*Statistical Science*
1994, Vol. 9, No. 2, 222–278

# DNA Fingerprinting: A Review of the Controversy

**Kathryn Roeder**

*Abstract.* Forensic scientists have used genetic material (DNA) as evidence in criminal cases such as rape and murder since the middle of the last decade. The forensic scientist's interpretation of the evidence, however, has been subject to some criticism, especially when it involves statistical issues (including relevant areas of population genetics in the realm of statistics). These issues include the appropriate method of summarizing data subject to measurement error, independence of events in a DNA pattern or profile; characterization of heterogeneity of populations; appropriate sampling methods to develop reference databases; and probabilistic evaluation of evidence under uncertainty of appropriate reference database. I review these issues, with the goal of making them accessible to the statistical community. My thesis in this article is that, for most cases, the tremendous genetic variability among individuals obviates concern arising from minor violations of modeling assumptions.

*Key words and phrases:* DNA profiling, forensic inference, population genetics, Hardy–Weinberg equilibrium, legal inference, population heterogeneity.

## 1. INTRODUCTION

*DNA fingerprinting* or *DNA profiling* are terms used to describe molecular techniques that have been employed by forensic scientists to draw inferences from bodily fluids and other materials found at crime scenes. The culpability of a suspect is based, in part, on the similarity of his DNA profile to that obtained at the scene of the crime. Berry (1991) introduced some of the statistical aspects of this topic to readers of *Statistical Science*. In his article he presented a new method for assessing the weight of the evidence and compared this method to those currently used in the courts. He also alluded to the controversy surrounding the subject at that time. In this article the history of the controversy will be reviewed, with emphasis on the confounding features of the data and the pertinent statistical issues.

One of the earliest uses of this technique occurred during 1987 in Narborough, England (Shapiro, 1991). A seventeen-year-old boy was accused of the rape and murder of two girls, one in 1983 and the other in 1987. Circumstantial evidence and his own previous history indicated that he could have been the murderer. During his confession he requested a blood test. By coincidence, the laboratory of noted molecular biologist Alec Jeffreys was located within 6 miles of the village. Using molecular methods pioneered by Jeffreys, Wilson and Thein (1985a, b), the suspect's DNA was compared to the sperm samples found on the victims. It was concluded that both girls were raped by the same individual, but the boy was not the perpetrator.

Such DNA profiles can be obtained from any material that contains nucleated cells: blood, semen, skin, hair roots and so on. A DNA profile consists of a set of measurements of discrete random variables, measured with error. The random variables occur in pairs, one inherited from each parent. Typically three to five different pairs of random variables are measured. Forensic scientists declare a *match* if the profiles from two samples are considered to be sufficiently similar. Two samples from the same individual differ only due to measurement error, while samples from different individuals rarely match.

If the samples from the suspect and the crime scene (*evidentiary sample*) do not meet the match criterion, then the case does not go to trial unless the nonmatch is immaterial in light of other evidence. Although there is a possibility of false exclusions, this has not been the focus of current debate. The controversy arises when the samples do

*Kathryn Roeder is Associate Professor of Statistics at Yale University, Box 208290, Yale Station, New Haven, Connecticut 08520.*

match and some measure of the weight of this evidence is presented to the jury. Typically they are presented with an estimate of the probability of obtaining a matching profile from a randomly selected individual from some appropriately selected population, called the *reference population*. This profile probability is usually obtained by multiplying the estimated matching probability for each component of the profile—that is, by assuming independence of the observations composing the profile. The estimates of the probability of matching for each component of the profile are based on the distribution of profiles of individuals that have been collected by forensic testing laboratories. Samples are available for each of the major ethnic groups (races).

The first flames of the controversy were ignited by the claim that there were likely to be gross violations of the assumption of independence (Lander, 1989; Cohen, 1990). Subsequent studies of the data by numerous authors refuted these initial claims (Devlin, Risch and Roeder, 1990, 1991a; Chakraborty and Kidd, 1991; Chakraborty and Jin, 1992; Weir, 1992a, b), yet the controversy did not completely subside.

Another key point of contention concerned the appropriate choice of the reference population (Weir and Evett, 1992; Lewontin, 1993), as well as the adequacy of the samples of the reference populations. These databases do not constitute stratified random samples of the populations of interest, but rather are convenience samples (Geisser, 1992). It has been argued that individuals tend to marry within their own region, religion and ethnic group and consequently the general population consists of subpopulations of individuals with radically different profile probabilities. The argument leads to the conclusion that some apparently rare profiles might be common in the context of the proper subpopulation (Lewontin and Hartl, 1991). This argument seems incompatible, as many authors have noted, with the observation that profile probabilities do not differ substantially across major ethnic groups (Evett, 1992a; Weir, 1992a, b) and differ even less across subpopulations within an ethnic group (Devlin and Risch 1992b; Devlin, Risch and Roeder, 1994).

A study sponsored by the National Research Council (NRC, 1992) to examine these issues, and presumably to quell the controversy, has instead fanned the flames, drawing criticism from those who generally support the forensic scientists' interpretation of DNA evidence: yet the NRC report endorsed the forensic use of DNA profiling. Indeed a flurry of articles have appeared critiquing the NRC report (Budowle and Monson, 1992; Cohen, 1992; Weir, 1992c, 1993; Balazs, 1993; Devlin, Risch and Roeder, 1993a; Evett, Scranage and Pinchin, 1993; Morton, Collins and Balazs, 1993).

Although many of the arguments put forth by the critics of current methods of evaluating DNA evidence are theoretically correct, my conclusions are that the data do not support their claims. The concern about both independence of the components of the profile and the choice of reference population are based on the assumption of extreme *population heterogeneity*. Such extreme heterogeneity could only occur if individuals within a population (say, Caucasians) tend to marry other individuals from the same subpopulation (say, Irish, Italian, German, etc.) and these subpopulations have very different profile probabilities; population geneticists would describe the situation as extreme population substructure. There is little doubt that some population heterogeneity exists, although it is not extreme for the major ethnic groups constituting the U.S. population. Moreover, because there is such a tremendous amount of variability among DNA profiles within a subpopulation, this heterogeneity is of little practical import in most cases. Claims to the contrary, supported by apparent deviations of model assumptions, have been based on incorrect analyses of the data, and hypothetical conjectures of gross violations have not been supported by empirical evidence.

In Section 2 the data are described, and a simple probability model for DNA profiles and the associated measurement errors is presented. From the probability model, methods for summarizing the weight of the evidence are first developed in Section 2 and expanded in Section 6. Due to measurement error, a profile consists of continuous measurements of discrete characters. A discrete probability model will be the focus of this article because the methods currently in use discretize the data. Methods for summarizing the data without discretizing it are discussed in Section 6.

In Section 3 the genetic model (heterogeneity among populations) that is most commonly assumed to be the basis for violations of independence is presented, and tests of independence are reviewed. Heterogeneity among populations is also the reason for concern about the choice of reference population and the quality of the existing nonrandom samples. The appropriate reference population to evaluate the evidence for a given crime is discussed in Section 4. Statistical methods that allow for various levels of relatedness between the suspect and donor of the evidentiary sample are also presented in Section 4. This discussion covers corrections that account for population heterogeneity. The adequacy of standard reference populations is discussed in Section 5. In Section 7, issues surrounding some open questions are outlined. Finally, in Section 8, I contrast my conclusions with those drawn by the NRC panel.

## 2. DNA PROFILES: THE DATA AND THE EVIDENCE

A *VNTR* (variable number of tandem repeats) *locus*, the genetic marker presently favored by forensic scientists, is a specific location of the genome where a core sequence of nucleotides is repeated in tandem numerous times (Wyman and White, 1980; Baird et al., 1986). DNA from a VNTR locus of two randomly chosen individuals generally differ in the number of these repeats, and hence in length. For each *locus* (location), the data on VNTR length are obtained by a sequence of molecular techniques. First, the DNA is cut into smaller pieces using a restriction enzyme that cuts outside (usually) the VNTR region. Next, fragments are separated by size using gel electrophoresis. The fragments are then denatured into single strands and blotted onto a membrane. A radioactive probe, designed to attach to the core sequence (the repeating segment), is applied. When the gel is exposed to an X-ray film, the radioactive sites appear as dark bands (Figure 1). The length of each DNA fragment is inferred from the distance the DNA has traveled on the gel. The genetic markers discussed are designed to yield a pair of fragments (one inherited from each parent) at each locus.

Consider the pair of measurements obtained at a given locus. Because of the large number of *alleles* (distinct types or lengths of DNA) segregating in the population, most individuals are *heterozygous* (two distinct alleles at the locus) and generate bands of two distinct lengths (Figure 1a). A single band is sometimes generated at a locus (Figure 1b); single-banded patterns may be due to *homozygosity* (two copies of the same allele), to difficulties in distinguishing fragments of similar lengths or to an allele too small to be measured. See Devlin, Risch and Roeder (1990) and Devlin and Risch (1992a) for statistical analysis and Steinberger, Thompson and Hartmann (1993) for molecular demonstration of these phenomena.

This section describes the independence model for DNA profiles and reviews the calculations which summarize the weight of the evidence presented to a jury. In Section 2.1 a model is presented for a simpler case in which the DNA profile can be observed without measurement error. The weight of the evidence is summarized in the form of a likelihood ratio. In Section 2.2 measurement error is introduced. With measurement error, the discrete DNA profile is transformed to a continuous set of random variables. In an effort to mimic the methods developed for discrete data, a method was developed, called match/binning, for discretizing continuous DNA profiles. This method is presented in Section 2.3. Finally, Section 2.4 reviews some of the recommendations made by the NRC committee concerning the probability calculations laid out in this section.

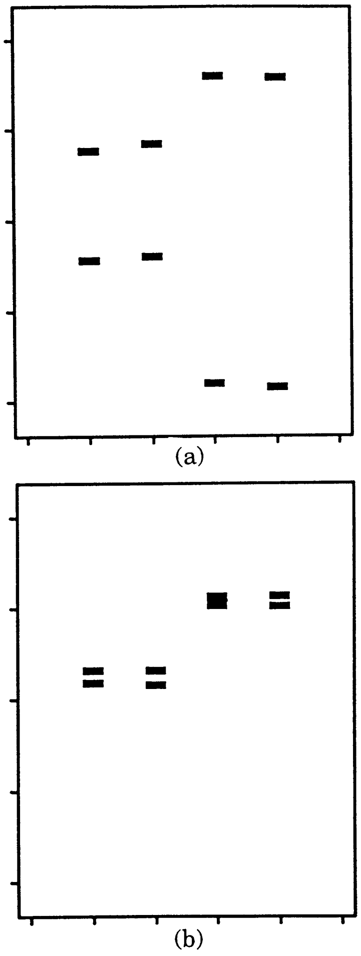


FIG. 1. *Schematic of autoradiographs of two loci (top, bottom): the first and second lanes (columns) on each autoradiograph are the suspect and evidentiary samples; the third and fourth lanes are victim samples. In the second autoradiograph's third lane, the two bands blurred together or coalesced.*

### 2.1 A Probability Model for Discrete Data

A *multilocus genotype* $\mathcal{G}$ for a particular individual consists of unordered pairs of fragments from each of $L$ loci:

$$\mathcal{G} = \{(A_1, A_2)_\ell, \ell = 1, \ldots, L\}.$$

At each locus, $A_i$ is a discrete random variable that takes on values, (alleles) $\{a(k), k = 1, \ldots, m\}$ with probabilities $\{\gamma(k), k = 1, \ldots, m\}$ that depend on $\ell$ and the population of interest. Provided that $A_1$ is independent of $A_2$ (*Hardy–Weinberg equilibrium*, hereafter H-W), the probability of observing the single-locus genotype $\{a(i), a(j)\}$ can be calculated as

$$\Pr\left(\{a(i),a(j)\}\right) = \begin{cases} 2\gamma(i)\gamma(j), & i \neq j \text{ (heterozygotes)}, \\ \gamma(i)^2, & i = j \text{ (homozygotes)}. \end{cases} \tag{1}$$

Furthermore, provided the genotypes are independent across loci (*linkage equilibrium*), then the multilocus genotype probability can be obtained by multiplying across loci. The assumption of independence is a key point of contention, which will be discussed further in Section 3.

Assume the genetic evidence consists of the multilocus genotype $\mathcal{G}_s$ obtained from the suspect and that obtained from the evidentiary sample $\mathcal{G}_e$. The objective is to distinguish between two competing hypotheses:

$H_0$: the samples were obtained from different individuals;

$H_1$: the suspect and evidentiary samples were obtained from the same individual.

Notice that neither of these complementary hypotheses contains an evaluation of guilt. In fact, formulations involving guilt and innocence are misleading: like dermal fingerprints, a DNA profile, even if unique, can only place the suspect's DNA at the crime scene.

Under $H_1$, because the same person was the donor for both the suspect and evidentiary sample, $\Pr(\mathcal{G}_s = \mathcal{G}_e) = 1$. Clearly, the calculation of genotype probabilities is only of interest when $\mathcal{G}_s = \mathcal{G}_e$. Let us use $\mathcal{G}$ to denote this matching genotype and let $\Pr(\mathcal{G})$ denote the probability of observing the multilocus genotype $\mathcal{G}$ in a random draw from a hypothetical population of potential donors of the evidentiary sample. This population is known as the reference population. Under $H_0$, assuming that the suspect and evidentiary samples are independently drawn from the reference population, $\Pr(\mathcal{G}_s = \mathcal{G}_e = \mathcal{G}) = \Pr(\mathcal{G})^2$. Under these model assumptions, the evidence for a crime can be summarized in the following likelihood ratio:

$$\mathcal{LR} = \frac{\Pr\left(\mathcal{G}_s, \mathcal{G}_e \mid H_1\right)}{\Pr\left(\mathcal{G}_s, \mathcal{G}_e \mid H_0\right)} = \begin{cases} \dfrac{\Pr(\mathcal{G}) \times 1}{\Pr(\mathcal{G})^2} = \dfrac{1}{\Pr(\mathcal{G})}, & \text{if } \mathcal{G}_s = \mathcal{G}_e = \mathcal{G}, \\ 0, & \text{if } \mathcal{G}_s \neq \mathcal{G}_e. \end{cases} \tag{2}$$

This calculation is based on modeling assumptions that have been the focus of some debate. These issues will be discussed in Section 4. The choice of reference population is probably the most controversial. At this point, it suffices to note that all of the major testing laboratories possess large databases that are used as reference populations. These samples have been obtained, for example, from the men and women tested in disputed paternity cases (see Section 5.2). For any particular crime the objective is to choose a database that is representative of the genotypes that we would expect to find in the pool of potential donors of the evidentiary sample. For example, if a crime occurred on a Navajo reservation and the particulars of the crime indicate the perpetrator was Navajo, then a suitable reference population would be a random sample of Navajos.

It has also been argued that frequently it is not reasonable to assume that the suspect and evidentiary samples are random draws from the reference population, under $H_0$. For such cases the model formulation can be expanded to allow the evidentiary sample and suspect sample to be drawn from closely related individuals or individuals from a cluster in the population.

Suppose $\mathcal{LR} = 10^6$; typically $1/\mathcal{LR} = \Pr(\mathcal{G})$ is presented to the jury and interpreted as the probability of observing $\mathcal{G}$ in a random draw from the population. This presentation can be confusing because the value of $\mathcal{LR}$ might be considerably larger than the number of people in the set of potential donors of the evidentiary sample. The appropriate interpretation is not as a probability at all, but as a likelihood ratio: the evidence is a million times more likely to have arisen if the crime scene material were left by the defendant than if it were left by some unrelated person. Another way to interpret the data is via culprit-based sampling. Let us condition on the suspect's genotype, $\mathcal{G}_s$. Under $H_1$, if $\mathcal{G}_s = \mathcal{G}_e = \mathcal{G}$, then the likelihood of observing $\mathcal{G}$ is 1; while under $H_0$, the evidentiary sample is obtained as a random draw from a hypothetical reference population (excluding the suspect) and the likelihood of observing a matching genotype is $\Pr(\mathcal{G})$: $1/\mathcal{LR} = \Pr(\mathcal{G}_e \mid \mathcal{G}_s, H_0) = \Pr(\mathcal{G})$, provided $\mathcal{G}_e = \mathcal{G}_s = \mathcal{G}$. Once again, this argument has nothing to do with the number of potential donors of the evidentiary sample. If one were to correct for a finite sample, then a small pool of potential perpetrators increases the probability of $H_1$. Moreover, if DNA profiles were unique, then $\Pr(\mathcal{G}) = 0$ for any-sized reference population, and $\mathcal{LR} = \infty$.

Multiplying $\mathcal{LR}$ by the prior odds of $H_1$ (assessed on the basis of all the pertinent evidence apart from the DNA profile) yields the posterior odds that the same individual was measured twice,

$$\text{odds}(H_1) = \frac{\Pr(H_1)}{\Pr(H_0)}\mathcal{LR}. \tag{3}$$

The presentation of posterior odds in the courts was argued for by Berry (1991) and cautioned against by

Kaye (1991), who notes that outside of civil suits, Bayesian methods have not gained much of a foothold in the courts (see also Geisser, 1990). For a review of the legal issues see Kaye (1988, 1991, 1993).

### 2.2 A Probability Model for Continuous Data

With conventional genetic markers (e.g., blood groups, serum proteins), the laboratory techniques are accurate enough that observations can usually be classified to distinct alleles and (2) is presented as the weight of the evidence for $H_1$.

Although a VNTR allele can be thought of as a categorical random variable, distinguished by the number of tandemly repeating core sequences, it is usually thought of as a discrete quantitative measurement, defined by its length. Observations of a pair of VNTR alleles are obtained by indirectly measuring the lengths of a pair of alleles, using gel electrophoresis and autoradiography. For VNTR's, because there is an enormous number of alleles (typically $m > 100$), even a small amount of measurement error makes it impossible to directly classify the fragments into alleles (Balazs et al., 1989); that is, the unobservable pair of fragments $(A_1, A_2)$, which are discrete random variables, are measured with error, yielding observable pairs of measurements of alleles $(X_1, X_2)$ at locus $\ell$ for a randomly chosen individual:

$$X_1 = A_1 + \varepsilon_1,\qquad X_2 = A_2 + \varepsilon_2, \tag{4}$$

where $\varepsilon_1$and$\varepsilon_2$ are the measurement errors. The random error $\varepsilon$ is a function of the allele size $A$. Given $A = a(j)$, it is commonly assumed that $\varepsilon$ is distributed $\mathcal{N}(0, \sigma_j^2)$, where $\sigma_j = c \times a(j)$. From experimental data, the normality assumption appears reasonable. The value of $c$ is determined experimentally and is known to be laboratory dependent (Devlin, Risch and Roeder, 1991b; Berry, Evett and Pinchin, 1992). In the laboratories involved in DNA testing, however, the standard error $\sigma_j$ is of the same magnitude as the difference between adjacent alleles, $|a(j) - a(j-1)|$.

While the alleles $(A_1, A_2)$ are commonly assumed to be independent, the molecular methodology causes the pairs of measurement errors $(\varepsilon_1, \varepsilon_2)$ to be correlated (Figure 2). This correlation, which is a function of the allele sizes and percent difference, can be estimated from repeated measurements of the same alleles. In brief, the correlation function can be described as follows: it is 1.0 for $A_1 = A_2$, ensuring that $X_1 = X_2$ (a homozygote always appears as a single band on an autoradiograph, and the forensic scientist records the same measurement for both alleles); the correlation decreases as $|A_1 - A_2|$ increases, but this decrease is slower for larger fragment pairs than for smaller fragment pairs. [See Devlin, Risch and Roeder (1992) for more details.]

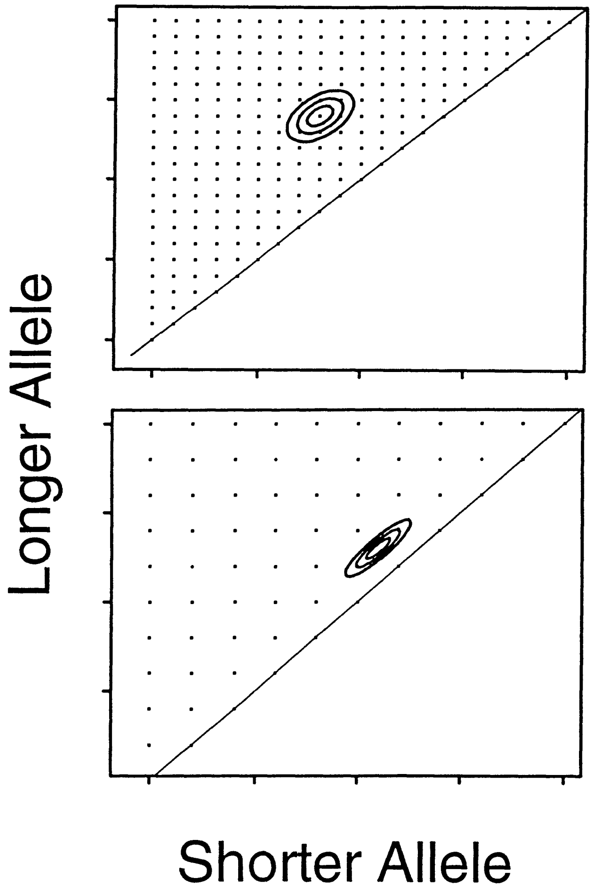


FIG. 2. *A portion of the two-dimensional space of allele pairs (grid) for two loci (top, bottom) with different measurement error relative to the allele spacing: the bivariate normal distribution of measurement error for a given allele pair is superimposed on the space of possible single-locus genotypes. Note the increased correlation in allele pairs of similar size.*

Given $(A_1, A_2) = (a(i), a(j))$, $(X_1, X_2)$ are assumed to follow a bivariate normal distribution with mean $(a(i), a(j))$ and variance-covariance matrix determined by the mean. The corresponding density evaluated at $(x_1, x_2)$ will be denoted by $\phi_{ij}(x_1, x_2)$. (Note that when $A_1 = A_2$, the distribution degenerates to a univariate density.) Let $\gamma(i, j)$ denote the probability of sampling the genotype $\{a(i), a(j)\}$; under H-W, $\gamma(i, j)$ is given in (1). It follows that the marginal distribution of $(X_1, X_2)$ is that of a normal mixture with density

$$f(x_1, x_2) = \sum_{i \le j} \gamma(i, j)\, \phi_{ij}(x_1, x_2). \tag{5}$$

Because fragments are visualized on an autoradiograph, the corresponding bands have substantial width. This introduces a complication if $A_1$ and $A_2$ are very similar but not identical in length. In this case, the bands may be indistinguishable because they blur together and only one length $z = (x_1 + x_2)/2$

is observed. This phenomenon is called *coalescence*. The probability of this event is a function $\delta$ of the mean $z$ and the difference $t = |x_1 - x_2|/2$ of the allele sizes. The coalescence probability $\delta(t,z)$ can be estimated from the data (Devlin, Risch and Roeder, 1990). Coalescence will not be discussed in detail here. The interested reader is referred to Devlin, Risch and Roeder (1990, 1991a, 1992).

Figure 3 illustrates the estimate of the allele distribution $\{\gamma(k), k = 1, \ldots, m\}$ for two loci commonly used by forensic laboratories. (These allele distributions are estimated from Lifecodes Corporation's Caucasian database; Lifecodes is an independent testing laboratory. See Section 6 for further exposition on the methods used to obtain the estimate.) Due to innate properties of these two VNTR loci (Devlin, Risch and Roeder, 1991b), the distribution on the top (D17S79) is estimated with substantial accuracy, while the estimator on the bottom (D2S44) has substantial variance. Notice that D17S79 is not especially informative: although it possesses about 53 alleles with $\gamma(k) > 0$, four of these alleles make up about 60% of the probability. D2S44 is much more informative: it possess about 172 alleles with $\gamma(k) > 0$, and none of them dominates the distribution. Furthermore, the estimated distributions do not differ substantially by ethnic group (race) (Devlin and Risch, 1992b).

The idea of using a likelihood ratio for continuous measurements to assess the weight of evidence in a forensic setting was first proposed by Lindley (1977). In this article, a brief review of methods for summarizing the evidence without discretizing the data is presented in Section 6. For an extensive review of methods used to summarize genetic information for both paternity and criminal cases, see Devlin (1993).

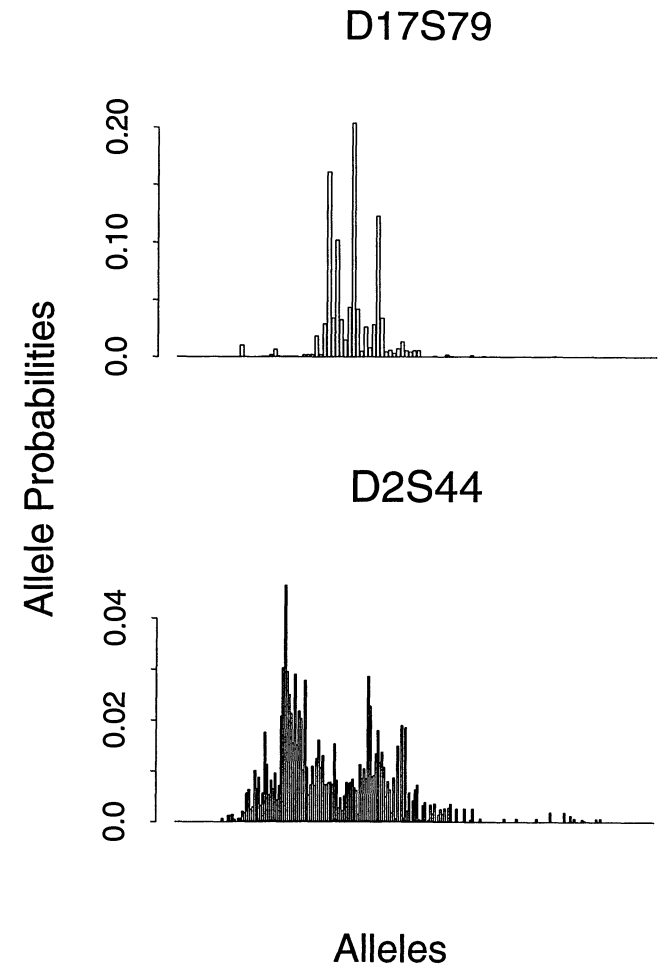


FIG. 3. *The allele distribution estimates of the* D17S79 *locus* (*top*) *and the* D2S44 *locus* (*bottom*): *the probabilities of* 98 *alleles for* D17S79 *and* 329 *for* D2S44 *are estimated. The horizontal axis is the number of repeating units.*

### 2.3 Match/Binning Methods

To apply (2) to discrete random variables, we first determine if $\mathcal{G}_e = \mathcal{G}_s$; if so, we calculate the probability of observing this genotype in the reference population. There is no obvious extension of this simple method for the continuous VNTR data. Because forensic scientists and the legal community have decades of experience with discrete genetic markers, however, they created a method that mimicked the discrete approach.

To apply analogues of (2) to continuous random variables, it is necessary to construct a rule that classifies a pair of profiles as potential observations of the same genotype (a *match*) or not (an *exclusion*). When a match is declared, the probability of observing a match is calculated using a method called binning. These methods are known as *match/binning* methods because of the two stages of the procedure. To illustrate an approach used in the courts (prior to the NRC report) let us consider the method used by Lifecodes Corporation. If the DNA of the suspect and evidentiary samples appear sufficiently similar, the samples are declared a match. Roughly speaking, if any fragments are separated by more than three standard deviations (SD) of the measurement error, the profiles are declared nonmatching, $\mathcal{LR} = 0$ and the suspect is excluded as a possible donor of the evidentiary sample. However, because the measurement errors are known to have positive correlations, the forensic experts exercise some judgment in their declarations, possibly declaring near misses as matches (which may or may not be accepted by the court) and ruling out apparent matches due to unlikely patterns of measurement error. (The term "match" is actually a convenient misnomer; a match is actually a failure to exclude. On rare occasions a locus is declared to be inconclusive.)

Because the standard error is proportional to the length of the allele, one frequently compares band sizes to evaluate "matching" in terms of a percent error. For example, the Lifecodes match criterion (3 SD) is referred to as a 1.8% match window because their standard deviation is $\sigma_y = 0.006y$. Confusion may occur because an equivalent window for the Forensic Science Service (a testing laboratory in Great Britain) would require a somewhat larger match window because their laboratory techniques lead to larger errors.

A *floating bin* method is used to obtain "allele" probabilities. This probability is estimated by the proportion of fragments in the reference population $\{z_{1j}, z_{2j}, j = 1, \ldots, r\}$ within three standard deviations of the evidentiary sample:

$$\widehat{\gamma}(y) = \frac{1}{2r}\sum_{j=1}^{r}\sum_{i=1}^{2}\mathbf{I}\{|y - z_{ij}| < 3\sigma_y\}.$$

The multilocus probability is obtained by multiplication within a locus and between loci with one exception. [A single-banded pattern may be due to homozygosity, to difficulties in distinguishing fragments of similar lengths or to an allele too small to be measured. The second problem does not occur when certain laboratory techniques are used. To account for the two physical artifacts which may have occurred, the following adjustment is made: replace $\widehat{\gamma}(y)^2$ with $2\widehat{\gamma}(y)$ or $2\widehat{\gamma}(y)^2$ if the methodology employed in the laboratory makes the second technical difficulty unlikely. As mentioned in Berry (1991), this adjustment more than compensates for the disturbance.] Finally, mimicking (2), $\mathcal{LR}$ is calculated.

A competing match/binning technique relies on a *fixed bin* (Budowle et al., 1991b) or histogram approach to calculate $\mathcal{LR}$. The histogram is formed based on arbitrary fixed boundaries which define the bins. Certain conservative adjustments are made with this method: for example, if any bin contains fewer than five observations from the reference data set, then adjacent bins are pooled until this lower limit is achieved.

The forensic laboratories have sampled large numbers of DNA profiles from each major ethnic group to use as reference populations (Caucasians, African Americans, Hispanics, Asians, etc.; see Section 5.2). $\mathcal{LR}$ is calculated for most, or all, of these reference populations, and usually all of the calculations are presented to the jury.

## 2.4 Recommendations of the NRC Report

The NRC report (NRC, 1992) made many recommendations concerning DNA profiling, including one chapter on the statistical interpretation of the evidence. This section will explore some of the statistically relevant suggestions that have affected the calculation of $\mathcal{LR}$.

The NRC panel dismissed methods for calculating $\mathcal{LR}$ based on (5), espoused by Berry (1991) and others, which bypass the matching step (see Section 6) on the basis that they are too complicated. They also opposed the use of expert judgment in the declaration of a match, endorsing only objective methods.

The NRC panel proposed a novel method for calculating genotype "probabilities." Their suggestion, dubbed the *ceiling principle*, is to obtain samples of 100 unrelated individuals from 15–20 "relatively genetically homogeneous" populations, with examples being English, Germans, Russians, Navajos, Puerto Ricans and West Africans, among others. One would estimate allele probabilities from these populations for the VNTR loci commonly used for forensics. Then for any particular DNA profile, one would choose the maximum allele probability found among the study populations (the ceiling). In addition, they add the condition that no allele probability should be below 10% (the floor). Until these samples are obtained, they recommended an interim ceiling method.

The committee was motivated by the following considerations. Strictly speaking, the assumption of independence depends on the absence of population heterogeneity, an issue discussed further in Section 3. The committee correctly inferred that "in a population that contains groups each with different allele [probabilities], the presence of one allele in a person's genotype can alter the statistical expectation of the other alleles in the genotype." Although the committee notes that empirical studies have detected no violations of independence within or across loci, they chose "to provide a method for estimating population frequencies in a manner that would adequately account for [population heterogeneity]."

When considerable population heterogeneity exists, the choice of the reference population has a significant impact on the weight of the evidence. Therefore, "the committee recommends approaches of making *sound* estimates that are independent of the race or ethnic group of the subject." They claim "the ceiling principle eliminates the need for investigating the perpetrator population, because it yields an upper bound to the [probability] that would be obtained by that approach."

The NRC panel suggested a second calculation, which they dubbed the "one-on-$N$" rule. That is, if the appropriate reference population database is of size $N$, and if no one in the database matches the evidentiary sample, then $1/N$ should be presented as an upper bound on $\Pr(\mathcal{G}_e)$. This calculation would be presented to the jury along with or instead of the ceiling principle calculation. As noted by the panel, however, the "one-on-$N$" rule seems to be at odds with analyses of existing databases: "pairwise comparisons of all five-locus DNA profiles in the FBI database showed no exact matches;

the closest match was a single three-locus match among 7.6 million pairwise comparisons. Those studies are interpreted as indicating that multiplication of gene [probabilities] across loci does not lead to major inaccuracies in the calculation of genotype [probabilities]."

## 3. THE QUESTION OF INDEPENDENCE

Forensic scientists typically assume approximate independence of alleles both within and between loci. This assumption stirred most of the initial controversy. A population or ethnic group (e.g. Caucasians) composed of subpopulations (Irish, Italian, French, etc.) having different allele probability distributions is considered heterogeneous. The most likely violations of independence result from population heterogeneity (also known as *population substructure*). Population heterogeneity causes dependencies of alleles within and between loci (Li, 1969). Several authors have suggested that population heterogeneity could lead to a serious underestimate of the probability of two DNA profiles matching (Lander, 1989, 1991a, b; Cohen, 1990; Cohen, Lynch and Taylor, 1991). Other geneticists and statisticians have countered that, while the argument that heterogeneity causes dependence is theoretically correct, human populations rarely exhibit enough heterogeneity to have a substantial impact on forensic calculations (Devlin, Risch and Roeder, 1990; Chakraborty and Kidd, 1991; Chakraborty and Jin, 1992). Exceptions exist, they argue, but these exceptions occur in extremely isolated populations such as a few Amerindian tribes.

In Section 3.1, a probability model for population heterogeneity is developed. This model predicts an excess of homozygotes. In fact, the first paper claiming violations of independence based this claim on an apparent excess of homozygotes. In Section 3.2, this claim is shown to have been the result of a physical artifact (coalescence). In the remaining subsections, various tests of independence are described. Valid tests for independence find (approximately) the predicted number of rejections when a large number of tests have been performed. Tests for independence that do not model the physical artifacts (coalescence and correlated measurement error) have routinely yielded spurious rejections of the null hypothesis. Of course, as the sample size increases, even minute violations of independence will eventually lead to a rejection of the independence hypothesis. In Section 3.6, a Monte Carlo experiment illustrates the negligible effect on profile probabilities observed when the dependence between a pair of loci is ignored.

### 3.1 The Model

To understand the effect that allele distribution variation has on the calculation of genotype probabilities, consider the single-locus case. The model of population substructure assumes independent assortment of alleles within a subpopulation (random mating) and limited matings between subpopulations. This population substructure model is probabilistically equivalent to assuming that the vector of allele probabilities for a given subpopulation $G = (G(1), G(2), \ldots, G(m))$ possibly varies by subpopulation and that, conditional on $G$, an individual's pair of alleles is sampled independently. For example, in Figure 4 the estimated allele probability distributions at D4S139 and D10S28 are depicted for four subpopulations of Asians: Chinese, Koreans, Japanese and Vietnamese. Ignoring the nonnegligible sampling error in these allele distributions, the probability of observing a particular genotype from a Japanese person can be calculated by multiplying the Japanese allele probabilities

$$(6)\quad \Pr\left(\{a(i), a(j)\}\right) = \begin{cases} 2G_J(i)G_J(j), & \text{if } i \neq j, \\ G_J(i)^2, & \text{if } i = j. \end{cases}$$

Comparing the result of this calculation to a similar calculation for a Korean person yields essentially identical genotype probabilities for locus D4S139, but somewhat different results for D10S28. (Much of this difference could be due to sampling error; data obtained from Orange County Coroner's office.)

Suppose the perpetrator of a crime is known to be Asian, but his subpopulation is unknown. If an adequate reference database for each subpopulation is available, the genotype probability can be calculated as a weighted average over the various subpopulations:

$$(7)\quad \begin{aligned} &\Pr\left(\{a(i), a(j)\}\right) \\ &\quad = \begin{cases} 2\sum_k w(k)\, G_k(i)G_k(j), & \text{if } i \neq j, \\ \sum_k w(k)\, G_k(i)^2, & \text{if } i = j, \end{cases} \end{aligned}$$

where $w(k)$ is the relative frequency of the $k$th subpopulation in the population. This calculation yields the true probability of drawing genotype $\{a(i), a(j)\}$ from the Asian population provided H-W holds in the defined subpopulations. If no information is available concerning the allele probabilities in the subpopulations, the forensic scientist might make his calculation based on the mixed Asian reference population, assuming H-W holds (approximately) in the mixed population. If $\gamma(\cdot)$ is the marginal probability of observing $a(\cdot)$ in the population, then the probabil-

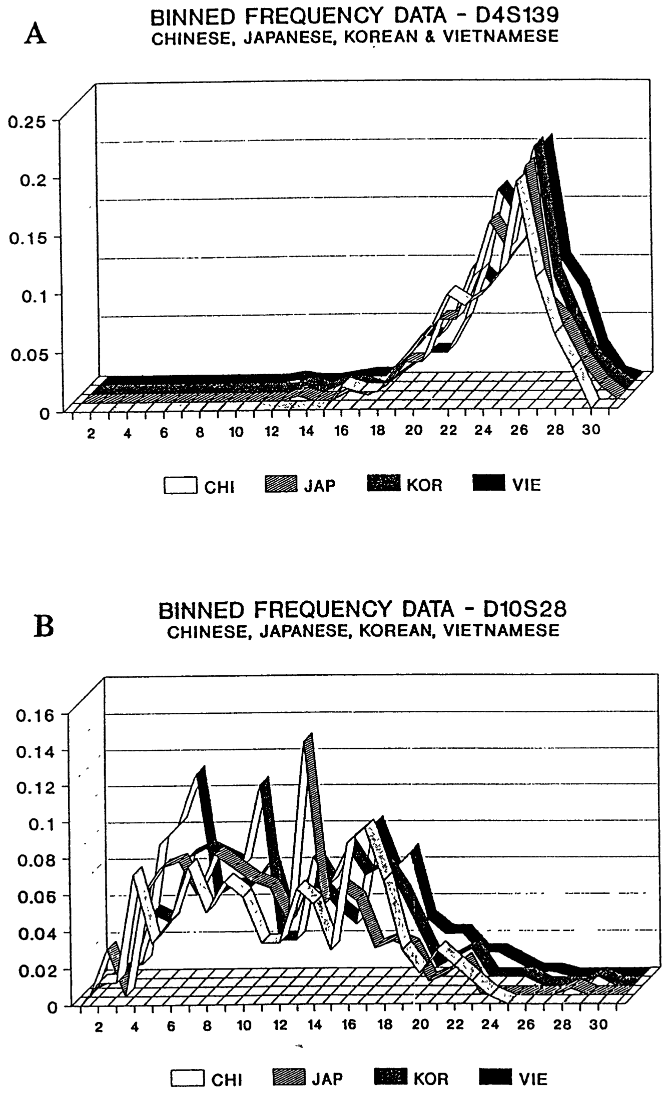


FIG. 4. *Fixed bin distribution (histogram) for two loci and four Asian subpopulations (used with permission from John Hartmann): the boundaries of the* 30 *bins (vertical axis) are determined by the FBI; these bins are not of equal length. Sample sizes (numbers of individuals) for Chinese, Japanese, Korean and Vietnamese are* 103, 125, 93 *and* 215 *for* D4S139 *and* 120, 137, 100 *and* 193 *for* D10S28. *The horizontal axis is the bin number; bins are not of equal length.*

ity of observing the genotype $\{a(i), a(j)\}$ is calculated using (1).

In general, assuming the population substructure model, the probability of observing the genotype $\{a(i), a(j)\}$ in a random draw from a population is correctly calculated as

$$\Pr\left(\{a(i), a(j)\}\right) = \begin{cases} 2\gamma(i)\gamma(j) + 2\operatorname{cov}[G(i), G(j)], & \text{if } i \neq j, \\ \gamma(i)^2 + \operatorname{var}[G(i)], & \text{if } i = j. \end{cases} \tag{8}$$

Assuming H-W in the entire population is equivalent to assuming there is no heterogeneity in the population: in other words, discarding the second term in both cases. Clearly this leads to an underestimate of the probability when $i = j$, and it will usually lead to an overestimate when $i \neq j$.

A genetic model, based on evolutionary theory, leads to a one-parameter model for the variances and covariances in (8),

$$\begin{aligned} E[G(i)] &= \gamma(i), \\ \operatorname{var}[G(i)] &= \theta_S\gamma(i)[1 - \gamma(i)], \\ \operatorname{cov}[G(i), G(j)] &= -\theta_S\gamma(i)\gamma(j) \end{aligned} \tag{9}$$

(see, e.g., Weir, 1990). In genetic parlance, $\theta_S$ is analogous to $F_{ST}$ of Wright (1951). A parametric model for $G$ with the same moments is the Dirichlet

$$\begin{aligned} G &= (G(1), G(2), \ldots, G(m)) \\ &\sim \text{Dirichlet}(\gamma(1), \gamma(2), \ldots, \gamma(m); \theta_S). \end{aligned}$$

## 3.2 An Excess of Homozygotes?

An obvious consequence of (8) is that population heterogeneity leads to an excess of homozygotes in the mixed population. This result is well known to geneticists. Indeed, the controversy over independence assumptions was ignited by a claim that an apparent excess of homozygotes indicated a "spectacular deviation from H-W" (Lander, 1989). Specifically, an excess of 9 and 13% homozygotes was claimed for D17S79 and D2S44, respectively. To obtain insight into the extreme heterogeneity implied by these figures, examine Figure 5. Consider a mixture of two populations: one population constructed by drawing pairs of alleles from the upper distribution and the other constructed by drawing pairs of alleles from the hanging distribution. A population composed of a 50-50 mixture of these two extremely disparate populations leads to only a 5% excess of homozygotes! In fact, no relevant populations have shown the amount of differentiation illustrated in Figure 5.

### 3.2.1 *A physical artifact masquerading as population substructure*

The data Lander based his claim upon were from Lifecodes Corporation. Devlin, Risch and Roeder (1990) reexamined these data, in large part because such a huge excess of homozygotes would be unusual for two reasons: (i) extreme heterogeneity would rarely lead to such an excess; and (ii) relevant human populations are not very heterogeneous (see Mourant, Kopec and Domainewska-Sobczak, 1976; Nei and Roychoudhury, 1982).

In the absence of measurement error, it would be simple to construct a test to determine if there are significantly more homozygotes in the population

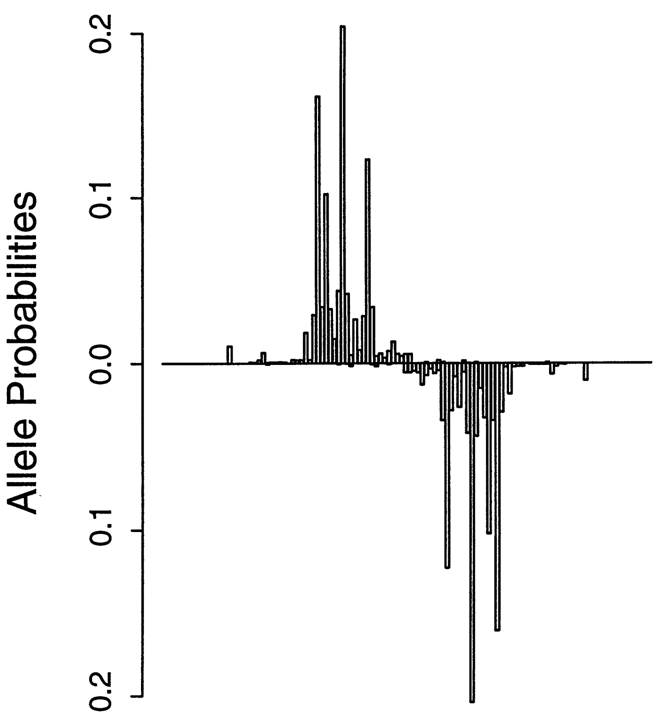


FIG. 5. *Two allele distributions used to create a mixed population: the distribution for* D17S79 *(upper distribution) and its mirror image (hanging distribution). The horizontal axis is the number of repeating units.*

than we would expect based on (8) and (9) (Levene, 1949). Let us consider a simple extension, which allows for continuous measurement error. Let

$$t_\varepsilon = \frac{O_\varepsilon - E_\varepsilon}{\sqrt{\operatorname{var}(O_\varepsilon - E_\varepsilon)}},$$

where

$$(10) \qquad O_\varepsilon = \frac{1}{r}\sum_{j=1}^{r} \mathbf{I}\{|z_{1j} - z_{2j}| < \varepsilon\}$$

is the observed number of *ε-homozygotes* (fragments which are separated by no more than ε). One can estimate the expected number of *ε-homozygotes* under H-W using a $U$-statistic

$$(11) \qquad E_\varepsilon = \frac{1}{2r-1}\sum_{i<j}^{2r} \mathbf{I}\{|z_i - z_j| < \varepsilon\},$$

where $\{z_i, i = 1, \dots, 2r\}$ is the collection of fragments from the population of interest, ignoring the pairings. If $t_\varepsilon$ is large relative to a standard normal, one can conclude there is an excess of homozygotes. For $\varepsilon = 0.0001$, $t_\varepsilon$ is highly significant for several loci and several populations. For example, $t_\varepsilon = 84.5$ for the Caucasians at D2S44; but if ε is increased, $t_\varepsilon$ decreases and quickly ceases to be significant (Devlin, Risch and Roeder, 1990). The most likely explanation for this phenomenon is coalescence (Section 2.2). With some effort, the coalescence function can be estimated from the data using natural generalizations of (10) and (11); a smooth logistic curve fits the data. For ε large enough that the probability of coalescence is less than 0.01, $t_\varepsilon$ is no longer significant for the data sets examined (Devlin, Risch and Roeder, 1990). Therefore the apparent excess of homozygotes can be explained by close heterozygotes masquerading as homozygotes, rather than a serious violation of model assumptions.

It is important to bear in mind that this composite test is not a complete test of H-W. Nevertheless, it has reasonable power to detect population substructure. To demonstrate this fact empirically, Devlin, Risch and Roeder (1991a) created mixed populations by randomly sampling from Lifecodes African American and Caucasian databases. For such mixtures of 1000 profiles from each ethnic group, they found nearly 100% power to detect mixture for at least one of the three loci (D17S79, D2S44 and D14S13). On the other hand, Devlin and Risch (1992b) found such mixtures induced only relatively small errors in genotype probability estimates. These experiments suggest that the amount of population heterogeneity within the forensic databases is substantially less than the amount of variability observed between Caucasians and African-Americans. One weakness of this test, however, is that it can be affected by correlated measurement error (Green and Lander 1991; Devlin, Risch and Roeder, 1991a) and so must be interpreted carefully.

### 3.2.2 *A Bayesian analysis of the data*

Slight violations of the independence assumptions are to be expected in any human population. Ideally we would like to obtain a posterior distribution of $\theta_S$ that would indicate the variability of $G$ over subpopulations. Based on the posterior distribution, adjustments to $\mathcal{LR}$ indicated by (8) could be made that rely on the covariance structure indicated in (9). Roeder et al. (1993) have obtained a result of this type using a hierarchical Bayes approach. Their model has five levels; each level is supported by theory and/or experimental data:

1. $(X_1, X_2)$ appear as $(X_1 + X_2)/2$ with probability determined by the coalescence function.
2. $(X_1, X_2)$, given $(A_1, A_2)$, are distributed as a bivariate normal with mean, variance and covariance determined by $(A_1, A_2)$.
3. $(A_1, A_2)$, given $G$, are independent samples from $G$.
4. $G$, given $(\gamma, \theta_S)$, is distributed as a Dirichlet $(\gamma; \theta_S)$.

5. A prior for $\gamma$ is obtained from experimental data, and a conservative prior for $\theta_S$, based on previous studies, is beta(1, 49).

In their analysis they obtain posterior distributions of $\theta_S$ with modes (and upper percentiles) near 0. For example, in D17S79, D2S44, D14S13 and D18S28 the posterior mode (95th percentile) for Caucasians is approximately 0.0015 (0.007). This indicates that there is almost no variability in the allele distributions of Caucasian subpopulations. It would be simple to incorporate an estimate of $\theta_S$ into corrections to the $\mathcal{LR}$ calculations. For example, assuming that the covariance structure in (9) is approximately correct and that $\theta_S$ is known, $\Pr(\{a(i), a(j)\})$ can obtained by substituting (9) into (8) to obtain

$$\Pr\left(\mathcal{G}_e, \mathcal{G}_s | H_1\right) = \begin{cases} \gamma(i)^2 + \theta_S \gamma(i)\left[1 - \gamma(i)\right], & \text{if } i = j, \\ 2\gamma(i)\gamma(j)(1 - \theta_S), & \text{if } i \neq j. \end{cases} \tag{12}$$

A full Bayesian correction involves integrating (12) over the posterior distribution of $\theta_s$. Corrections to (2) that allow for population heterogeneity are developed further in Section 4.

### 3.3 Classical Tests of Independence

Of course, mixture of subpopulations is not the only possible violation of the H-W assumption, and efforts have been made to check for other violations of the independence assumption. Perhaps the simplest test relevant to VNTR data is Fisher's intraclass correlation. This test is equivalent to testing for correlation in the set of pairs of fragments in the reference population $\{(z_{1j}, z_{2j}), (z_{2j}, z_{1j}), j = 1, \ldots, r^*\}$ and $\{(z_{1j}, z_{2j}), j = r^*, \ldots, r\}$, where the first $r^*$ measurements denote the heterozygotes. Such analyses reveal little, if any, relationship between paternal and maternal fragments (Budowle et al., 1991a; Berry, Evett and Pinchin, 1992; Weir, 1992a; Chakraborty, Srinivasan and de Andrade, 1993). An advantage of the intraclass correlation is its insensitivity to some of the electrophoretic phenomena discussed in Section 2.2, unlike other methods below.

Testing the assumption of H-W for a population is often straightforward for classical genetic marker data. These tests measure the difference between the observed and expected number of each distinct genotype in a sample from the population (Hernandez and Weir, 1989). Initial efforts to test H-W by applying standard $\chi^2$ methods to binned data were plagued by difficulties. At first, ad hoc techniques were applied in which the number of bins and the position of bins were selected to maximize or minimize the test statistic. Not surprisingly, contradictory results were obtained.

Presumably to circumvent these difficulties, Geisser and Johnson (1992) recently proposed a quantile $\chi^2$ test for H-W. The steps of the test are as follows: (i) Order the $2r$ fragments by size $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(2r)}$. (ii) Decide a priori how many quantiles will be used; call this number $q$. (iii) Choose bin boundaries $b_1, \ldots, b_{q-1}$ that divide the ordered variates into $q$ subsets, each with $f = 2r/q$ members. (iv) Since $z_{1i} \leq z_{2i}$, in two dimensions the $q$ boundaries divide the space into $q(q+1)/2$ bins. The number of fragments falling in each bin is called $O_{ij}$. Define $E_{ij} = r/q^2$ if $i = j$ and $E_{ij} = 2r/q^2$ if $i \neq j$. (v) Under reasonable choice of $q$, the Pearson's $\chi^2$ statistic obtained from this procedure is distributed $\chi^2$ with $q(q-1)$ degrees of freedom under the H-W hypothesis.

Geisser and Johnson (1992) claim that the only assumption of their test is bivariate exchangeability. This claim is challenged by Devlin and Risch (1993), who show via simulation that physical artifacts confound the test of H-W. Their arguments are based on the complications introduced by coalescence and correlated measurement error that are not allowed for in Geisser and Johnson's test. Coalesence causes an excess of mass in the bins on the diagonal and a dearth of mass in the adjacent off-diagonal bins. The effect of correlated errors is more complicated, although it is clear that the potential for confounding is present from (4). The goal is to assess violations of independence in $(A_1, A_2)$; however, the test is based on $(X_1, X_2)$ which are clearly correlated through $(\varepsilon_1, \varepsilon_2)$. The test is more sensitive to this effect when the allele distribution is very uneven, with a few very common alleles creating "ridges" in the two-dimensional surface (Figure 2).

Weir (1992a) also examines the question of independence using binned data. He uses the fixed bins set by the FBI, discarding the bins on the diagonal to remove the effect of coalescence. (Weir correctly points out that there is no need to be concerned about the H-W assumption for homozygotes, using the FBI's method, because H-W is not assumed for homozygotes — only one allele is used in the calculation.) He forms a test statistic using the likelihood ratio test. Because the data are sparse, he uses a bootstrap resampling procedure to determine the null distribution of the test statistic. Unfortunately this test is also sensitive to correlated errors and coalescence because of the dearth of mass in the off-diagonal cells. Notably, even with this disturbance in the data, Weir shows most loci and most databases conform to H-W expectations. In a similar analysis of the Lifecodes database, he obtains results generally supporting H-W and linkage equilibrium (Weir, 1992b).

### 3.4 Tests of Linkage Equilibrium

Mixture of subpopulations, as well as evolutionary forces such as selection, may induce associations between alleles at different loci. Weir (1992a, b), using tests similar to those described above, has thoroughly analyzed the VNTR databases for violation of linkage equilibrium. Again his analyses reveal no large deviations from independence.

### 3.5 Composite Tests of Independence between Loci

Population substructure causes an association among alleles in VNTR profiles. According to those critical of the forensic scientists' evaluation of the DNA evidence, this association can be substantial. If the association is strong, some VNTR profiles would occur substantially more often in the mixed database than would be predicted under independence. For discrete data, a classical test of independence for two loci requires the construction of a two-way table with rows consisting of possible genotypes for the first locus and columns consisting of possible genotypes for the second locus. Using a nonparametric approach that reduces this extremely large, sparse table to a $2 \times 2$ table, Risch and Devlin (1992a, b) examined both the FBI and the Lifecodes databases to determine if an excess of matching occurs in the databases.

To estimate match probabilities for individual loci, they made all $r(r-1)/2$ possible comparisons of the $r$ individuals in each database. A match was declared if the corresponding measurements differed by no more than four standard deviations. This is an estimate of the probability that a random, innocent suspect and an evidentiary sample would be declared a match by chance.

Under the independence assumption, the occurrence of genotypes at pairs of loci should be independent. Therefore, the probability that two individuals have matching genotypes at a pair of loci should be the product of the single-locus match probabilities. To test for violations of pairwise independence, they constructed $2 \times 2$ tables from the $r(r-1)/2$ comparisons, with match–no match at the first locus being the two rows and match–no match at the second locus being the columns. The expected values for the cells are obtained from the product of the single-locus match probabilities. Although the test statistic they used has the form of Pearson's $\chi^2$, it does not follow a $\chi^2$ distribution under the null hypothesis. Because the entries in the table are $U$-statistics, the test has greater variance than a classical test of independence. Using the method of bootstrapping to obtain the null distribution of the test statistic, they did not find evidence for significant violations of independence. See Herrin (1993) for similar results using different databases.

Under the population substructure hypothesis, more matches are expected than are predicted by independence, provided the different subpopulations actually have different allele probability distributions. The analyses suggest that, whatever disequilibrium exists among loci, it has little effect on the probability of two random individuals having matching genotypes. Risch and Devlin explain their results as indicating that subpopulations do not differ much in their allele distributions, at least for the ethnic groups they studied. In fact, they argue (see also Devlin, Risch and Roeder, 1993a, b) that the genotype probabilities differ more between ethnic groups than within ethnic groups—contrary to the arguments of Lewontin and Hartl (Lewontin and Hartl, 1991; Lewontin, 1993). Indeed other results support this supposition (see Section 5.1). For instance, in mixed databases formed by randomly selecting equal numbers of African American and Caucasian single-locus profiles (500 and 1000 each), a greater number of two- and three-locus matches are observed than are predicted by the independence model, usually significantly greater (Devlin, Risch and Roeder, 1994).

To evaluate the power of this test to detect population substructure, Devlin, Risch and Roeder (1994) again create artificial populations by mixing equal proportions of Lifecodes African American and Caucasian profiles (see Section 3.2). For each subpopulation, they randomly sampled and combined single-locus profiles, thereby ensuring equilibrium in each subpopulation. When each subpopulation consisted of 500 individuals, the power to detect mixture was 69, 61 and 58% for the locus pairs D2S44–D17S79, D2S44–D14S13 and D14S13–D17S79, and 33% for the three loci. For 1000 individuals in each subpopulation, power exceeded 95% for each pair of loci and was 81% for the triplet.

It is important to recall two facts regarding these results: (i) such mixture does not cause large (relative) errors in genotype probabilities (Devlin and Risch, 1992b); and (ii) Risch and Devlin (1992a) detected no violations of independence in the actual Lifecodes data for these loci, consisting of over 3000 Caucasians and 2000 African Americans. (Unfortunately this test, like the other composite test, can be affected by correlated measurement error.)

### 3.6 Sensitivity of $\mathcal{LR}$ to Assumptions

The implicit assumption required for multiplying genotype probabilities across loci is either homogeneous populations or random mating. Critics argue correctly that the populations of interest are neither homogeneous nor randomly mating, and then argue that it must be wrong to multiply probabilities. As Evett, Scranage and Pinchin (1993) point out, classical tests of significance do not address the issue of interest directly: tests may fail to reject the independence hypothesis because they lack power or, on the other hand, they may reject the independence hypothesis even though there are no practical consequences of the failure of the assumption.

Therefore, Evett, Scranage and Pinchin (1993) performed a Monte Carlo experiment to examine the practical significance of independence violations. The Metropolitan Police Forensic Science Laboratory database consists of a file of $r$ individuals, each with a three-locus profile (D1S7, D7S21, D12S11). For any two loci, they calculated $\mathcal{LR}$, under $H_0$, for all possible pairwise comparisons [i.e., $r(r-1)/2 \approx 1.2 \times 10^6$] for two settings: (i) the database itself (dependence assumed) and (ii) a database constructed by randomly reassigning the profiles of one locus (independence certain). The expectation is that the vast majority of the $\mathcal{LR}$'s will be essentially zero under either (i) or (ii), but $\mathcal{LR}$ will be stochastically larger for (i) under the population substructure model because, if there is an association between alleles in a subpopulation, a comparison between persons in the same subpopulation should lead to inflated values of $\mathcal{LR}$.

In the three between-locus comparisons, they found only one significant difference between distributions (D1S7, D7S21), but even for this pair of loci these distributions did not differ to a degree deemed practically significant. For example, the observed frequency of reporting an $\mathcal{LR}$ in excess of 1000 is about 3.7 cases per 100,000, compared with an expected rate of 2.7 cases per 100,000.

See also Brookfield (1992) for a discussion of the effect of population subdivision at a single locus.

## 4. REFERENCE POPULATIONS

There is little doubt that one of the thorniest issues in the DNA fingerprinting debate is that of selecting an appropriate reference population. A basic question is treated in Section 4.1: Should the reference population consist of individuals of the same ethnic group as the suspect? Assuming that an innocent suspect is unrelated to the donor of the evidentiary sample, the answer is no, his ethnicity is irrelevant unless his guilt is presupposed. In some cases it can be argued that this assumption should be weakened, allowing for various degrees of relatedness between the culprit and the suspect, from assuming that the culprit and suspect are members of the same subpopulation to assuming that the culprit and suspect are brothers. It is possible to adjust $\mathcal{LR}$ to allow for these weaker assumptions. For reasons of grammatical ease, I refer to the individual donating the evidentiary sample as the culprit, throughout this section, although this is legally incorrect. In Section 4.2 the suspect is assumed to be from the same subpopulation as the culprit, and in Section 4.3 the suspect is assumed to be a close relative of the culprit.

The conclusions of this section can be summarized as follows. The ethnicity of the suspect is not relevant, unless the suspect and culprit are assumed to be from the same subpopulation. Because usually little is known about the culprit, it is generally assumed that the suspect and culprit are possibly, but not necessarily, from the same subpopulation (unrelated). Some people argue that they should be assumed to be from the same subpopulation in all cases (Nichols and Balding, 1991). This distinction is not as important as one might initially expect. The correction for the suspect and culprit being from the same subpopulation has little effect on the weight of the evidence, unless the heterogeneity is considerably larger than that observed in most forensically important populations. When the suspect and culprit are close relatives, the correction may have a large impact. Brothers are considerably more likely to have matching DNA profiles than are unrelated individuals. Ignoring relatedness of first cousins also leads to an inflation of the weight of the evidence; however, it is not dramatic.

### 4.1 Whose Ethnicity Matters?

In Vermont, the defense succeeded in blocking the admission of DNA evidence by demonstrating that the suspect was of mixed Amerindian, Italian and French ancestry and then arguing that the FBI lacked an appropriate reference population for this case because it did not have a database of individuals of this ethnic mix (Weir and Evett, 1992). Was this ruling based on fallacious thinking?

In order to clarify the key issues, assume that the genotypes in (2) can be ascertained without error. A question arises: is $\Pr(\mathcal{G})$ to be calculated based on a reference population composed of individuals of the same ethnic mix as the suspect (suspect-based sampling) or individuals who could have committed the crime by virtue of having access to the crime scene, fitting eyewitness description and so on (culprit-based sampling)? The NRC report states that "Some legal commentators have pointed out that frequencies should be based on the population of possible perpetrators, rather than on the population to which a particular suspect belongs. Although this argument is formally correct, practicalities often preclude use of that approach." Regardless of any "practicalities," it is critical that the legal system give serious consideration to the appropriate reference population in any particular case. To clarify the issues, let us look at the formal logic put forth by Evett and Weir (1992), which supports culprit-based sampling. The argument is based on two key assumptions that have been the subject of debate.

Let $\mathcal{I}_e$ and $\mathcal{I}_s$ denote relevant information about the culprit and suspect, respectively, such as the ethnic background. The objective is to calculate the likelihood ratio

$$\mathcal{LR} = \frac{\Pr(\mathcal{G}_e, \mathcal{G}_s \mid H_1, \mathcal{I}_e, \mathcal{I}_s)}{\Pr(\mathcal{G}_e, \mathcal{G}_s \mid H_0, \mathcal{I}_e, \mathcal{I}_s)}. \tag{13}$$

Assumption 1. The relevant information about the culprit is consistent with the information about the suspect: $\mathcal{I}_e \subset \mathcal{I}_s$.

The consequence of this assumption is that the information must not be contradictory, such as the suspect is white and the culprit is known to be black, because then the probability of $H_1$ is 0. Taking the viewpoint of suspect-based sampling, write $\mathcal{LR}$ as

$$\frac{\Pr(\mathcal{G}_e \mid \mathcal{G}_s, H_1, \mathcal{I}_e, \mathcal{I}_s)}{\Pr(\mathcal{G}_e \mid \mathcal{G}_s, H_0, \mathcal{I}_e, \mathcal{I}_s)} \frac{\Pr(\mathcal{G}_s \mid H_1, \mathcal{I}_e, \mathcal{I}_s)}{\Pr(\mathcal{G}_s \mid H_0, \mathcal{I}_e, \mathcal{I}_s)}. \tag{14}$$

$\mathcal{LR}$ can be equivalently represented as

$$\frac{\Pr(\mathcal{G}_s \mid \mathcal{G}_e, H_1, \mathcal{I}_e, \mathcal{I}_s)}{\Pr(\mathcal{G}_s \mid \mathcal{G}_e, H_0, \mathcal{I}_e, \mathcal{I}_s)} \frac{\Pr(\mathcal{G}_e \mid H_1, \mathcal{I}_e, \mathcal{I}_s)}{\Pr(\mathcal{G}_e \mid H_0, \mathcal{I}_e, \mathcal{I}_s)}, \tag{15}$$

which may be a more convenient representation from the viewpoint of culprit-based sampling.

Assumption 2. *Independence model.* If $H_0$ holds, then $\mathcal{G}_e$ is independent of $\mathcal{G}_s$ and $\mathcal{I}_s$; furthermore, $\mathcal{G}_s$ is independent of $\mathcal{G}_e$ and $\mathcal{I}_e$.

By Assumption 2, if $\mathcal{G}_e = \mathcal{G}_s$, then (14) and (15) simplify to

$$\frac{\Pr(\mathcal{G}_s \mid H_1, \mathcal{I}_e, \mathcal{I}_s)}{\Pr(\mathcal{G}_s \mid \mathcal{I}_s) \Pr(\mathcal{G}_e \mid \mathcal{I}_e)}, \tag{16}$$

and

$$\frac{\Pr(\mathcal{G}_e \mid H_1, \mathcal{I}_e, \mathcal{I}_s)}{\Pr(\mathcal{G}_e \mid \mathcal{I}_e)) \Pr(\mathcal{G}_s \mid \mathcal{I}_s)}, \tag{17}$$

respectively. Because (14) and (15) are identical, $\Pr(\mathcal{G}_s \mid H_1, \mathcal{I}_e, \mathcal{I}_s) = \Pr(\mathcal{G}_e \mid H_1, \mathcal{I}_e, \mathcal{I}_s)$. From Assumption 1 the lack of symmetry in the problem comes into play—the numerators of both equations simplify to $\Pr(\mathcal{G}_s \mid \mathcal{I}_s)$. We conclude that $\mathcal{LR} = 1/\Pr(\mathcal{G}_e \mid \mathcal{I}_e)$: the information about the suspect's ethnicity cancels out of the equation. The court's reasoning was fallacious, under these assumptions.

### 4.2 Subpopulations

By Assumption 2, the suspect and culprit are assumed to be unrelated, that is, they are not from the same family nor necessarily from the same subpopulation. Occasionally a good case can be made for some sort of relatedness between the suspect and the culprit under $H_0$. For example, suppose a crime occurred in Seneca, Kansas, where most residents are of German descent, and both the victim and the suspect are Senecans. If the circumstances of the crime indicate that the culprit is a local, chances are fairly high that the culprit is of German descent.

The statistical properties of population substructure were first described by Wright (1951), and further elaborated by Cockerham (1969, 1972); these theoretical developments have been recently used to extend the calculation of $\mathcal{LR}$ to allow for some relatedness (Nichols and Balding 1991; Balding and Nichols, 1994; Morton, 1992; Weir, 1994).

Assumption 3. *Affinal model.* The culprit and suspect derive from the same subpopulation.

Under the affinal model, the appropriate reference population is the subpopulation of the suspect; however, typically, the crime laboratory possesses an insufficient amount of data for this approach. Nevertheless, calculations can be based on the larger reference population. We illustrate these ideas using a one locus marker. Assume $\mathcal{G}_e = \mathcal{G}_s = \{a(i), a(j)\}$ and calculate $1/\mathcal{LR} = \Pr(\mathcal{G}_e \mid \mathcal{G}_s, H_0)$ by extending the reasoning used to obtain (9),

$$\Pr(\mathcal{G}_e \mid \mathcal{G}_s, H_0) = \begin{cases} 2\dfrac{[\theta_s + (1-\theta_s)\gamma(i)]\,[\theta_s + (1-\theta_s)\gamma(j)]}{(1+\theta_s)(1+2\theta_s)}, & \text{if } i \neq j, \\[2ex] \dfrac{[2\theta_s + (1-\theta_s)\gamma(i)][3\theta_s + (1-\theta_s)\gamma(i)]}{(1+\theta_s)(1+2\theta_s)}, & \text{if } i = j. \end{cases} \tag{18}$$

These formulas can be derived based on the moments of the Dirichlet, which happen to agree with evolutionary theory (Balding and Nichols, 1994); similar formulas have been derived directly from evolutionary theory by Weir (1994).

If $\theta_S$ is bigger than 0, then this calculation yields less evidence for $H_1$ than the independence model. However, as noted in Section 3, for Caucasians $\theta_S \approx 0.0015$. This small estimate for $\theta_S$ is typical; see Morton (1992) for a table of values of $\theta_S$ obtained from demographic information concerning human populations that do not include U.S. populations. U.S. ethnic groups, being a melting pot of subpopulations, are expected to have smaller values of $\theta_S$. See also Chakraborty and Jin (1992). The results of these studies, and others, can be summarized as follows: some major ethnic groups in the United States ex-

TABLE 1

*Allele distribution for Poles and Italians, obtained from Lewontin and Hartl* (1991; *LH*) *and Chakraborty and Kidd* (1991; *CK*) *MGP =* Pr($\mathcal{G}$), *assuming independence* (*H-W*) *and population substructure* (*True*)

| | | LH | | CK | |
|---|---|---|---|---|---|
| Locus | Allele | Poles | Italians | Poles | Italians |
| Rh | cDe | 0.047 | 0.006 | 0.042 | 0.033 |
| | Cde | 0.044 | 0.015 | 0.030 | 0.011 |
| Kell | K | 0.058 | 0.015 | 0.043 | 0.049 |
| | k | 0.942 | 0.985 | 0.957 | 0.951 |
| ABO | A | 0.370 | 0.370 | 0.259 | 0.239 |
| | B | 0.220 | 0.070 | 0.145 | 0.142 |
| MGP | H-W | $3.69 \times 10^{-6}$ | | $5.24 \times 10^{-6}$ | |
| MGP | True | $1.19 \times 10^{-6}$ | | $5.69 \times 10^{-6}$ | |

hibit almost no heterogeneity (African Americans, Caucasians), some exhibit a minor amount of heterogeneity (Hispanics), while some exhibit enough heterogeneity to have an impact on the calculation of $\mathcal{LR}$ (Amerindians). For the latter, separate databases are generally kept for the tribes that are largely isolated, such as the Navajos.

The calculations in this section highlight a logical flaw commonly made by critics of the DNA methodology (e.g., Lewontin and Hartl, 1991; Hartl and Lewontin, 1993). Arguing from the position that a subpopulation is the "appropriate" reference population, they take the ratio of the allele probabilities in the appropriate reference population to those *in another subpopulation* as a measure of the error induced by typical forensic calculations. This ratio is not relevant because it is a general database, composed of many subpopulations, that is used by forensic scientists. Therefore the error in calculation is the ratio of a calculation like (6) to (1), not the ratio of particular allele probabilities of subpopulations. Take the example presented by Lewontin and Hartl (1991) and rebutted by Chakraborty and Kidd (1991). Allele probabilities for Poles and Italians are presented for several blood group loci (see Table 1). Although Lewontin and Hartl express concern that the ratio of the Polish genotype probability to the Italian genotype probability is 247, the appropriate comparison is the genotype probability of Poles ($7.4 \times 10^{-5}$) to the genotype probability obtained assuming H-W in the mixed Polish-Italian population ($3.7 \times 10^{-5}$), only a 2.0-fold difference. In a number of this magnitude, this is hardly an error of practical importance. In fact the data presented in Lewontin and Hartl are flawed. [These data are taken from an outdated reference and include typographical errors (Chakraborty and Kidd, 1991; Morton, Collins and Balazs, 1993). Repeating the experiment with the full set of published data, the results change dramatically. The multilocus genotype probabilities are $5.74 \times 10^{-6}$, $4.73 \times 10^{-6}$, $5.24 \times 10^{-6}$ and $5.69 \times 10^{-6}$ for the Polish, Italian, H-W and true mixed probabilities, respectively (Chakraborty and Kidd, 1991). See Morton, Collins and Balazs (1993) for further discussion. Apparently there is little harm in assuming H-W and/or using a general mixed population rather than a subpopulation in this instance.]

In general, if the appropriate reference population is a particular subpopulation and only an estimate of $\theta_S$ is available, then the magnitude of the error can be obtained from a comparison of $\Pr(\mathcal{G}_e \mid \mathcal{G}_s, H_0)$ calculated using (1) versus (18).

## 4.3 Relatives

A more serious concern is based on the defense "the culprit was my brother" (Evett, 1992b). Using standard genetic principles, $\mathcal{LR}$ can be calculated when the culprit is assumed to be a relative of the suspect. For example, at a particular locus, identical twins share both alleles *identical by descent* (inherited from the same parent); full sibs (regular brothers) have a 25, 50 and 25% chance of sharing both alleles, one allele and no alleles identical by descent, respectively. Of course they can also share alleles by chance, which is called *identical by state*. From this we can infer that no matter how polymorphic the genetic markers, there is at least a $(0.25)^L$ chance that an individual matches his brother at $L$ loci [the exact probability for full sibs is $[1+2\Sigma_k\gamma(k)^2+2(\Sigma_k\gamma(k)^2)^2-\Sigma_k\gamma(k)^4]/4$ (Weir, 1993)].

ASSUMPTION 4. *Cognate model.* The culprit is a relative of the suspect with probability $c_p$ of having $p$ alleles identical by descent (Cotterman, 1940) and genotypes of the parents of the suspect are unknown.

If $\mathcal{G}_e = \mathcal{G}_s = \{a(i), a(j)\}$, then

$$\Pr(\mathcal{G}_e \mid \mathcal{G}_s, H_0) \tag{19}$$
$$= \begin{cases} c_2 + c_1\gamma(i) + c_0\gamma(i)^2, & \text{if } i = j, \\ c_2 + c_1\left[\gamma(i) + \gamma(j)\right]/2 + 2c_0\gamma(i)\gamma(j), & \text{if } i \neq j. \end{cases}$$

Clearly $\mathcal{LR}$ obtained from this calculation is considerably larger than that obtained in (2) when the culprit is allowed to be as closely related as a brother (for full sibs $c_2 = 1/4$, $c_1 = 1/2$, $c_0 = 1/4$). For individuals as closely related as first cousins, the effect is already relatively unimportant as $c_2 = 0$ and $c_1 = 1/4$.

Suppose the courts want to entertain two null hypotheses:

$H_{00}$: the samples were obtained from unrelated individuals;

$H_{01}$: the samples were obtained from brothers.

In this situation the posterior odds of $H_1$ no longer factor as in (3), hence the probability of the competing hypotheses enter into the calculation in a meaningful way,

$$\text{odds}(H_1) = \frac{\Pr(H_1)\Pr(\mathcal{G}_s, \mathcal{G}_e|H_1)}{\Pr(H_{00})\Pr(\mathcal{G}_s, \mathcal{G}_e|H_{00}) + \Pr(H_{01})\Pr(\mathcal{G}_s, \mathcal{G}_e|H_{01})}.$$

Two legal avenues have been used by the prosecution in situations where relatives of the suspect are also under suspicion. The first is to verify that all near relatives either have a solid alibi and/or their DNA does not match. If this avenue is not viable, then the jury is presented with $\mathcal{LR}$'s calculated under the two competing hypotheses separately.

## 5. GENETIC DIVERSITY AND SAMPLING ISSUES

For a trial involving matching DNA profiles, the forensic scientist, the prosecuting and defense attorneys and the judge are charged with conveying the significance of the match to the jurors. The significance depends on the reference population(s) that are appropriate to the case. The initial decision on the proper reference population is made by the forensic scientist and the prosecuting attorney; others can be advanced by the defense or, less frequently, by the judge. For the majority of cases, forensic scientists believe the general population is the appropriate reference population, evaluating the profile probability in each of several ethnic databases gathered for this purpose. Frequently, even if the crime is committed in a region where the vast majority of the people belong to a subpopulation (e.g., German ancestry) of an ethnic group (e.g., Caucasian), the forensic scientist evaluates the evidence based on the ethnic group, not the subpopulation. To justify this approach two arguments are made: (i) individuals of other subpopulations had access to the crime scene; (ii) there is little difference between the subgroup and ethnic group profile probabilities, and this difference is minor relative to their conservative evaluation of the evidence.

It is true that binning methods generally yield smaller $\mathcal{LR}$'s than the statistical methods (see Section 6). Nevertheless, from the statistical perspective, the forensic scientist's argument invites two questions: Are the differences among subgroups small relative to the variation among individuals within a subgroups (known to be vast) and the variation among groups (known to be relatively small)? Are the reference databases a reasonably accurate reflection of the ethnic group? These questions will be addressed in Sections 5.1 and 5.2, respectively.

One basic conclusion can be drawn from this section. Despite concern about heterogeneity among subpopulations, the data show DNA profile probabilities are quite similar across subpopulations: the dominating source of genetic diversity can be attributed to differences among individuals within a subpopulation. The relevant literature demonstrates diversity among subpopulations is less than diversity among ethnic groups. Because of the tremendous amount of genetic variability among individuals, variability among ethnic groups has little practical impact on the calculation of DNA profile probabilities for most cases. Moreover, this variability can be observed and reported. Variability among subpopulations has even less effect than variability among ethnic groups. Results of a study that purports to have shown otherwise are explained by a bias in the statistical methodology; the study did not account for sampling error.

Concern about the quality of reference populations obtained by convenience sampling can be eased by the same reasoning. Obviously, a reasonable effort must be taken to obtain a representative sample of unrelated individuals; however, it appears that a well-designed experiment may not yield results that differ measurably from a reasonable convenience sample. This point is illustrated by examining a historical database; allele probabilities obtained from a stratified random sample are remarkably similar to results obtained from a convenience sample.

### 5.1 Partitioning Diversity

From the existing reference populations that have been classified by major ethnic group (e.g., Caucasian, Hispanic and African American), it is apparent that there is tremendous variability between individuals within an ethnic group and limited variation between ethnic groups. In fact the variability within an ethnic group is so great that the probability of two unrelated individuals matching is vanishingly small (Risch and Devlin, 1992a). One would like to conclude from this that the choice of reference population is of little importance. The relevance of this result is questionable, however, if there is substantially greater variability between subpopulations within an ethnic group (*subpopulation diversity*) than the variability between ethnic groups (*ethnic diversity*). Under this scenario there exist clusters of individuals with similar DNA profiles. Hence, although two individuals picked at random from the mixed population may be unlikely to match, two people picked from the cluster may be considerably more likely to match.

The authors of the NRC report base their recommendations on this conjecture, espoused by Lewontin and Hartl (1991) and based on research by Lewontin (1972). Much attention has focused on the historical records of markers such as blood antigens which have been classified by both ethnic group and subpopulation. These results are relevant to the debate because the results reflect the mating and demographic characteristics of human populations. The analyses of these historical databases, as well as some modern VNTR databases, are discussed below.

Let us start by establishing some notation: let $j$ index subpopulations within ethnic groups, $j = 1, \dots, J$; let $i$ index ethnic groups, $i = 1, \dots, I$; and let $c$ index sampled alleles within subpopulation $j$ and ethnic group $i$, $c = 1, \dots, 2C$. For simplicity of notation, assume an equal number of individuals $C$ is randomly sampled per subpopulation, and likewise an equal number of subpopulations $J$ is randomly sampled per ethnic group.

For the $c$th allele, let $W_{ijc}(k)$ be 1 if the allele is $a(k)$ and 0 otherwise. Random mating is assumed within a subpopulation so there is no need to distinguish between alleles within or between individuals.

Within the $(i,j)$th level, the probability $W_{ijc}(k) = 1$ is $\gamma_{ij}(k)$. As in Section 3.1, I assume $\gamma_{ij}(k)$ varies over the subpopulations and has conditional expectation $\gamma_{i\cdot}(k)$ and variance $\gamma_{i\cdot}(k)(1 - \gamma_{i\cdot k}(k))\theta_S$. Such variability induces correlation in the genes of individuals of a common group:

$$\begin{aligned} &E\left[W_{ijc}(k) \mid \gamma_{i\cdot}(k)\right] = \gamma_{i\cdot}(k), \\ &E\left[W_{ijc}(k)W_{ijc'}(k) \mid \gamma_{i\cdot}(k)\right] \\ &\quad = \gamma_{i\cdot}(k)^2 + \gamma_{i\cdot}(k)\left(1 - \gamma_{i\cdot}(k)\right)\theta_S. \end{aligned}$$

Following the same pattern we allow $\gamma_{i\cdot}(k)$ to vary over ethnic groups with expectation $\gamma_{\cdot\cdot}(k)$. This induces correlation between alleles sampled from the same ethnic group with parameter $\theta_E$ denoting coancestry of alleles in common ethnic groups, but different subpopulations,

$$\begin{aligned} &E\left[W_{ijc}(k) \mid \gamma_{\cdot\cdot}(k)\right] = \gamma_{\cdot\cdot}(k), \\ &E\left[W_{ijc}(k)W_{ij'c'}(k) \mid \gamma_{\cdot\cdot}(k)\right] \\ &\quad = \gamma_{\cdot\cdot}(k)^2 + \gamma_{\cdot\cdot}(k)\left(1 - \gamma_{\cdot\cdot}(k)\right)\theta_E. \end{aligned}$$

Lewontin (1972) proposed a partition of diversity based on a hierarchical application of the Shannon Information criterion. To measure individual diversity, Lewontin suggested

$$\phi_1 = -\frac{1}{IJ}\sum_i\sum_j\sum_k \gamma_{ij}(k)\log\gamma_{ij}(k).$$

To measure subpopulation diversity, he suggested

$$-\frac{1}{I}\sum_i\sum_k \gamma_{i\cdot}(k)\log\gamma_{i\cdot}(k) - \phi_1 = \phi_2 - \phi_1.$$

Finally, to measure population diversity, he suggested

$$-\sum_k \gamma_{\cdot\cdot}(k)\log\gamma_{\cdot\cdot}(k) - \phi_2 = \phi_3 - \phi_2.$$

Because each level of this partition is a convex combination of the level before, the estimated diversities must be nonnegative.

Of course, to obtain estimates of the components of diversity, $\widehat{\phi}_1, \widehat{\phi}_2$ and $\widehat{\phi}_3$ are obtained by replacing the unknown allele probabilities by their corresponding estimates. To study the effect of using estimates of allele frequencies rather than the true values, we use a Taylor series expansion of $\widehat{\gamma}(k)$ about $\gamma(k)$, to obtain the expected value of $\widehat{\gamma}(k)$:

$$\begin{aligned} &E\left[\widehat{\gamma}(k)\log\widehat{\gamma}(k)\right] \\ &\quad \approx \gamma(k)\log\gamma(k) + \frac{1}{2}\frac{1}{\gamma(k)}\operatorname{var}\left[\widehat{\gamma}(k)\right]. \end{aligned} \tag{20}$$

It follows from the approximation (20) that $\widehat{\phi}_1, \widehat{\phi}_2$ and $\widehat{\phi}_3$ are biased estimates. Because the number of alleles sampled to obtain the estimates $\widehat{\gamma}_{ij}(k)$, $\widehat{\gamma}_{i\cdot}(k)$ and $\widehat{\gamma}_{\cdot\cdot}(k)$ increases rapidly, the variances decrease rapidly. The ultimate effect is that the subpopulation diversity estimate tends to possess a strong positive bias (Devlin, Risch and Roeder, 1993b, 1994). This bias depends on $\theta_S$ and $\theta_E$ and is reduced for large $C$.

Applying this method to small samples of subpopulations from around the world, including groups as remote as Pygmies, Lewontin obtained the following breakdown: ethnic diversity and subpopulation diversity accounted for 6.3 and 8.3%, respectively, of the total diversity, while the remaining 85.4% was attributable to variation among individuals of the same subpopulations.

In the 21 years since this surprising result was published, standard methods for partitioning variance have been developed (Cockerham, 1969, 1972) and a large body of consistent results exists (Smouse, Spielman and Park, 1982; Nei and Roychoudhury, 1982; Chakraborty and Jin, 1992; Morton, Collins and Balazs, 1993). These results indicate that subpopulation diversity is dominated by ethnic diversity. Even a study (Latter, 1980) using Lewontin's methods, as well as similar populations and loci, but with substantially larger samples found that 5.6% of the diversity was attributable to differences between subpopulations and 10.4% to differences among ethnic groups.

Neither Lewontin's nor Latter's results are completely applicable to industrialized societies like the United States because they give as much weight to small isolated populations (e.g., tribes of Pygmies) as to the large open subpopulations that have populated the United States. In this regard, Nei and Roychoudhury's results are more applicable. They also found that ethnic groups accounted for about 10% of the diversity, but only 0.5% or less was attributed to differences among English, Germans and Italians. In industrialized societies like the United States, the estimate of diversity based on variance of VNTR allele frequencies among subpopulations is usually quite small—approximately 0.1% (Morton, Collins and Balazs, 1993; see also Budowle et al., 1994).

The calculations and empirical results presented in this section suggest that Lewontin's contradictory results can be explained, in part, by sampling error. His method for partitioning diversity fails to account for sampling error and hence leads to biased results. Competing methods (e.g., Cockerham, 1969, 1972; Weir and Cockerham, 1984) are based on standard partitioning of variance and do account for sampling error.

### 5.2 Construction of Reference Databases

From the statistical perspective, the construction of a reference database (say, U.S. Caucasians) is a straightforward if onerous proposition: simply collect a stratified random sample from U.S. Caucasians and obtain the genotypes of that sample. Human population geneticists and forensic scientists are less fastidious. To construct a Caucasian database, they generally collect and type samples that are convenient; for example, from blood banks and law enforcement officers for the FBI database, and mothers and putative fathers from paternity cases for the Lifecodes database. The forensic scientists (and human geneticists in general) argue, based in part on results presented in the previous section, that there is little difference between a stratified random sample and a convenience sample for their purposes. Statisticians without knowledge of human genetics are naturally skeptical of this claim.

There are reasons, however, to believe this claim is correct; for instance, consider the implications of the observations about genetic diversity in the previous subsection. There are also data that bear directly on this issue, in particular, the results of a stratified random sample of traditional genetic markers (DHEW 1980), which are described below.

A study that obtained a stratified random sample of traditional genetic markers (ABO; RH-C, D, E; Secretor status; haptoglobin; and transferrin) was sponsored by the Department of Health, Education and Welfare during the years 1967–1970. The study stratified according to ethnic group (African American, Caucasian), sex, geographic region (Northeast, Midwest, South and West), income and education. Justification for the study was that, while there existed numerous studies of these markers from U.S. samples (compiled in Mourant, Kopec and Domainewska-Sobczak, 1976), "none had been random or systematic, and none had been systematic representative samples of the U.S. population." Indeed, the statement is true, but the authors perhaps overstate the need for their study (especially given their results) when they argue that published allele frequencies could not "be considered a representative sample of the U.S. or any U.S. geographic region."

What is striking about the results of this study is how little variability there is among geographic regions. See Table 2 for the regional data for Caucasians. None of the regional differences in allele frequencies have a probability of occurrence smaller than 0.05 for either ethnic group. There are also striking similarities between the results of the stratified random sample and the studies published in Mourant, Kopec and Domainewska-Sobczak (1976). As an example, one such study (Niederman, Gilbert and Spiro, 1962) examines two markers from a sample of 1000 Yale students, all Caucasian males. The results of this study are given in Table 3. Apparently one must conclude that there is not much genetic heterogeneity in U.S. Caucasian populations or that Yale was doing an excellent job of obtaining a representative sample of U.S. Caucasians during the years 1958–1959, or both.

Regarding the VNTR databases themselves, there are again striking similarities between allele distributions from different regions of the United States for all major ethnic groups (Budowle et al., 1994), despite the fact that sampling error must be large for these relatively small databases. Chakraborty

TABLE 2
*Genotype distributions for Caucasian population, by locality [NE=Northeast (n = 1428), MW=Midwest (n = 1582), S=South (n = 1206), W=West (n = 1519); U/nd=Unsatisfactory/not done]*

| Locus | Genotype | NE | MW | S | W |
|---|---|---|---|---|---|
| ABO | $O$ | 0.430 | 0.410 | 0.463 | 0.474 |
| | $A_1$ | 0.356 | 0.358 | 0.334 | 0.353 |
| | $A_2$ | 0.060 | 0.071 | 0.081 | 0.062 |
| | $B$ | 0.114 | 0.115 | 0.093 | 0.083 |
| | $A_1B$ | 0.029 | 0.032 | 0.016 | 0.019 |
| | $A_2B$ | 0.011 | 0.012 | 0.012 | 0.008 |
| Rh-D | $D^+$ | 0.850 | 0.859 | 0.829 | 0.846 |
| | $D^-$ | 0.147 | 0.141 | 0.167 | 0.152 |
| | $D^u$ | 0.003 | 0.000 | 0.004 | 0.002 |
| Rh-C | $CC$ | 0.180 | 0.175 | 0.165 | 0.188 |
| | $Cc$ | 0.495 | 0.510 | 0.484 | 0.512 |
| | $cc$ | 0.325 | 0.315 | 0.350 | 0.300 |
| Rh-E | $EE$ | 0.031 | 0.030 | 0.037 | 0.027 |
| | $Ee$ | 0.254 | 0.259 | 0.266 | 0.294 |
| | $ee$ | 0.714 | 0.711 | 0.696 | 0.679 |
| Secretor status | $Se^+$ | 0.762 | 0.780 | 0.719 | 0.754 |
| | $Se^-$ | 0.202 | 0.203 | 0.249 | 0.225 |
| | U/nd | 0.037 | 0.017 | 0.031 | 0.021 |
| Haptoglobin | 1-1 | 0.150 | 0.179 | 0.152 | 0.171 |
| | 2-1 | 0.460 | 0.470 | 0.464 | 0.475 |
| | 2-2 | 0.350 | 0.325 | 0.321 | 0.320 |
| | U/nd | 0.024 | 0.010 | 0.035 | 0.018 |
| Transferrin | $CC$ | 0.959 | 0.967 | 0.947 | 0.966 |
| | $BC$ | 0.007 | 0.020 | 0.011 | 0.011 |
| | $DC$ | 0.005 | 0.003 | 0.007 | 0.005 |
| | Other | 0.002 | 0.000 | 0.001 | 0.000 |
| | U/nd | 0.028 | 0.010 | 0.034 | 0.018 |

TABLE 3
*Genotype distributions for Yale Caucasian male students (n = 1000) versus the stratified random sample of Caucasian population*

| Locus | Genotype | Yale | NE | MW | S | W |
|---|---|---|---|---|---|---|
| ABO | $O$ | 0.431 | 0.430 | 0.410 | 0.463 | 0.474 |
| | $A$ | 0.422 | 0.416 | 0.431 | 0.415 | 0.415 |
| | $B$ | 0.110 | 0.114 | 0.115 | 0.093 | 0.083 |
| | $AB$ | 0.037 | 0.040 | 0.044 | 0.038 | 0.027 |
| Secretor status | $Se^+$ | 0.773 | 0.791 | 0.793 | 0.743 | 0.770 |
| | $Se^-$ | 0.227 | 0.209 | 0.207 | 0.257 | 0.230 |

(1993), who analyzed these regional data using a method to partition diversity that does not account for sampling error, found that only 0.4% of the diversity occurred among localities within an ethnic group. Contrast this with Lewontin and Hartl's (1991) characterization of the similarity of the regional data: "vague claims of the similarity in the 'shape' of the VNTR distribution ... are irrelevant."

The FBI's database has been subject to the most criticism, in large part because it is reputed to be composed predominantly of FBI agents. (In fact, about 25% of the Caucasian samples are FBI agents.) It has been argued that analyses demonstrating the rarity of DNA profiles matches (Risch and Devlin, 1992a) were meaningless because they involve the FBI database. These criticisms were based on the belief that there are large (genetic) differences between Caucasian FBI agents and the average Caucasian American (Geisser, 1992). An analysis of the VNTR data suggests otherwise. Comparing the FBI agents to the remainder of the Caucasian database (blood bank samples), no locus demonstrates a significant difference between the fixed-bin allele distributions. Figure 6 plots the allele probabilities for FBI agents against the remaining sample for two loci, D2S44 and D4S139.

It appears that human geneticists are correct: the databases constructed by convenience sampling are reasonable proxies for stratified random samples. Does this mean that the evaluation of DNA profiles could not benefit from statistical considerations? Of course not. The regional databases are small and therefore subject to substantial effects of sampling error, as well as minor inherent regional differences. It would seem that this is an ideal setting for application of ideas from empirical Bayes methods. A forensic scientist in Minnesota may wish to use her Minnesota data, yet draw strength from the numerous other databases from around the country. With an estimate of $\theta_S$ and an estimate of the the sampling error, empircal Bayes estimates of allele frequencies are directly available.

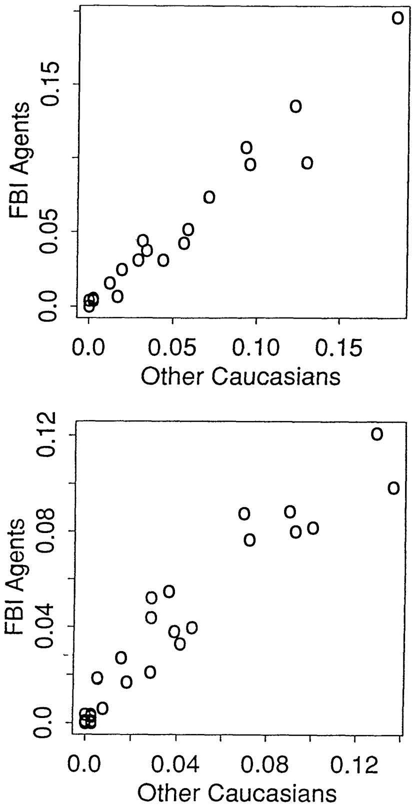


FIG. 6. *A plot of the fixed-bin allele probabilities of* D2S44 (*top*) *and* D4S139 (*bottom*) *for a partition of the FBI's Caucasian database: this partition is all FBI agents in one group, and the remaining Caucasians in the other group.*

## 6. CALCULATING $\mathcal{LR}$ WITHOUT DECLARING A MATCH

Berry (1991) introduced a method for obtaining $\mathcal{LR}$ that obviates the need to declare a match. In Section 6.1, I review the current literature on this topic, motivating it from a perspective somewhat different from Berry's. Methods of this type bypass the matching step of the match/binning method, thereby potentially avoiding a great deal of argument in the courts. Although, in theory, $\mathcal{LR} \neq 0$ for any profile, in practice, when there are large differences between comparable fragments, $\mathcal{LR} \approx 0$. Methods based on continuous versions of the likelihood ratio enjoy some use in the United Kingdom. Unfortunately, in the United States, their use is limited to theoretical calculations and comparisons.

### 6.1 The Likelihood Ratio

To derive $\mathcal{LR}$ for VNTR data, first consider the single locus case. For simplicity of exposition, ignore the complications introduced by coalescence. From the physical properties of the data, it is known that if $A_1 \leq A_2$, then $X_1 \leq X_2$. This holds (approximately) in the bivariate normal model, too, because the correlation is large for alleles of similar size. When $(X_1,X_2)$ are measurements of $a(i) < a(j)$, then with high probability $X_1 \leq X_2$ (as in Figure 2). Henceforth, by convention, take $X_1$ to be the smaller of the two measurements.

Let $(x_1,x_2)$ and $(y_1,y_2)$ denote the suspect and the evidentiary samples, respectively. Under $H_0$, the samples are obtained as a random draw from the population and hence the likelihood is

$$\text{lik}(x_1,x_2,y_1,y_2 \mid H_0) = f(x_1,x_2)\, f(y_1,y_2). \tag{21}$$

Under $H_1$ the samples must be from the same alleles; they may differ due to measurement error only. Therefore

$$\begin{aligned} &\text{lik}(x_1,x_2,y_1,y_2|H_1) \\ &\quad = \sum_{i\leq,j} \gamma(i,j)\, \phi_{ij}(x_1,x_2)\, \phi_{ij}(y_1,y_2). \end{aligned} \tag{22}$$

Finally,

$$\mathcal{LR} = \frac{\text{lik}(x_1,x_2,y_1,y_2 \mid H_1)}{\text{lik}(x_1,x_2,y_1,y_2 \mid H_0)} \to \frac{1}{\Pr(\mathcal{G}_e)} \tag{23}$$

as the measurement error goes to zero. Notice that, when measurement error is greater than zero, it is no longer clear whether or not the samples have matching genotypes and hence $\mathcal{LR}$ is never precisely zero.

To calculate $\mathcal{LR}$ an estimate of $\gamma$, the genotype probability distribution, is required. A method for approximating the likelihood ratio was pioneered by Gjertson et al. (1988) for paternity cases using

VNTR's. A few years later the ideas were further developed in three articles (Berry, 1991; Berry, Evett and Pinchin, 1992; and Devlin, Risch and Roeder, 1992).

Devlin, Risch and Roeder (1992, subsequently DRR) take a direct modeling approach to obtain the maximum likelihood estimate of the genotype distribution, $\gamma(\cdot,\cdot)$. An approximation to the likelihood ratio is then obtained by substituting $\hat{\gamma}(\cdot,\cdot)$ into (5) and (22).

Because they assume the alleles are independent, it suffices to estimate the allele distribution and invoke (1). This method requires an estimate of the location of the grid of alleles $\{a(k), k = 1,\ldots,m\}$ and the associated weights $\{\gamma(k), k = 1,\ldots,m\}$. For the VNTR's they studied, the distance between supports $a(k) - a(k-1)$ is a known quantity, hence the support can be estimated with great accuracy. For the purpose of exposition, assume the supports $\{a(k), k = 1,\ldots,m\}$ are known quantities.

Let $(z_{1j} \le z_{2j})$, $j = 1,\ldots,r$, represent the $r$ ordered pairs of observations of fragment lengths in a reference population of size $r$. Recall from (4) that these measurements are comprised of allele pairs convolved with measurement errors $(\varepsilon_1, \varepsilon_2)$ which are assumed to be normally distributed. Assuming H-W and ignoring the correlated measurement error (for the purpose of estimating $\gamma$ only), the likelihood of the data in the reference population is that of a normal mixture,

$$\text{lik}(\gamma) = \prod_{j=1}^{r}\prod_{i=1}^{2}\sum_{k=1}^{m}\gamma(k)\frac{1}{\sqrt{2\pi\sigma_k^2}} \cdot \exp\left\{-\frac{1}{2\sigma_k^2}\left[z_{ij} - a(k)\right]^2\right\}.$$

The maximum likelihood estimate for $\gamma$ is unique and consistent. The properties of this estimator are explored in Devlin, Risch and Roeder (1991b). For a given genotype, what is the effect of estimating $\mathcal{LR}(\gamma)$ with $\mathcal{LR}(\hat{\gamma})$? As one would expect, DRR found that the variance increased as the observed genotype became more rare. Specifically, the variance increased almost linearly as a function of $\mathcal{LR}(\hat{\gamma})$. This had little practical effect, however, as a 95% confidence interval about $\mathcal{LR}(\hat{\gamma})$ continued to be large if $\mathcal{LR}(\hat{\gamma})$ was large.

The DRR approach could also be used to obtain an estimate of the genotype distribution without assuming independence of alleles; however, the variance of $\hat{\gamma}(\cdot,\cdot)$ would be substantially increased, perhaps to the point of having practical impact.

Berry, Evett and Pinchin (1992) obtain a smooth estimate of the joint distribution of $(A_1,A_2)$ by convolving the pairs of observations in the reference population with a bivariate normal kernel, obtaining a continuous estimate of the genotype distribution. The genotype distribution is estimated by

$$H(u,v) = \frac{1}{r}\sum_{i=1}^{r}\mathcal{N}\left\{\begin{pmatrix} u - z_{1j} \\ v - z_{2j}\end{pmatrix}, (bc)^2\begin{pmatrix} z_{1j}^2 & 0 \\ 0 & z_{2j}^2\end{pmatrix}\right\},$$

where $b$ is a smoothing parameter. The authors suggest that setting $b = 1$ accounts for measurement error while setting $b > 1$ accounts for sampling variation as well [but see Chernoff (1991) and DRR].

Berry, Evett and Pinchin (1992, BEP) estimate (21) with

$$2\int\int_{u<v}\phi_{uv}(x_1,x_2)\,dH(u,v) \cdot 2\int\int_{u<v}\phi_{uv}(y_1,y_2)\,dH(u,v)$$

and (22) with

$$2\int\int_{u<v}\phi_{uv}(x_1,x_2)\phi_{uv}(y_1,y_2)\,dH(u,v).$$

By assuming independence across loci, the multilocus $\mathcal{LR}$ can be obtained by multiplying the single locus $\mathcal{LR}$s obtained from either method.

### 6.2 Comparison and Performance

This subsection examines the performance of the various statistical methods for calculating $\mathcal{LR}$. Because of the tremendous variability in the population, $\mathcal{LR} < 10^{-15}$ in the vast majority of cases when the DNA profiles of two randomly selected individuals are compared, even with as few as two loci. For example, using the BEP method and a three-locus system, Evett, Scranage and Pinchin (1993) found that, under $H_0$, $\mathcal{LR} > 0$ in only eight cases out of a million while, under $H_1$, $\mathcal{LR}$ is typically as large as a million and often as large as a billion. Similar results were obtained for other two- and three-locus systems (Berry, Evett and Pinchin, 1992; Devlin, Risch and Roeder, 1992).

The BEP and DRR methods perform similarly, even though they are mathematically quite different in their formulation. While technically it is correct to deconvolve (DRR) rather than convolve (BEP) to obtain an estimate of $\gamma$, the DRR method has the disadvantage of having a larger variance and being more complicated than the BEP method. The BEP method, on the other hand, reduces its variance at the expense of introducing a slight bias: it overestimates the weight of the evidence for common phenotypes and underestimates the weight of the evidence for rare phenotypes.

In principle the methods designed for continuous data are better at discriminating between the two competing hypotheses than the match/binning method because they avoid the ambiguity of declaring a match and hence have zero chance of false negatives. In addition, when $H_1$ holds, the match/binning leads to an estimate of $\mathcal{LR}$ that is typically one to several orders of magnitude smaller than the value of $\mathcal{LR}$ obtained using (23) (Berry, Evett and Pinchin, 1992; Devlin, Risch and Roeder, 1992; Evett, Scranage and Pinchin, 1993). It is difficult to compare the match/binning method to the continuous methods when $H_0$ holds because forensic scientists use subjective and objective criteria to declare matches. In an experiment to assess the probability of false negatives (and to underscore the inadvisability of the NRC report's recommendation to use strict objective criteria), Evett, Scranage and Pinchin (1993) found that a method with a match window of three standard deviations falsely declares an exclusion for 6% of the cases. Under $H_0$, they also found that the match/binning method was less conservative than the BEP method when a strictly objective match criterion was implemented: 27 and 9 cases out of a million obtained $\mathcal{LR} > 1$ for the match/binning and BEP methods, respectively. Using subjective criteria many of these false positives would be removed from consideration (Evett, 1993). Unfortunately, the NRC report calls for objective match criteria, denying the value of expert judgment. In the absence of expert judgment, the match/binning method is an unnecessarily crude statistical tool.

## 7. OPEN ISSUES

In this section, I describe a few topics that have not received extensive coverage in the literature. These topics include the issue of when DNA profiles can be assumed to identify unique genomes, database searching to find suspects, and laboratory error.

### 7.1 Uniqueness

No two people have the same set of dermal fingerprints. This fact is relied upon routinely by the courts despite the fact that fingerprints are only technically unique if a large portion of the pattern is measured. In reality only a small fraction of the pattern (a set number of points) is verified. Likewise, no two people have the same genome (with the possible exception of identical twins). Paralleling dermal fingerprints, sequencing the entire genome for any given crime is not feasible or necessary. Although some loci are considerably more informative than others (Devlin, Risch and Roeder, 1992), much evidence suggests that five locus profiles may be variable enough to approach uniqueness for persons who are not closely related (henceforth, unrelated persons). This invites the question, What constitutes a proof of uniqueness?

This question is partly motivated by the suggestion that the courts use the "one-on-$N$" rule (Lewontin and Hartl, 1991); that is, if the suspect's profile does not match any of the $N$ profiles in the appropriate reference population, then the courts present $1/N$ instead of $1/\mathcal{LR}$. The motivation behind this argument goes as follows. Of the $N$ possible genotypes that have been observed, the suspect does not match any of them; therefore, the probability of observing this genotype in a random draw from the population is conservatively estimated as $1/N$. Contrast this approach with the opposite extreme. Suppose we know that there are $n$ possible genotypes and that they are all equally probable. Then the appropriate probability of drawing any given profile by chance would in fact be $1/n$. If profiles of unrelated persons are unique, then $n \to \infty$.

Some preliminary attempts have been made to estimate how many distinct genotypes are possible for a given battery of loci and how one might evaluate an upper bound for the probability of the most common genotype. If the number of possible genotypes is enormous relative to the population and if none of the genotypes is overwhelmingly abundant, then this evidence supports near uniqueness among unrelated individuals. The original work in this area appears to be due to Risch and Devlin (1992a) and was supported by a careful study by Herrin (1993).

Using techniques described in Section 3.5, Risch and Devlin (1992a) argued that the number of five-locus genotype-equivalents was so large that the probability of a chance match was vanishingly small. [The genotype-equivalent depends on their choice of matching rule—2.4%. They subsequently expanded the matching rule to 5% (Risch and Devlin, 1992b) and reached the same conclusion.] When independence of match probabilities across loci is assumed, they estimated probability of a five-locus match for each of the racial groups studies was approximately $10^{-12}$. If one were to consider genotypes as discrete entities, the number of different genotypes must be at least as great as the inverse of the match probability—the minimum number of genotypes would occur only when the genotypes are all equally likely. Another approach to determining the number of different genotype patterns, which does not rely on independence across loci, is to use the distribution of observed genotypes. This approach is similar to that of estimating the number of unseen species (Fisher, Corbet and Williams, 1943) and gives an estimate which is $O(10)^{11}$. From these analyses, it appears that the number of possible genotypes easily exceeds the total U.S. population size. They also obtained a crude upper bound to the most common genotype of approximately $10^{-6}$.

Herrin (1993) repeated portions of this experiment with other databases. He carefully analyzed the effect of the match window, allowing it to extend up to 20%. The effect of this was a drastic increase in the probability of matches. Nevertheless, he estimated that the number of genotype-equivalents is extremely large.

The question of a proof of uniqueness remains open and depends strongly on the complications of relatedness among individuals. The analyses of Risch and Devlin (1992a, b) and Herrin (1993) are of unrelated individuals only.

On a related topic, the methods discussed in Section 2 apply when the suspect was selected based on nongenetic evidence. By contrast, if a suspect is identified only through evidence obtained from a DNA profile data bank, where a multitude of individuals have their DNA profiles on file (analogous to dermal fingerprint data banks), additional issues arise. Because DNA profiles are not considered unique, the data must be interpreted carefully.

Similar issues have arisen in other court cases, for instance, the Collins case; for details, see Solomon (1982). In this case, a suspect was identified solely based on the fact that he had a number of characteristics in common with the culprit (e.g., model and color of automobile.) This situation differs from cases where the DNA is found to match, after the suspect has been identified using other evidence. It has been argued that the relevant probability in this case is not the probability of observing a suspect with the same characteristics as the culprit due to chance alone. Rather, the appropriate calculation is the probability that at least one more individual exists in the reference population who has the characteristics of the culprit, given that an individual has already been observed with those characteristics.

Instances where a database of DNA profiles has been searched to obtain a suspect have already arisen in Minnesota, Illinois and Virginia. In the Minnesota case, the DNA evidence used to search the database was not entered into court. This approach, while conservative (and recommended by the NRC report), does not make full use of the information. The appropriate statistical solution to this type of problem is an open question.

### 7.2 Laboratory Error

Here we consider a laboratory error to have occurred if, instead of comparing two samples to determine if they match, the same sample was accidentally analyzed twice. For a particular case, information is available that affects the probability of a handling error: chain of custody documentation, number of different profiles analyzed, various controls and other biological information.

There has been very little statistical research on laboratory error rates. The emphasis has been on proficiency testing. Actually there is little information on proficiency testing, and what little exists is frequently misinterpreted. For example, consider the data on laboratory error rates for the California Crime Laboratory Directors tests (California Association of Crime Laboratory Directors, 1988). Based on these tests, it is often suggested that one laboratory's error rate is $1/50 = 0.02$ (e.g., Geisser, 1992). It is true that 50 samples were sent to each testing laboratory, and that one laboratory had one false match; however, the tests were structured so that the laboratory had to draw an inference for all possible pairwise comparisons of the 50 samples. Even ignoring the greater difficulty of these tests, compared to the standard forensic case, the error rate for this lab would be $1/1225 < 0.0008$ if the laboratory indeed made all possible comparisons.

I believe this emphasis on proficiency testing is misplaced. A great deal of useful information concerning laboratory rates under standard conditions is available in the form of paternity suit data. In situations where it is impossible to do proficiency testing (e.g., probability of a space shuttle disaster), methods have been developed for incorporating related error rates into the calculation (e.g., Hartigan 1990). Presumably a method could be constructed by which information from error rates in paternity testing, proficiency testing, and information specific to the case could be combined to form a posterior probability of a handling error for a particular case.

## 8. CONCLUDING REMARKS: CONTRASTING THE VIEWPOINTS OF A STATISTICIAN AND THE NRC PANEL

I have now covered many of the statistical issues inherent in the forensic use of DNA. As I summarized in Section 2.4, a committee of the National Research Council (NRC) recently discussed many of the same issues in their 1992 report. Unfortunately there were no statisticians on the committee, even though a major focus of the report was statistical in nature. The consequences of this omission were numerous minor and major errors regarding statistical theory and interpretation of population genetic results, which are now having an impact on the judicial process (Harmon, 1993). I will summarize the key issues covered in this review by contrasting my interpretation of the data and the literature with the interpretation voiced by the NRC panel.

The NRC report dismisses the statistically based likelihood ratio methods (Section 6) as Bayesian methods that would be unacceptable to the courts, favoring instead the match binning methods devised by forensic scientists (Section 2.3). The report misses the point that the binning methods simply yield discretized approximations to a likelihood ratio (Section 2.1) that are somewhat less efficient (Section 6.2). The arbitrary matching rule is a source of contention in most cases that come to trial. I recommend that the courts move toward adopting a continuous version of $\mathcal{LR}$ so as to avoid unnecessary argument about matching.

To evaluate the significance of matching DNA profiles, the report recommends the ceiling principle, a method designed to be independent of the heritage of the suspect (Section 2.4). This is not a satisfying solution to this complex problem. If the suspect and evidentiary sample can be assumed to be drawn independently, then it is the population of possible perpetrators, not the heritage of the suspect, that is relevant (Section 4.1). Even if the suspect and culprit are assumed to be related, there are better ways to adjust the calculation than the ceiling principle (Sections 4.2 and 4.3).

The other motivation for the ceiling principle of match/binning is population heterogeneity. In fact, population heterogeneity has been the crux of the arguments about quantifying the weight of the evidence. One consequence of population heterogeneity is that it leads to dependencies among events comprising a VNTR profile (Section 4.2). Another consequence is that some DNA profiles will tend to be more common in their cognate subpopulation than in the mixed population (Sections 3.1 and 4.2).

Regarding population heterogeneity, the report relies on only one study. Based on the findings of this study, the NRC report claimed that there are larger genetic differences between individuals of British, French and Italian heritage than there are between individuals of African American, Caucasian and Oriental heritage. This curious finding about human genetics results is based on a 1972 study that used a method to partition genetic diversity that did not account for sampling error and is inherently biased toward providing such a result (Section 5.1). More recent studies demonstrate the opposite is true: there is greater diversity among African Americans, Caucasians and Orientals, but this diversity is dwarfed by individual variability (Section 5.1). The consequence is that DNA profiles obtained from three or more loci are rare in all ethnic groups. Exactly how rare depends somewhat on the choice of reference population. For instance, the range may span one to two orders of magnitude, but such a range will have little practical impact on $\mathcal{LR}$'s as large as several million.

Violations of independence also are indicative of population heterogeneity. Tests of independence rarely detect any deviations from independence (Sections 3.2–3.5). Moreover, when statistically significant deviations are observed, they frequently do not lead to deviations of practical importance (Section 3.6). The report argues, based on a two-allele locus, that statistical tests of independence lack power to detect dependencies induced by population substructure. In fact, such tests (Sections 3.2 and 3.5), while not sensitive to weak dependencies, are powerful if mixture induces meaningful differences between true genotype frequencies and those estimated assuming independence.

Finally, statistical methods to account for any existing heterogeneity are readily developed and spring naturally from population genetic theory (Sections 3.1 and 4.2). I recommend that corrections, when necessary, be based on population genetic and statistical theory. Ad hoc corrections such as the ceiling principle are difficult to justify and sometimes obscure the need for sensible corrections such as those described in Section 4.3 where the suspect and culprit are assumed to be close relatives.

It is my hope that this review has informed the reader about the major statistical issues in this area, as well as the results. Open statistical issues on the forensic use of DNA exist (Section 7), and others will become obvious as the technologies and databases continue to evolve.

## ACKNOWLEDGMENTS

I would like to thank Bernie Devlin, Ian Evett, Jay Kadane, John Hartigan and the Editor, Rob Kass, for helpful comments on a previous draft of the manuscript; John Hartmann, Ivan Balazs and Bruce Budowle for the use of their data; and Ivan Balazs, Bruce Budowle and Rockne Harmon for comments on a draft of my rejoinder. This research was supported by NSF Grants DMS-92-01211 and DMS-92-57006.

## REFERENCES

Baird, M. I., Balazs, I., Guiste, A., Miyazaki, L., Nicholas, L., Wexler, K., Kanter, E.,Glassberg , J., Allen, F., Rubenstein, P. and Sussman L. (1986). Allele frequency distribution of two highly polymorphic DNA sequences in three ethnic groups and its application to the determination of paternity. *American Journal of Human Genetics* **39** 484–501.

Balazs, I. (1993). Population genetics of 14 ethnic groups using phenotypic data from VNTR loci. In *Second International Conference on DNA Fingerprinting* (S.D.J. Pena, R. Chakraborty, J. T. Epplen and A. J. Jeffries eds.). Birkhäuser, Boston.

Balazs, I., Baird, M., Clyne, M. and Meade, E. (1989). Human population genetic studies of five hypervariable DNA loci. *American Journal of Human Genetics* **44** 182–190.

Balding, D. J. and Nichols, R. A. (1994). DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Unpublished manuscript.

Berry D. A. (1991). Inferences using DNA profiling in forensic identification and paternity cases. *Statist. Sci.* **6** 175–205.

Berry, D. A., Evett, I. W. and Pinchin, R. (1992). Statistical inference in crime investigations using DNA profiling: single locus probes. *J. Roy. Statist. Soc. Ser. C* **41** 499–531.

Brookfield, J. (1992). The effect of population subdivision on estimates of the likelihood ratio in criminal cases using the single loucs DNA probes. *Heredity* **69** 97–100.

Budowle, B., Chakraborty, R., Giusti, A. M., Eisenberg, A. J. and Allen, A. J. (1991a). Analysis of the VNTR locus D1S80 by the PCR followed by high resolution PAGE. *American Journal of Human Genetics* **48** 137–144.

Budowle, B., Giusti, A. M., Wayne, J. S., Baechtel, F. S., Fourney, R. M., Adams, D. E., Presley, L. A., Deadman, H. A. and Monson, K. L. (1991b). Fixed bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci for use in forensic comparisons. *American Journal of Human Genetics* **48** 841–855.

Budowle, B. and Monson, K. L. (1992). Perspectives on the fixed bin method and the floor approach/ceiling principle. In *Proceedings of the International Symposium on Human Identification*. Promega Corp., Madison, WI.

Budowle, B., Monson, K. L., Giusti, A. M. and Brown, B. L. (1994). The assessment of frequency estimates of Hae III–generated VNTR profiles in various reference databases. *American Journal of Forensic Sciences* **39** 319–352.

California Association of Crime Laboratory Directors (1988). Report to the Directors.

Chakraborty, R. (1993). DNA forensics and population structure. *Science* **260** 1059–1060.

Chakraborty, R. and Jin, L. (1992). Heterozygote deficiency, population substructure and their implications in DNA fingerprinting. *Human Genetics* **88** 267–272.

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

Chakraborty, R., Srinivasan, M. R. and de Andrade, M. (1993). Intraclass and interclass correlations of allele sizes within and between loci in DNA typing. *Genetics* **133** 411–419.

Chernoff, H. (1991). Comment on "Inference using DNA profiling in forensic identification and paternity cases," by D. A. Berry. *Statist. Sci.* **6** 192–196.

Cockerham, C. C. (1969). Variance of gene frequencies. *Evolution* **23** 72–84.

Cockerham, C. C. (1972). Analysis of gene frequencies. *Genetics* **76** 669–700.

Cohen, J. E. (1990). DNA Fingerprinting for forensic identification: potential effects on data interpretation of subpopulation heterogeneity and band number variability. *American Journal of Human Genetics* **46** 358–368.

Cohen, J. E. (1992). The ceiling principle is not always conservative in assigning genotype frequencies for forensic DNA testing. *American Journal of Human Genetics* **51** 1165–1167.

Cohen, J. E., Lynch, M. and Taylor, C. E. (1 991). Forensic DNA tests and Hardy–Weinberg equilibrium. *Science* **253** 1037–1038.

Cotterman, C. W. (1940). *A Calculus for Statistical Genetics*. Ohio State Univ., Columbus.

Devlin, B. (1993). Forensic inference from genetic markers. *Statistical Methods in Medical Research* **2** 241–262.

Devlin, B. and Risch, N. (1992a). A note on Hardy–Weinberg equilibrium of VNTR data using the FBI's fixed bin method. *American Journal of Human Genetics* **51** 549–553.

Devlin, B. and Risch, N. (1992b). Ethnic differentiation at VNTR loci, with special reference to forensic applications. *American Journal of Human Genetics* **51** 534–548.

Devlin, B. and Risch, N. (1993). Physical properties of VNTR data and its impact on tests of allelic independence. *American Journal of Human Genetics* **53** 324–329.

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

Devlin, B., Risch, N. and Roeder, K. (1991a). Forensic DNA tests and Hardy–Weinberg equilibrium. *Science* **253** 1039–1041.

Devlin, B., Risch, N. and Roeder, K. (1991b). Estimation for allele frequencies for VNTR loci. *American Journal of Human Genetics* **48** 662–676.

Devlin, B., Risch, N. and Roeder, K. (1992). Forensic inference from DNA fingerprints. *J. Amer. Statist. Assoc.* **87** 337–349.

Devlin, B., Risch, N. and Roeder, K. (1993a). Statistical evaluation of DNA fingerprinting: A critique of the NRC's report. *Science* **259** 748–749; 837.

Devlin, B., Risch, N. and Roeder, K. (1993b). NRC report on DNA typing. *Science* **260** 1057–1059.

Devlin, B., Risch, N. and Roeder, K. (1994). Statistical comments on the NRC's report on DNA typing. *Journal of Forensic Sciences* **39** 29–41.

DHEW (1980). Selected genetic markers of blood secretions for youths, 12–17 years of age, United States. Series 11, Number 168. U.S. Dept. Health, Education and Welfare, Hyattsville, MD. (DHEW publication [PHS] 80-1661.)

Evett, I. W. (1992a). DNA statistics: putting the problems into perspective. *Justice of the Peace* **156** 583–586.

Evett, I. W. (1992b). Evaluating DNA profiles in the case where the defense is: It was my brother. *Journal of the Forensic Science Society* **32** 5–14.

Evett, I. W. (1993). Personal communication.

Evett, I. W., Scranage, J. and Pinchin, R. (1993). An illustration of the advantages of efficient statistical methods for RFLP analysis in forensic science. *American Journal of Human Genetics* **52** 498–505.

Evett, I. W. and Weir, B. S. (1992). Flawed reasoning in court. *Chance* **4** 19–21.

Fisher, R. A., Corbet, A. B. and Williams, C. B. (1943). The relationship between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* **12** 42–58.

Geisser, S. (1990). Some remarks on DNA fingerprinting. *Chance* **3** 8–9.

Geisser, S. (1992). Some statistical issues in medicine and forensics. *J. Amer. Statist. Assoc.* **87** 607–614.

Geisser, S. and Johnson, W. (1992). Testing Hardy–Weinberg equilibrium on allelic data from VNTR loci. *American Journal of Human Genetics* **51** 1084–1088.

Gjertson, D. W., Mickey, M. R., Hopfield, J., Takenouchi, T. and Terasaki, P. (1988). Calculation of the probability of paternity using DNA sequences. *American Journal of Human Genetics* **43** 860–869.

Green, P. and Lander, E. S. (1991). Forensic DNA tests and Hardy–Weinberg equilibrium. *Science* **253** 1038–1039.

Harmon, R. P. (1993). DNA evidence. *Science* **261** 13.

Hartigan, J. A. (1990). Partition models. *Comm. Statist. Theory Methods* **19** 2745–2756.

Hartl, D. L. and Lewontin, R. C. (1993). NRC Report on DNA typing. *Science* **260** 473–474.

Hernandez, J. L. and Weir, B. S. (1989). A disequilibrium coefficient approach to Hardy–Weinberg testing. *Biometrics* **45** 53–70.

Herrin, G., Jr. (1993). Probability of matching RFLP patterns from unrelated individuals. *American Journal of Human Genetics* **52** 491–497.

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

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

Kaye, D. H. (1988). What is Bayesianism? In *Probability and Inference in the Law of Evidence* (P. Tillers and E. Green, eds.) 1–19. Kluwer, Hingham, MA. [Reprinted in *Jurimetrics* **28** (1988) 161–177.]

KAYE, D. H. (1991). Comment on "Inference using DNA profiling in forensic identification and paternity cases," by D. A. Berry. *Statist. Sci.* **6** 196–200.

KAYE, D. H. (1993). DNA evidence: probability, population genetics and the courts. *Harvard Journal of Law and Technology* **7** 101–172.

LANDER, E. (1989). DNA fingerprinting on trial. *Nature* **339** 501–505.

LANDER, E. S. (1991a). Research on DNA typing catching up with courtroom applications. *American Journal of Human Genetics* **48** 819–823.

LANDER, E. S. (1991b). Lander reply. *American Journal of Human Genetics* **49** 899–903.

LATTER, B. H. D. (1980). Genetic differences within and between populations of the major human subgroups. *The American Naturalist* **116** 220–237.

LEVENE, H. (1949). On a matching problem arising in genetics. *Ann. Math. Statist.* **20** 91–94.

LEWONTIN, R. C. (1972). The apportionment of human diversity. *Evolutionary Biology* **6** 381–398.

LEWONTIN, R. C. (1993). Which population? *American Journal of Human Genetics* **52** 205.

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

LI, C. C. (1969). Population subdivision with respect to multiple alleles. *Annals of Human Genetics* **33** 23–29.

LINDLEY, D. V. (1977). A problem in forensic science. *Biometrika* **64** 207–213.

MORTON, N. E. (1992). Genetic structure of reference populations. *Proc. Nat. Acad. Sci. U.S.A.* **89** 2556–2560.

MORTON, N. E., COLLINS, A. and BALAZS, I. (1993). Bioassay of kinship for hypervariable loci in blacks and caucasians. *Proc. Nat. Acad. Sci. U.S.A.* **90** 1892–1896.

MOURANT, A. E. KOPEC, A. C. and DOMAINEWSKA-SOBCZAK, K. (1976). *The Distribution of Human Blood Groups and Other Polymorphisms*. Oxford Univ. Press.

NATIONAL RESEARCH COUNCIL (NRC). (1992). DNA typing: Statistical basis for interpretation. In *DNA Technology in Forensic Science* 74–96. National Academy Press, Washington, DC.

NEI, M. and ROYCHOUDHURY,, A. K. (1982). Genetic relationships and evolution of human races. *Evolutionary Biology* **14** 1–59.

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

NIEDERMAN, M. D., GILBERT, E. C. and SPIRO, H. M. (1962). The relationship between blood pepsin level, ABO blood group, and secreter status. *Annals of Internal Medicine* **4** 564–569.

RISCH, N. and DEVLIN, B. (1992a). On the probability of matching DNA fingerprints. *Science* **255** 717–720.

RISCH, N. and DEVLIN, B. (1992b). DNA fingerprint matches. *Science* **256** 1744–1746.

ROEDER, K., ESCOBAR, M., KADANE, J. and BALAZS, I. (1993) Measuring heterogeneity in forensic databases: a hierarchical approach to the question. Technical report, Carnegie Mellon Univ.

SHAPIRO, R. (1991). *The Human Blueprint: The Race to Unlock the Secrets of our Genetic Script*. St. Martin's, New York.

SMOUSE, P. E., SPIELMAN, R. S. and PARK, M. H. (1982). Multiple-locus allocation of individuals to groups as a function of the genetic variation within and the differences among populations. *The American Naturalist* **119** 445–463.

SOLOMON, H. (1982). Measurement and burden of evidence. In *Some Recent Advances in Statistics* (J. Tiago de Oliveira and B. Epstein, eds.) 1–22. Academic, New York.

STEINBERGER, E. M., THOMPSON, L. D. and HARTMANN, J. M. (1993). On the use of excess homozygousity for subpopulation detection. *American Journal of Human Genetics* **52** 1275–1277.

WEIR, B. S. (1990). Genetic data analysis. Sinauer Associates, Sunderland, MA.

WEIR, B. S. (1992a). Independence of VNTR alleles defined by fixed bins. *Genetics* **130** 873–887.

WEIR, B. S. (1992b). Independence of VNTR alleles defined as floating bins. *American Journal of Human Genetics* **51** 992–997.

WEIR, B. S. (1992c). Population genetics in the forensics debate. *Proc. Nat. Acad. Sci. U.S.A.* **89** 11,654–11,659.

WEIR, B. S. (1993). Forensic population genetics and the NRC. *American Journal of Human Genetics* **52** 437–439.

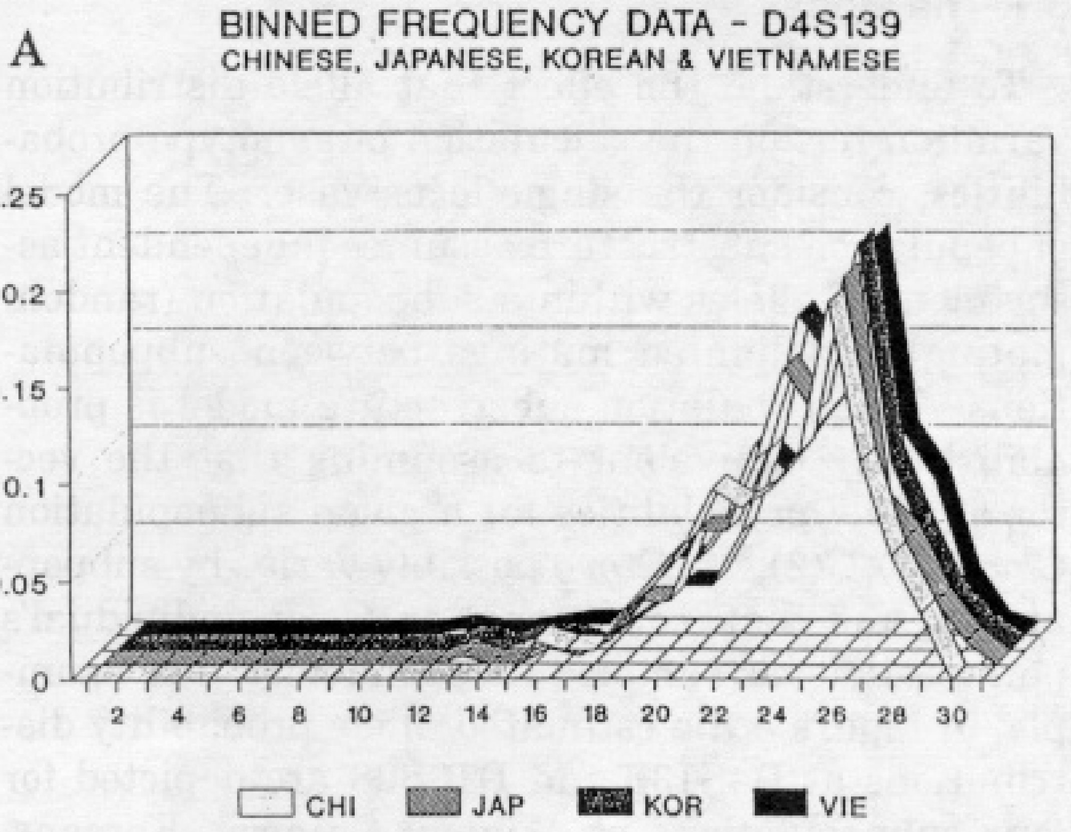
WEIR, B. S. (1994). The effects of inbreeding on forensic calculations. *Annual Review of Genetics* **28**. To appear.

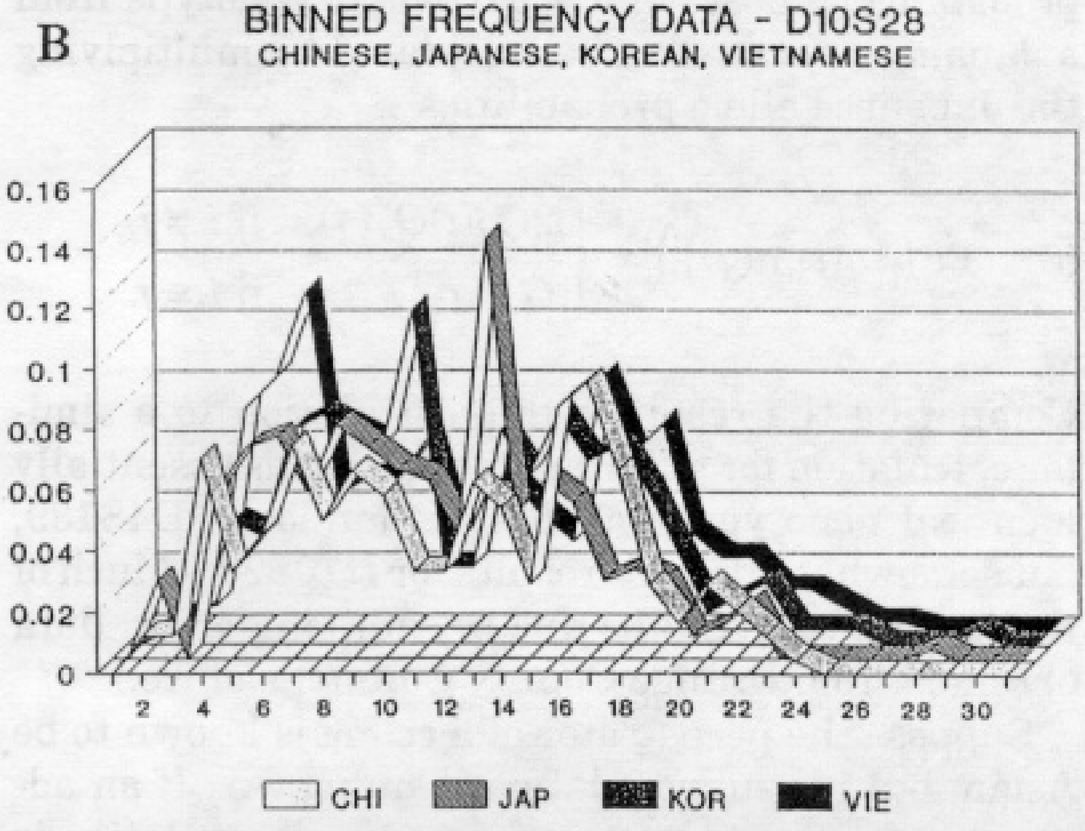
WEIR, B. S. and COCKERHAM, C. C. (1984). Estimating *F*-statistics for the analysis of population structure. *Evolution* **38** 1358–1370.

WEIR, B. S. and EVETT, I. W. (1992). Whose DNA? *American Journal of Human Genetics* **50** 869.

WRIGHT, S. (1951). The genetical structure of populations. *Annals of Eugenics* **15** 323–354.

WYMAN, A. R. and WHITE, R. (1980). A highly polymorphic locus in human DNA. *Proc. Nat. Acad. Sci. U.S.A.* **77** 6754–6758.


A
BINNED FREQUENCY DATA - D4S139
CHINESE, JAPANESE, KOREAN & VIETNAMESE
0.25
0.2
0.15
0.1
0.05
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
CHI
JAP
KOR
VIE


B
BINNED FREQUENCY DATA - D10S28
CHINESE, JAPANESE, KOREAN, VIETNAMESE
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
CHI
JAP
KOR
VIE