boston

Li CC (1976) First course in population genetics. Boxwood, Pacific Grove, CA

Mohrenweiser HW, Wurzinger KH, Neel JV (1987) Frequency and distribution of rare electrophoretic mobility variants in a population of newborns in Ann Arbor, Michigan. Ann Hum Genet 51:303–316

Morris JW, Sanda AI, Glassberg J (1989) Biostatistical evaluation of evidence from continuous allele frequency distribution deoxyribonucleic acid (DNA) in reference to disputed paternity identity. J Forensic Sci 34:1311–1317

Morton NE (1992) Genetic structure of forensic populations. Proc Natl Acad Sci USA 89:2256–2560

Mourant AE, Kopec AC, Domaniewska-Sobczak K (1976) The distribution of the human blood groups and other polymorphisms, 2d ed. Oxford University Press, Oxford

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, et al (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616–1622

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

Odelberg SJ, Platke R, Eldridge JR, Ballard L, O'Connell P, Nakamura Y, Leppert M, et al (1989) Characterization of eight VNTR loci by agarose gel electrophoresis. Genomics 5:915–924

Rao CR (1973) Linear statistical inference and its applications. John Wiley, New York

Reed TE (1969) Caucasian genes in American Negroes. Science 165:762–768

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

Roberts L (1991) Fight erupts over DNA fingerprinting. Science 254:1721–1723

Solomon E, Rawlings C (eds) (1991) Human Gene Mapping 11: Eleventh International Workshop on Human Gene Mapping. Cytogenet Cell Genet 52:1–2000

Steinberg AG, Cook CE (1981) The distribution of human immunoglobin allotypes. Oxford University Press, Oxford

Tills D, Kopec AC, Tills RE (1983) The distribution of the human blood groups and other polymorphisms, suppl 1. Oxford University Press, Oxford

Weir BS (1990) Genetic data analysis. Sinauer, Sunderland, MA

——— (1992) Independence of VNTR alleles defined as fixed bins. Genetics 130:873–887