Intraclass and Interclass Correlations of Allele Sizes Within and Between Loci in DNA Typing Data

R. Chakraborty, M. R. Srinivasan and M. de Andrade

Center for Demographic and Population Genetics, The University of Texas Graduate School of Biomedical Sciences, Houston, Texas 77225

Manuscript received May 1, 1992
Accepted for publication October 17, 1992

ABSTRACT

Nonparametric measures of correlations of DNA fragment lengths within and between variable number of tandem repeat (VNTR) loci are proposed to test the hypothesis of random association of allele sizes at VNTR loci. Transformations of these nonparametric correlation measures are suggested to detect deviations of their null expectations caused by population subdivision and errors of measurement of VNTR fragment lengths. Analytic and permutation-based computer simulation studies are performed to show that under the hypothesis of independence of allele sizes the transformed correlation measures are normally distributed, irrespective of the VNTR fragment size distribution in the population even when the number of individuals samples is as low as 100. Power calculations are performed to establish that the current population data on six VNTR loci in the US Hispanic sample are in accordance with the hypothesis of random association of allele sizes within and between loci. Implications of these results in the context of forensic use of DNA typing are also discussed.

IT is now well established that the human genome contains probably thousands of interspersed variable number of tandem repeat (VNTR) loci each of which provides a degree of polymorphism greater than the traditional serologically and biochemically detectable polymorphic loci (JEFFREYS, WILSON and THEIN 1985; NAKAMURA *et al.* 1987). Such hypervariable DNA loci now are being commonly used for human identification, determination of relatedness between individuals and disease-gene mapping (BALLENTYNE, SENSABAUGH and WITKOWSKI 1989; BURKE *et al.* 1991). The abundance of allelic and genotypic possibilities at these loci, raises a challenge for testing the validity of the classic population genetic models (such as Hardy-Weinberg equilibrium and gametic phase equilibrium) because no feasible sample size can be prescribed which will encompass all possible genotypes in a sample (CHAKRABORTY 1992). Since interpretation of DNA typing data generally relies on the assumption of independence of occurrences of DNA fragments of different length within an individual, WEIR (1992) recently studied the correlation of DNA fragment lengths within and between loci in individuals, using an analysis of variance-based intraclass correlation approach. He found little evidence for correlations between pairs of VNTR fragments, within or between loci in a database consisting of 2046 individuals divided by geographic location of sampling within the United States and general ethnic groups or categories (Caucasian, Blacks and Hispanics).

The distributions of quasi-continuous random variables such as VNTR fragment lengths are not known to follow any standard distribution (BALAZS *et al.* 1989, BUDOWLE *et al.* 1991b). In the context of studying familial resemblance of quantitative traits, some authors claim that alternative estimators of intra- and interclass correlations may be preferred than the analysis of variance (ANOVA) model-based estimators for such distributions (see *e.g.*, KARLIN, CAMERON and WILLIAMS 1981; SRIVASTAVA 1984). The purpose of this research is to show that WEIR's (1992) estimators are virtually identical to the estimators proposed by KARLIN, CAMERON and WILLIAMS (1981), even though the assumptions of ANOVA models may be violated for VNTR fragment size data. Since VNTR fragment sizes measured by Southern gel electrophoresis methods (BUDOWLE *et al.* 1991a) involve measurement errors, we study the effect of measurement errors on the estimates of intraclass and interclass correlations. Furthermore, we show that the standard transformation of these correlations yield normally distributed variables through which the significant departures of these correlations from their random expectations may be tested. Computer simulation results show the appropriateness of the normal approximations of transformed correlations using the DNA typing data in the pooled Hispanic sample from Florida and Texas gathered by the Forensic Laboratory of the Federal Bureau of Investigation. Such normal approximations also allow evaluation of the statistical power of the test procedures.

INTRACLASS AND INTERCLASS CORRELATIONS OF ALLELE SIZES

For single locus VNTR probes, an individual's (say, i) DNA profile may be represented by a pair (x_{i1}, x_{i2}) of fragment sizes where x_{ij}'s may have a quasi-continuous distribution of complex shape (multimodal, skewed, etc.) due to unknown etiology. Following WEIR (1992), three expectations of functions of the x_{ij} values define the intraclass correlation (ρ_x) between fragment sizes, given by $E(x_{ij}^2) = \mu_x^2 + \sigma_x^2$, $E(x_{ij}x_{ij'}) = \mu_x^2 + \rho_x\sigma_x^2$; for $j \neq j'$ and $E(x_{ij}\, x_{i'j'}) = \mu_x^2$; for $i \neq i'$, where x_{ij}, $j = 1, 2$ are assumed to have been drawn from a distribution with mean μ_x and variance σ_x^2.

With the notations

$$\bar{x} = \sum_{i=1}^{n} \sum_{j=1}^{2} x_{ij}/2n \tag{1a}$$

$$s_x^2 = \sum_{i=1}^{n} \sum_{j=1}^{2} x_{ij}^2 - 2n\bar{x}^2 \tag{1b}$$

$$s_{xx} = \sum_{i=1}^{n} x_{i1}x_{i2} - n\bar{x}^2; \tag{1c}$$

where n is the number of individuals in a sample, KARLIN, CAMERON and WILLIAMS (1981) nonparametric estimator $\hat{\rho}_x$ of ρ_x becomes

$$\hat{\rho}_x = 2s_{xx}/s_x^2, \tag{2}$$

while WEIR's (1992) ANOVA-based estimator $\tilde{\rho}_x$ is

$$\tilde{\rho}_x = \frac{2(2n-1)s_{xx} + s_x^2}{2s_{xx} + (2n-1)s_x^2}. \tag{3}$$

Therefore $\hat{\rho}_x$ is related with $\tilde{\rho}_x$ by the equation

$$\tilde{\rho}_x = \frac{1 + (2n-1)\hat{\rho}_x}{(2n-1) + \hat{\rho}_x}. \tag{4}$$

For large n these two estimators are virtually identical, even without any assumption regarding the distributions of x_{ij}'s.

Similarly the association of fragment sizes between pairs of loci can also be studied by introducing parameters ρ_{xy_1} and ρ_{xy_2} which indicate the correlations between two fragment sizes at two loci depending on whether they are from the same or different gametes. In doing so, when the fragment size for a pair of loci for the ith individual is represented by (x_{i_1}, x_{i_2}), (y_{i_1}, y_{i_2}); $(i = 1, 2, \ldots, n)$, so that if the y_{ij}'s are from a distribution with mean μ_y and variance σ_y^2, the parameters ρ_{xy_1} and ρ_{xy_2} are defined by (WEIR 1992):

$E(x_{ij}y_{ij}) = \mu_x\mu_y + \rho_{xy_1}\sigma_x\sigma_y$ and $E(x_{ij}y_{ij'}) = \mu_x\mu_y + \rho_{xy_2}\sigma_x\sigma_y$ for $j \neq j'$. WEIR (1992) estimated the average correlation between the two loci by estimating $(\rho_{xy_1} + \rho_{xy_2})/2$. With $\bar{y}$, s_y^2, and s_{yy} defined by expressions analogous to eqns. (1a) ~ (1c), and by defining

$$s_{xy} = \sum_{i=1}^{n}\sum_{j=1}^{2}\sum_{j'=1}^{2} x_{ij}y_{ij'} - 4n\bar{x}\bar{y},$$

the ANOVA-based estimator WEIR's (1992) of interclass correlation is given by

$$\tilde{\rho}_{xy} = \frac{ns_{xy}}{[2s_{xx} + (2n-1)s_x^2]^{1/2}[2s_{yy} + (2n-1)s_y^2]^{1/2}}, \tag{5}$$

while KARLIN, CAMERON and WILLIAMS (1981) generalized nonparametric correlation between two sets of variables, the (x, y)-set in this context, becomes

$$\hat{\rho}_{xy} = s_{xy}/[2(s_x^2 s_y^2)^{1/2}]. \tag{6}$$

When $x_{i1} = y_{i1}$ and $x_{i2} = y_{i2}$, for all i, $\hat{\rho}_{xy}$ and $\hat{\rho}_x$ becomes identical [see section 5 of KARLIN, CAMERON and WILLIAMS (1981)]. This implies that the principle of deriving the two nonparametric estimators are essentially identical, even though the estimator $\hat{\rho}_{xy}$ resembles an interclass correlation measure.

Equations 5 and 6 show that the ANOVA-based estimator $\tilde{\rho}_{xy}$ of ρ_{xy} is related to the generalized nonparametric estimator $\hat{\rho}_{xy}$ as

$$\tilde{\rho}_{xy} = \frac{2n\hat{\rho}_{xy}}{[(2n-1) + \hat{\rho}_x]^{1/2}[(2n-1) + \hat{\rho}_x]^{1/2}}. \tag{7}$$

As in the case of correlation of fragment sizes within loci, this establishes that the ANOVA-based estimator of ρ_{xy} is virtually identical to the generalized nonparametric estimator of KARLIN, CAMERON and WILLIAMS (1981) when the sample size n is large, without any assumption regarding the fragment size distributions.

ASYMPTOTIC EXPECTATIONS OF CORRELATIONS

Under the assumption of random association: If the fragment sizes measured are determined by an unknown number of discrete alleles $A_1, A_2, \ldots, A_r$, where the true size of the kth allele is a_k ($k = 1, 2, \ldots, r$), the paired allele sizes for individuals in a sample, (x_{i1}, x_{i2}), $i = 1, 2, \ldots, n$ will have the independent and identical distribution given by

$$(x_{i1}, x_{i2}) = \begin{cases} (a_k, a_k) \text{ with probability } p_k^2 & \text{for } k = 1, 2, \ldots, r, \\ (a_k, a_l) \text{ with probability } p_k p_l & \text{for } k \neq l = 1, 2, \ldots, r, \end{cases}$$

where p_k is the frequency of the kth allele in the population.

Under the assumption of random association of alleles within a locus, $E(x_{ij})$, $E(x_{ij}^2)$, $E(x_{i1}x_{i2})$ and $V(\bar{x})$ are $\sum_{k=1}^{r} a_k p_k$, $\sum_{k=1}^{r} a_k^2 p_k$, $(\sum_{k=1}^{r} a_k p_k)^2$ and $[\sum_{k=1}^{r} a_k^2 p_k - (\sum_{k=1}^{r} a_k p_k)^2]/2n$ respectively.

Using these, we obtain

$$2E(s_{xx}) = -\left[\sum_{k=1}^{r} a_k^2 p_k - \left(\sum_{k=1}^{r} a_k p_k\right)^2\right]$$

and

$$E(s_x^2) = (2n-1)\left[\sum_{k=1}^{r} a_k^2 p_k - \left(\sum_{k=1}^{r} a_k p_k\right)^2\right]$$

Therefore, under the assumption of random association of alleles within a locus the asymptotic expectation of $\hat{\rho}_x$ becomes,

$$E(\hat{\rho}_x) \simeq -(2n-1)^{-1}, \tag{8}$$

while the ANOVA-based estimator $\tilde{\rho}_x$ is asymptotically unbiased. These asymptotic results are invariant of the true allele sizes (a_k's) as well as the allele frequency distribution, showing that the distribution-free property of KARLIN, CAMERON and WILLIAMS' estimator also holds for the ANOVA-based estimator. FISHER's (1928) transformation of intraclass correlation indicates that the transformed variable $t = \ln\sqrt{(1+\hat{\rho}_x)/(1-\hat{\rho}_x)}$ has an asymptotic expectation under the hypothesis of random association

$$E(t) \simeq \ln\sqrt{(n-1)/n},$$

with an approximate variance $(n - 2)^{-1}$, so that a transformation

$$z_x = \sqrt{(n-2)}.\ \ln\sqrt{\frac{1+\hat{\rho}_x}{1-\hat{\rho}_x} \times \frac{n}{n-1}} \tag{9}$$

yields a standard variate (*i.e.*, with mean 0 and variance 1). Similar transformation applied to the ANOVA-based estimator has the same property, with the change that the factor $n/(n - 1)$ is no longer necessary, since the transformed statistic t defined for $\tilde{\rho}_x$ has a zero expectation. In the sequel we show that the distribution of z_x can be approximated by a standard normal distribution for VNTR data on allele sizes with a fairly good precision, irrespective of the VNTR fragment size distributions.

Similar argument holds for the correlation of allele sizes of two loci. Representing the alleles of the second locus by $B_1, B_2, \ldots, B_s$ with frequencies $q_1, q_2, \ldots, q_s$ in the population, the marginal distribution of (y_{i1}, y_{i2}) can be written as

$$(y_{i1}, y_{i2}) = \begin{cases} (b_u, b_u) \text{ with probability } q_u^2 & \text{for } u = 1, 2, \ldots, s, \\ (b_u, b_v) \text{ with probability } q_u q_v & \text{for } u \neq v = 1, 2, \ldots, s, \end{cases}$$

where b_u is the true size of the u-th allele ($u = 1, 2, \ldots, s$) at the B-locus in the population. Under the independence assumption, $E(x_{ij}y_{i'j'}) = (\sum_{k=1}^{r} a_k p_k)(\sum_{u=1}^{s} b_u q_u)$; for all $i, i' = 1, \ldots, n$; $j, j' = 1, 2$ which is also equal to $E(\bar{x}\bar{y})$. This shows that the expectation of $\hat{\rho}_{xy}$ is zero irrespective of the sample size. The asymptotic variance of $\hat{\rho}_{xy}$ also differs from that of $\hat{\rho}_x$ or $\hat{\rho}_y$, because it is an interclass correlation and there is twice as much information (with respect to sample size) in it compared with the two intraclass correlations ($\hat{\rho}_x$ and $\hat{\rho}_y$). The large sample variance of $\hat{\rho}_{xy}$ can be approximated by $(4n - 3)^{-1}$ to get the FISHER's transformation

$$z_{xy} = \sqrt{(4n-3)} \ln \sqrt{(1+\hat{\rho}_{xy})/(1-\hat{\rho}_{xy})}, \tag{10}$$

whose distribution can also be well approximated by a standard normal distribution (shown in the simulation section). Equation 10 applies without any change for the ANOVA-based estimator $\tilde{\rho}_{xy}$.

Under the assumption of nonrandom association: A general treatment of nonrandom association of alleles within and between loci is difficult, since VNTR loci involve multiple alleles. However, a simple formulation may be provided where nonrandom association is measured by a single parameter, f, signifying excess of homozygotes in the population (ROBERTSON and HILL 1984). In this case, the within locus allele size distribution may be revised as

$$(x_{i1}, x_{i2}) = \begin{cases} (a_k, a_k) \text{ with probability } p_k^2 + fp_k(1 - p_k) \\ \quad \text{for } k = 1, 2, \ldots, r, \\ (a_k, a_l) \text{ with probability } (1 - f)p_k p_l \\ \quad \text{for } k \neq l = 1, 2, \ldots, r. \end{cases}$$

The partitioning of the parameter f for different causes of nonrandom association (such as inbreeding, population structure) are discussed in ROBERTSON and HILL (1984), WEIR and COCKERHAM (1984), but for our purpose, $f \neq 0$ would suffice. Following the derivations of the previous section, we obtain

$$E\left(\sum_{i=1}^{n}\sum_{j=1}^{2} x_{ij}^2\right) = 2n\sum_{k=1}^{r} a_k^2 p_k, \tag{11}$$

$$E\left(\sum_{i=1}^{n} x_{i1}x_{i2}\right) = nf\sum_{k=1}^{r} a_k^2 p_k + n(1-f)\left(\sum_{k=1}^{r} a_k p_k\right)^2, \tag{12}$$

$$E(\bar{x}^2) = \left[(1+f)\sum_{k=1}^{r} a_k^2 p_k + (2n-1-f)\left(\sum_{k=1}^{r} a_k p_k\right)^2\right]/2n, \tag{13}$$

so that the asymptotic expectation of $\hat{\rho}_x$ becomes

$$E(\hat{\rho}_x) \simeq [(2n - 1)f - 1]/[(2n - 1) - f]. \tag{14}$$

Since Equation 14 leads to

$$\frac{1 + E(\hat{\rho}_x)}{1 - E(\hat{\rho}_x)} = \frac{(n-1)(1+f)}{n(1-f)}$$

FISHER's transformation (Equation 9) leads to

$$E(z_x) = \sqrt{(n-2)}\ \ln\sqrt{\frac{1+f}{1-f}} \tag{15}$$

the asymptotic expectation of standardized variable in the presence of nonrandom association of alleles within a locus. Following RAO (1973) and invoking the asymptotic normality of z_x, the power of detection of nonrandom association is given by

$$\begin{aligned} \beta_f &= \Pr(z_x \geqslant z_\alpha \mid f > 0) \\ &= 1 - \Phi\left(z_\alpha - \sqrt{(n-2)}\ \ln\sqrt{\frac{1+f}{1-f}}\right), \end{aligned} \tag{16}$$

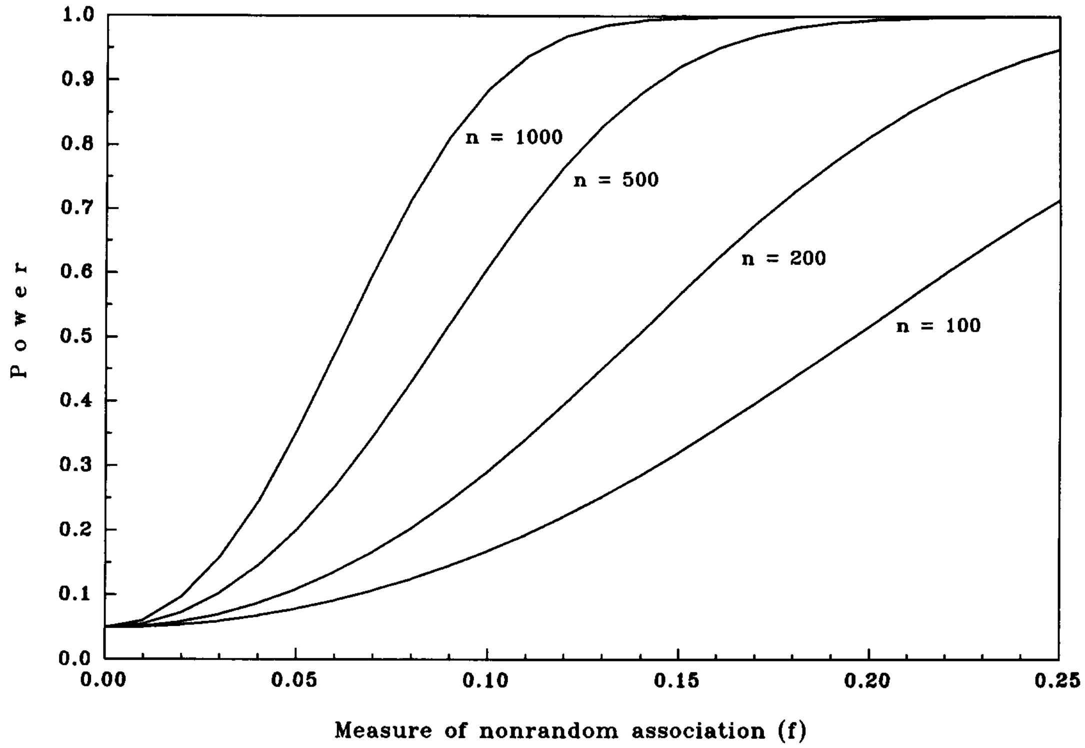

FIGURE 1.—Power of detection of nonrandom association of VNTR fragment sizes within individuals in the presence of population subdivision. The power function (Equation 15) is plotted against the f, generalized index of nonrandom association, for sample sizes $n = 100$, 200, 500 and 1000.

where $\Phi(c)$ is the cumulative probability of a standard normal distribution up to c, and z_α is the upper $100\ (1 - \alpha)\%$ value of a standard normal variate for an α-level test procedure.

Figure 1 shows some numerical calculations for this power function for different sample sizes ($n = 100$, 200, 500 and 1000). It is true that for $n \leq 500$, it would be difficult to detect deviations from the random association yielding $f \leq 0.05$ from a single-locus data, but this limitation of power of detectability is not critical for forensic applications of VNTR data for several reasons. First, since for $f = 0.05$ and $n = 500$, the power (β_f) is approximately 0.3; deviation from random association would remain undetected in data on six loci only with a probability of 12%. Second, $f \neq 0$ may also be generated from the existence of nondetectable alleles of extreme (low or high) size alleles, which is shown to yield an upward bias in DNA profile frequency estimates if the assumption of random association is invoked (CHAKRABORTY *et al.* 1992a). Third, when $f \neq 0$, methods are available (LESLIE 1990) to obtain conservative (upwardly biased) DNA profile frequencies incorporating such small departure from random association of alleles within individuals (see also CHAKRABORTY *et al.* 1992a, b).

A general expression for the non-null expectation of $\hat{\rho}_{xy}$ to examine the association of alleles between loci is complicated, since the two-locus expected genotype frequencies in a population are functions of $r + s$ allele frequencies and $(r - 1)(s - 1)(rs - r - s + 2)/2$ di-, tri-, and quadrigenic disequilibria measures (WEIR 1979). Ignoring terms involving the tri- and quadrigenic disequilibria and approximating up to expressions of order n^{-1}, the asymptotic expectation of $\hat{\rho}_{xy}$ in the presence of non-random association is

$$E(\hat{\rho}_{xy}) \simeq \left(\frac{n-1}{n}\right)\left[\left(1 - \frac{1+f_A}{2n}\right)\left(1 - \frac{1+f_B}{2n}\right)\right]^{-1/2} \cdot \frac{\sum_{k=1}^{r}\sum_{u=1}^{s}\delta_{ku}a_k b_u}{(V_a V_b)^{1/2}} \quad (17)$$

where $V_a = \sum a_{kpk}^2 - (\sum a_k p_k)^2$, f_A and f_B are parameters that represent nonrandom association of alleles within A and B loci and δ_{ku} is the digenic gametic disequilibrium coefficient for alleles A_k and B_u at these loci. Clearly, unlike $\hat{\rho}_x$ and $\hat{\rho}_y$, the non-null expectation of $\hat{\rho}_{xy}$ depends on the intra-locus variances of allele sizes as well as allele frequencies, which makes any power calculation for $\hat{\rho}_{xy}$ difficult. We might, however, surmise that in the context of VNTR allele size distributions since individual allele frequencies are small and variation of allele sizes is large [see *e.g.*, BAIRD *et al.* (1986) and FLINT *et al.* (1989)], the last term of Equation 17 is generally very small, suggesting that even if f_A and f_B are significantly different from zero, the expectation of $\hat{\rho}_{xy}$ will not be substantially different from zero. This result is particularly important, since there are many claims suggesting that factors such as population substructuring that cause intralocus association of alleles should also generate substantial association between alleles of unlinked loci (LANDER 1989; COHEN 1990).

EFFECT OF MEASUREMENT ERRORS ON ASSOCIATION ESTIMATORS

In DNA typing sizing of alleles involve approximations and such measuremental errors obscure the correspondence between actual genotypes and allele size profiles (see also DEVLIN, RISCH and ROEDER 1990). Several quality control experiments suggest that the errors of allele size measurements in Southern blot protocol of DNA typing can be theoretically modelled (BUDOWLE *et al.* 1991a; RISCH and DEVLIN 1992). These studies provide an empirical rationale of the assumption that the measurement errors are distributed with mean zero and standard deviation proportional to the true size, with the constant of proportionality that range generally below 0.025. With this assumption, when an allele of true size X_{ij} is measured as x_{ij}, we can write

$$x_{ij} = X_{ij} + \epsilon_{x_{ij}}, \tag{18}$$

where $\epsilon_{x_{ij}}$ is distributed with zero mean, and variance $c_x X_{ij}^2$. Similarly, $y_{ij} = Y_{ij} + \epsilon_{y_{ij}}$ with $E[\epsilon_{y_{ij}}] = 0$ and $V[\epsilon_{y_{ij}}] = c_y Y_{ij}^2$.

Uncorrelated measurement errors within individuals: First consider the case when $\epsilon_{x_{ij}}$, $\epsilon_{y_{ij}}$ are all mutually independent for all $i = 1, 2, \ldots, n$ and $j = 1, 2$. Under these assumptions note that in the estimation of $\hat{\rho}_x$, the numerator as well as the denominator of Equation 1 can be biased estimators of the respective functions of parameters they intend to measure. In fact, taking expectations over the error distributions we observe

$$E_\epsilon\left(\sum_{i=1}^{n} x_{i1}x_{i2} - n\bar{x}^2\right) = \sum_{i=1}^{n} X_{i1}X_{i2} - n\bar{X}^2 - c_x \sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}^2/4n, \tag{19}$$

$$E_\epsilon\left(\sum_{i=1}^{n}\sum_{j=1}^{2} x_{ij}^2\right) = (1 + c_x)\sum_{i=1}^{n}\sum_{j=1}^{2} X_{ij}^2, \tag{20}$$

$$E_\epsilon\left(\sum_{i=1}^{n}\sum_{j=1}^{2} x_{ij}^2 - 2n\bar{x}^2\right) = \sum_{i=1}^{n}\sum_{j=1}^{2} X_{ij}^2 - 2n\bar{X}^2 + (2n-1)c_x \sum_{i=1}^{n}\sum_{j=1}^{2} X_{ij}^2/2n, \tag{21}$$

so that in the presence of measurement errors we have a revised estimator

$$\hat{\rho}_x^* = \frac{2s_{xx} + d_{x1}\sum_{i=1}^{n}\sum_{j=1}^{2} x_{ij}^2}{s_x^2 - d_{x2}\sum_{i=1}^{n}\sum_{j=1}^{2} x_{ij}^2}, \tag{22}$$

where $d_{x1} = c_x/2n(1 + c_x)$ and $d_{x2} = (2n - 1)c_x/[2n(1 + c_x)]$. Comparison of Equations 2 and 22 indicates that in the presence of measurement errors, Equation 2 underestimates the actual correlation ρ_x. Furthermore, since $E(\hat{\rho}_x^*) = E_X E_\epsilon(\hat{\rho}_x^*) = E_X E_\epsilon(\hat{\rho}_x^* | X)$, $\hat{\rho}_x^*$ has the same asymptotic expectations under the assumption of random as well as non-random associations.

In examining the effect of measurement errors on $\hat{\rho}_{xy}$, the above formulation yields $E_\epsilon[s_{xy}] = s_{XY}$ and hence the revised estimator of $\hat{\rho}_{xy}$ is

$$\hat{\rho}_{xy}^* = \frac{\sum_{i=1}^{n}\sum_{j,k=1}^{2} x_{ij}y_{ik} - 4n\bar{x}\bar{y}}{2[s_X^2 s_Y^2]^{1/2}}, \tag{23}$$

where $s_X^2 = s_x^2 - d_{x2}\sum_{i=1}^{n}\sum_{j=1}^{2} x_{ij}^2$ and $s_Y^2 = s_y^2 - d_{y2}\sum_{i=1}^{n}\sum_{j=1}^{2} y_{ij}^2$. The asymptotic expectations of $\hat{\rho}_x^*$ and $\hat{\rho}_{xy}^*$ remain unaltered under all models of allelic associations. As in the case of $\hat{\rho}_x^*$, comparison of Eqs. (6) and (23) suggests that $\hat{\rho}_{xy}^* > \hat{\rho}_{xy}$, but the effect of measurement errors of $\hat{\rho}_{xy}$ is substantially smaller, shown in the numerical calculations below.

Correlated measurement errors within individuals: There is some evidence that VNTR fragment length variants at a locus within individuals may be correlated due to band shifting and the other technical phenomena associated with sizing of fragment sizes from southern blot gels (SHAPIRO 1991; EVETT 1991; BERRY, EVETT and PINCHIN 1992). In such events, the effect of measurement errors on association estimators can again be modelled as above with the simple change that for two fragment sizes for an individual at a locus the errors of measurement, $\epsilon_{x_{i1}}$ and $\epsilon_{x_{i2}}$ are correlated. With the notations, $E(\epsilon_{x_{i1}}\epsilon_{x_{i2}}) = r_x c_x X_{i1}X_{i2}$ and $E(\epsilon_{y_{i1}}\epsilon_{y_{i2}}) = r_y c_y Y_{i1}Y_{i2}$, for all $i = 1, 2, \ldots n$; and under the assumption of independence across individuals and loci, we have

$$E_\epsilon(\bar{x}^2) = \bar{X}^2 + (c_x/4n^2)\sum_{i=1}^{n}\sum_{j=1}^{2} X_{ij}^2 + (2r_x c_x/4n^2)\sum_{i=1}^{n} X_{i1}X_{i2}$$

$$E_\epsilon\left(\sum_{i=1}^{n} x_{i1}x_{i2}\right) = (1 + r_x c_x)\sum_{i=1}^{n} x_{i1}x_{i2}$$

$$E_\epsilon\left(\sum_{i=1}^{n}\sum_{j=1}^{2} x_{ij}^2\right) = (1 + c_x)\sum_{i=1}^{n}\sum_{j=1}^{2} X_{ij}^2.$$

The revised estimator of intracorrelation becomes

$$\hat{\rho}_x^{**} = \frac{2s_{xx} + d_{x1}\sum_{i=1}^{n}\sum_{j=1}^{2} x_{ij}^2 - d_{x3}\sum_{i=1}^{n} x_{i1}x_{i2}}{s_x^2 - d_{x2}\sum_{i=1}^{n}\sum_{j=1}^{2} x_{ij}^2 + d_{x4}\sum_{i=1}^{n} x_{i1}x_{i2}}, \tag{24}$$

where $d_{x3} = (2n - 1)2r_x c_x/2n(1 + r_x c_x)$ and $d_{x4} = 2r_x c_x/2n(1 + r_x c_x)$. Furthermore, since the errors are uncorrelated across the loci, the revised interclass correlation for a pair of loci becomes

TABLE 1

Intraclass correlation of VNTR fragment sizes within individuals in the Hispanics population of the FBI database for six VNTR loci and their statistical significance

Locus	No. of individuals (n)	$\phi\rho_x(Z_x)$	Levels of significance	
			Simulation	Normal approximation
D2S44	521	0.089 (2.07)	0.04	0.04
D17S79	521	0.035 (0.81)	0.45	0.42
D1S7	517	−0.000 (0.01)	1.00	1.00
D4S139	517	−0.035 (−0.77)	0.43	0.46
D14S13	494	0.070 (1.58)	0.12	0.12
D10S28	440	0.060 (1.27)	0.21	0.20

The levels of significance were determined by 2000 shuffling of all alleles across individuals (simulation) and by using the normal approximate (Equation 8) to detect how often the simulated correlations (absolute value) exceeded the observed (absolute value). The column $\hat{\rho}_x$ represents the observed intraclass correlation (Equation 1) and in parentheses are their normalized values (Equation 8).

$$\hat{\rho}^{**}_{xy} = \frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j,k=1}^{2} x_{ij}y_{ik} - 4n\bar{x}\bar{y}\right) \cdot \left([s_x^2 - d_{x2}\sum_{i=1}^{n}\sum_{j=1}^{2} x_{ij}^2 + d_{x4}\sum_{i=1}^{n} x_{i1}x_{i2}\right)^{-1/2} \cdot \left(s_y^2 - d_{y2}\sum_{i=1}^{n}\sum_{j=1}^{2} y_{ij}^2 + d_{y4}\sum_{i=1}^{n} y_{i1}y_{i2}\right)^{-1/2} \qquad (25)$$

Equations 24 and 25 indicate that when intraindividual measurement errors of two VNTR fragments for a locus are correlated, the original estimators (Equations 2 and 6) may not be systematically biased. However, as shown below the effects of measurement errors, remain small.

DATA ANALYSIS AND SIMULATION RESULTS

WEIR (1992) and BUDOWLE *et al.* (1991a,b) provided details of the blood samples and laboratory protocols which form the basis of DNA typing data gathered by the Forensic Sciences research unit of FBI Academy. For illustrative purpose of the above methods, we consider the pooled Hispanics from Florida and Texas, scored for six loci, to demonstrate the appropriateness of the large sample approximations and the impact of measurement errors. Table 1 shows the observed intraclass correlations and their levels of significance for the six loci. To determine the levels of significance all $2n$ VNTR fragment sizes were randomly shuffled for each locus and new DNA profiles were constructed by pairing the shuffled alleles to generate random observations on $\hat{\rho}_x$ (or its transformed value). Simulated levels of significance designate the empirical probability of how often the absolute value of simulated correlations exceeded that of observed one in 2000 replications. Clearly, the levels of significance observed in simulation and by using the normal approximation are virtually identical. The normal approximation holds for all VNTR loci currently in use in the forensic context, even when their fragment size distributions are complex (*i.e.*, skewed, bi- or multimodal, as in the case of D1S7, D2S44 and D10S28; see binned fragment size distributions in BUDOWLE *et al.* 1991b). Therefore, we conclude that the test of random association of VNTR fragment sizes within individuals for a locus can be done using the normal approximation.

Table 2 shows the results of interclass correlations of VNTR fragment sizes for all pairs of loci in the sample of Hispanics. These computations also show that the random distributions of shuffled allele size associations (intraclass) for each pair of loci can be approximated fairly well with a normal distribution. There is little evidence of fragment size association across pairs of loci, even though the pooled Hispanics sample may actually represent a heterogeneous population (BUDOWLE *et al.* 1991b). This observation completely parallels the findings of WEIR (1992). The only correlation ($\hat{\rho}_x$ = 0.089) which is significant at 5% level is the intraclass correlation for the D2S44 locus. While this could be due to pooling data from a truly heterogeneous population, one should also note that at D2S44 and D17S79 loci the frequencies of nondetectable alleles is appreciable (3% or larger), revealed by RFLP analysis with *Pvu*II and *Pst*I restriction digestion analysis (CHAKRABORTY *et al.* 1992a; R. CHAKRABORTY, unpublished data).

Table 3 shows the effects of measurement errors on the intraclass correlation estimates for the six loci in the sample of Hispanics. In computing $\hat{\rho}_x^*$ (Equation 22) we used c_x = 0.025 since in repeat measurements of the same DNA fragments, size differences are generally below 2.5% for all loci (BUDOWLE *et al.* 1991a; RISCH and DEVLIN 1992). For correlated measurement errors we used r_x = 0.904, since BERRY, EVETT and PINCHIN (1992) estimates that the errors of two VNTR fragment sizes occurring within individuals may be correlated due to factors such as band shifting. These computations show that the inference with regard to the significance of intraclass correlations changed only for the D17S79 locus, which exhibited a non-significant correlation ($\hat{\rho}_x$ = 0.035) when measurement errors were ignored, but $\hat{\rho}_x^* \simeq 0.1$ became significant (at 5% level) once errors of measurement were invoked. As mentioned before, the probable cause of this significant correlation is the presence of non-detectable alleles, although substructuring within the pooled Hispanic sample cannot be ruled out. The last two columns of this table indicate that all $\hat{\rho}_x^*$ values are virtually identical to $\hat{\rho}_x^{**}$, suggesting that the phenomenon of band shifting (resulting in $r_x \neq 0$) does not affect the intraclass correlations once

TABLE 2

Interclass correlation of VNTR fragment sizes in the Hispanics population of the FBI database for six VNTR loci and their statistical significance

			Descriptive statistics of transformed ($\hat{\rho}_{xy}$ z_{xy} of Equation 10)				Levels of significance	
Loci[a]	Sample size (n)	$\hat{\rho}_{xy}$	Mean	Variance	Skewness	Kurtosis	Simulation	Normal approximation
2, 17	499	−0.008	−0.016	1.000	−0.069	3.026	0.71	0.72
2, 1	489	0.038	0.024	1.047	0.051	2.982	0.10	0.10
2, 4	489	0.030	−0.008	0.978	0.082	2.908	0.17	0.18
2, 14	468	−0.026	−0.018	1.002	0.072	2.973	0.24	0.26
2, 10	415	−0.002	0.014	1.010	0.020	2.900	0.91	0.92
17, 1	488	−0.009	0.006	1.002	−0.109	2.970	0.70	0.70
17, 4	490	0.031	0.020	1.056	−0.007	2.897	0.17	0.16
17, 14	465	−0.044	−0.038	0.982	0.011	2.917	0.06	0.06
17, 10	413	−0.037	0.031	0.993	−0.086	2.991	0.12	0.14
1, 4	486	−0.016	−0.020	1.029	−0.011	2.927	0.46	0.46
1, 14	467	−0.037	−0.021	0.981	0.041	2.960	0.10	0.10
1, 10	417	0.025	−0.007	0.981	0.021	2.798	0.30	0.30
4, 14	463	−0.006	0.018	1.015	0.089	3.195	0.78	0.78
4, 10	410	−0.005	0.014	1.002	0.124	2.950	0.84	0.86
14, 10	392	0.009	−0.004	0.999	0.046	3.039	0.69	0.70

[a] 2: D2S44; 17: D17S79; 1: D1S7; 4: D4S139; 14: D14S13; 10: D10S28.

TABLE 3

Effect of measurement errors on intraclass correlation of VNTR fragment sizes for each locus in the pooled Hispanic sample of FBI database

		Intraclass correlation		
Locus	n	ρ_x	ρ_x^*	ρ_x^{**}
D2S44	521	0.089*	0.109*	0.109*
D17S79	521	0.035	0.100*	0.098*
D1S7	517	−0.000	−0.000	−0.000
D4S139	517	−0.035	−0.041	−0.041
D14S13	494	0.070	0.075	0.075
D10S28	440	0.060	0.068	0.068

The corrected intraclass correlations are computed assuming $c_x = 0.025$ for $\hat{\rho}_x^*$ and $r_x = 0.904$ and $c_x = 0.025$ for $\hat{\rho}_x^{**}$.

* $P < 0.05$.

TABLE 4

Effect of measurement errors on interclass correlation of VNTR fragment sizes for each locus in the pooled Hispanic sample of FBI database

		Interclass correlation		
Loci[a]	n	ρ_{xy}	ρ_{xy}^*	ρ_{xy}^{**}
2, 17	499	−0.008	−0.015	−0.015
2, 1	489	0.038	0.044	0.044
2, 4	489	0.030	0.037	0.037
2, 14	468	−0.026	−0.031	−0.031
2, 10	415	−0.002	−0.003	−0.003
17, 1	488	−0.009	−0.017	−0.017
17, 4	490	0.031	0.057	0.057
17, 14	465	−0.044	−0.076	−0.076
17, 10	413	−0.037	−0.064	−0.064
1, 4	486	−0.016	−0.019	−0.019
1, 14	467	−0.037	−0.040	−0.040
1, 10	417	0.025	0.028	0.028
4, 14	463	−0.006	−0.007	−0.007
4, 10	410	−0.005	−0.006	−0.006
14, 10	392	0.009	0.011	0.011

The corrected interclass correlations are computed assuming $c_x = 0.025 = c_y$ for $\hat{\rho}_{xy}^*$ and $r_x = 0.904 = r_y$ and $c_x = 0.025 = c_y$ for $\hat{\rho}_{xy}^{**}$.

[a] 2: D2S44; 17: D17S79; 1: D1S7; 4: D4S139; 14: D14S13; 10: D10S28.

the errors of measurement for each individual VNTR fragment length are invoked in estimation.

In contrast, the interclass correlations of VNTR fragment sizes for all pairs of loci are relatively unaffected by measurement errors. This is shown in Table 4, where again $c_x = 0.025 = c_y$ and $r_x = 0.904 = r_y$ were used. None of the correlations changed substantially to affect the inference that no evidence of nonrandom association of fragment sizes across loci exist even in the pooled Hispanic sample.

DISCUSSION AND CONCLUSIONS

Although the methodology developed in this work is almost parallel to the previous work of BERRY, EVETT and PINCHIN (1992) and WEIR (1992), there are several distinctive features of the present analysis.

First, the near equivalence of KARLIN's generalized correlations (KARLIN, CAMERON and WILLIAMS 1981) and the ANOVA-based measures of correlations establishes that the conclusions reached by WEIR (1992) are not affected by the complexity of the VNTR fragment size distributions, since WEIR's tests of significance results are not based on any distributional assumptions. Second, our analytical study and computer simulation indicate that the correlation approach has a substantial power of detectability of nonrandom association, particularly when multiple loci are available for a sampled population. Third, the appropriateness of large sample approximations for the transformed correlations shown in this work remains valid even when the sample size (n) is reduced to 100. This avoids extensive computer simulations for judging the significance of any observed correlation in a given data. Even for simulation, a simple shuffling would be more appropriate as, unlike jackknifing or bootstrapping, it does not disturb the fragment size distribution.

One might argue that the demonstration of the lack of significant correlations in individuals' VNTR fragment sizes does not necessarily ensure the validity of the forensic calculations of DNA profile frequency, since the calculations involve grouped data (binned fragment size frequency) and not the raw data on the individual fragment frequencies. This concern is not serious, because for grouped data one might consider exactly the same analysis, using group (bin) averages of VNTR fragment sizes weighted by the frequencies of each group and perform a correlation analysis such as the one suggested here. Of course, for grouped data a likelihood ratio test criterion of dependence may also be adopted (WEIR 1992) where significance should be assessed by permutation-based simulations. GUO and THOMPSON (1992) suggested an alternative exact test procedure for determining the levels of significance of such likelihood-based test criteria which takes into account infrequent bin frequencies in any observed data. Application of their method to the present data leads to qualitative conclusions similar to the ones reported here.

In summary, we showed that the conclusions reached by WEIR (1992) remain virtually the same, even when more robust estimators of correlations are used, and the measurement error of fragment sizes is taken into account in data analysis. In particular, we showed that there is little evidence of non-random association of VNTR fragment sizes within individuals even when the sample constitutes individuals that are of mixed ancestry (such as the Hispanics). Conservative (upwardly biased) estimates of DNA profile frequencies may, therefore, be obtained employing the chain multiplication rule using the frequencies of grouped (binned) fragment sizes for each locus.

We thank B. BUDOWLE and his laboratory staff for providing the data for analysis reported in this paper. Comments and suggestions of S. P. DAIGER and E. BOERWINKLE are greatly appreciated. This work is supported by U.S. Public Health Service research grants NIH-GM-41399 and NIJ-90-CX-0038.

LITERATURE CITED

BALAZS, I., M. BAIRD, M. CLYNE and E. MEADE, 1989 Human population studies of five hypervariable DNA loci. Am. J. Hum. Genet. **44:** 182–190.

BALLANTYNE, J., G. SENSABAUGH and J. WITKOWSKI, 1989 DNA Technology and Forensic Science. Banbury Report 32, New York.

BAIRD, M., I. BALAZS, A. GIUSTI, L. MIYAZAKI, L. NICHOLAS, K. WEXLER, E. KANTER, J. GLASSBERG, F. ALLEN and P. RUBINSTEIN, 1986 Allele frequency distribution of two highly polymorphic DNA sequences in three ethnic groups and its application to the determination of paternity. Am. J. Hum. Genet. **39:** 489–501.

BERRY, D. A., I. W. EVETT and R. PINCHIN, 1992 Statistical inference in crime investigations using deoxyribonucleic acid profiling. Appl. Statist. **41:** 499–531.

BUDOWLE, B., A. M. GIUSTI, J. S. WAYE, F. S. BAETCHEL, R. M. FOURNEY, D. E. ADAMS, L. A. PRESLEY, H. A. DEADMAN and K. L. MONSON, 1991a 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., K. L. MONSON, K. S. ANOE, D. L. BAECHTEL, D. L. BERGMAN, E. BUEL, P. A. CAMPBELL, M. E. CLEMENT, H. W. COEY, L. A. DAVIS, A. DIXON, P. FISH, A. M. GIUSTI, T. L. GRANT, T. M. GRONERT, D. M. HOOVER, L. JANKOWSKI, A. J. KILGORE, W. KIMOTO, W. H. LANDRUM, H. LEONE, C. R. LONGWELL, D. C. MACLAREN, L. E. MEDLIN, S. D. NARVESON, M. L. PIARSON, J. M. POLLOCK, R. J. RAQUEL, J. M. REZNICEK, G. S. ROGERS, J. E. SMERICK and R. M. THOMPSON, 1991b A preliminary report on binned general population data on six VNTR loci in Caucasians, Blacks and Hispanics from the United States. Crime Laboratory Digest **18:** 9–26.

BURKE, T., G. DOLF, A. J. JEFFREYS and R. WOLFF, 1991 *DNA Fingerprinting: Approaches and Applications*. Birkhäuser Verlag, Basel, Switzerland.

CHAKRABORTY, R., 1992 Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. Hum. Biol. **64:** 141–159.

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

CHAKRABORTY, R., M. R. SRINIVASAN, L. JIN and M. DE ANDRADE, 1992b Effects of population subdivision and allele frequency differences on the interpretation of DNA typing data for human identification. In: Proceedings of Third International Symposium on Human Identification (Promega Corporation, Madison, WI) (in press).

COHEN, J. E., 1990 DNA fingerprinting for forensic identification: potential effects on data interpretation of subpopulation of heterogeneity and band number variability. Am. J. Hum. Genet. **46:** 358–368.

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

EVETT, I. W., 1991 Trivial error. Nature **354:** 114.

FISHER, R. A., 1928 *Statistical Methods for Research Workers*. Oliver & Boyd, London.

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

GUO, S. W., and E. A. THOMPSON, 1992 Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics **48:** 361–372.

JEFFREYS, A. J., V. WILSON and S. L. THEIN, 1985 Individual specific "fingerprints" of human DNA. Nature **316:** 76–79.

KARLIN, S., E. C. CAMERON and P. T. WILLIAMS, 1981 Sibling and parent-offspring correlation estimation with variable family size. Proc. Natl. Acad. Sci. USA **78:** 2664–2668.

LANDER, E. S., 1989 DNA fingerprinting on trial. Nature **339:** 501–505.

LESLIE, P. W., 1990 Demographic behavior, mating patterns, and the distribution of inbreeding, pp. 63–79 in *Convergent Issues in Genetics and Demography*, edited by J. ADAMS, A. HERMALIN,

D. LAM and P. SMOUSE. Oxford University Press, New York.

NAKAMURA, Y., M. LEPPERT, P. O'CONNELL, R. WOLFF, T. HOLM, M. CULVER, C. MARTIN, E. FUJIMOTO, M. HOFF, E. KUMLIN and R. L. WHITE, 1987 Variable number of tandem repeat (VNTR) markers for human gene mapping. Science **235:** 1616–1622.

RAO, C. R., 1973 *Linear Statistical Inference and Its Applications.* John Wiley & Sons, New York.

ROBERTSON, A., and W. G. HILL, 1984 Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. Genetics **107:** 703–718.

RISCH, N., and B. DEVLIN, 1992 On the probability of matching DNA fingerprintings. Science **255:** 717–720.

SHAPIRO, M. M., 1991 Imprints on DNA fingerprints. Nature **353:** 121–122.

SRIVASTAVA, M. S., 1984 Estimation of interclass correlations in familial data. Biometrika **71:** 177–85.

WEIR, B. S., 1979 Inferences about linkage disequilibrium. Biometrics **25:** 235–254.

WEIR, B. S., 1992 Independence of VNTR alleles defined as fixed bins. Genetics **130:** 873–887.

WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating F-statistics for the analysis of population structure. Evolution **38:** 1358–1370.

Communicating editor: B. S. WEIR