Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 11654–11659, December 1992

Review

Population genetics in the forensic DNA debate

B. S. Weir

Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh NC 27695-8203

ABSTRACT **The use of matching variable number of tandem repeat (VNTR) profiles to link suspects with crimes is potentially very powerful, but it has been quite controversial. Initial debate over laboratory procedures has largely given way to debate over the statistical and population genetic issues involved in calculating the frequency of a profile for a random member of a population. This frequency is used to weight the evidence of a match between suspect and crime scene material when the suspect denies responsibility for that material. A recent report from the National Research Council, intended to put to rest some of the issues, has instead raised further debate by advocating a procedure based on maximum frequencies of profile components over several different populations.**

No two people have the same set of fingerprints, and this fact has proved invaluable in identifying criminals. Likewise, no two people who are not identical twins have the same sequence of base pairs in their DNA, and this has equal forensic potential. Although complete DNA sequencing of a human is a goal of the human genome project, it is not likely to become a routine endeavor. Fortunately, there are features of DNA sequences that may prove to be virtually unique to individuals, and the one that currently is being used for human identification rests on regions in which there are variable numbers of short tandem repeats (VNTRs). The number of repeat units for a particular system can range from a few to a few hundred, so that any VNTR locus can exist in one of several hundred forms, or alleles. A collection of several VNTR loci leads to many millions of different combinations, or DNA profiles, and hence to the possibility of distinguishing between people.

When Jeffreys and his colleagues (1–3) introduced VNTRs as a means of human identification in 1985 they thrust the fields of population and statistical genetics into unaccustomed prominence. The debate that has followed, starting with Lander's commentary (4) following his experiences with the celebrated Castro case, and attracting widespread attention in two opposing articles in *Science* late in 1991 (5, 6), has been phrased in very strong language. Even in this journal, statements such as "To protect the suspect, the Lander and Lewontin–Hartl defenses argue that the calculation of matching probability should be absurdly conservative" (7) have appeared.

Language has been harsh because, in some instances, the inferences drawn from VNTR profiles have been used in the process of convicting or acquitting persons accused of violent crimes. Arguments have arisen that could have been avoided if the deliberate pace with which scientific investigation usually proceeds had been applied to the forensic uses of DNA evidence. The chief complaint of opponents is that "The disjunction between scientific and judicial standards of evidence has allowed novel forensic methods to be used in criminal trials prematurely or without verification" (8).

This article traces the arguments surrounding the population and statistical genetic issues arising in the use of VNTR profiles for human identification. This use has faced several other challenges (4, 9) dealing with laboratory procedures, the degradation of DNA in samples, and so forth, but such experimental issues will not be considered here. Reference can be made to refs. 10–13. Particular attention here will be paid to a report issued by a committee established by the Board on Biology of the National Research Council. This NRC report (14) was intended to meet calls from the scientific and legal communities for an examination of the forensic uses of DNA.

It is somewhat surprising that there has been so much debate over the statistical questions, because they are neither complex nor subtle. That they are not well understood, however, becomes apparent when court decisions are issued containing such statements as "Appellant contends that the database of 710 samples is too small to be statistically significant" (15). A court may be excused for such nonstatistical language, but a report issued by the NRC (14) that contains (p. 76) "the traditional 95% confidence limit, whose use implies the true value has only a 5% chance of exceeding the upper bound" must lose some credibility with statisticians. Berry *et al.* (16), in criticizing the match/binning procedure described below, state: "As far as we know there were no statisticians involved in developing match/binning. This is hardly the first time that science has taken a wrong turn for lack of guidance from statisticians." The same theme was sounded by Geisser (17): "It is unconscionable that the three laboratories that provide almost all of these analyses have not yet seen fit to employ statistical help in improving their procedures." Fortunately, Geisser's suspicion that the laboratories "have little interest in testing their independence assumptions or having others do it" (18) is put to rest by several recent publications (19–31).

Most of the debate has taken place at pretrial hearings held to determine the admissibility of DNA evidence. These "Frye" hearings ("Kelly–Frye" hearings in California) are based on guidelines established in 1923. The court has a set of guidelines to determine whether novel scientific evidence should be admissible, and prominent among the criteria is the need for general acceptance in the relevant scientific community. Both "general acceptance" and "relevant community" are open to interpretation. It has been the practice for the defense to introduce the results of a small telephone survey to bolster the claim that the majority of population geneticists do not support current methods of calculating population frequencies, and for the prosecution to respond that none of the surveyed critics has performed any calculations on forensic databases to support their criticisms. This debate runs counter to the usual scientific practice of referring to published analyses and interested parties having the opportunity to verify these analyses independently.

As recently as August 1992 an appeal court in California felt compelled to say, "We conclude that one element of the current DNA analysis—the determination of the statistical significance of a match between a defendant's DNA and the DNA in bodily material found at the crime scene—does not satisfy the Kelly–Frye test" (32). The court apparently based this decision on the NRC report (14), two opposing articles (5, 6), and on a *Science* reporter's news story (33), and concluded that the statistical calculations were not generally accepted by popula-

Abbreviations: FBI, Federal Bureau of Investigation; NRC, National Research Council; VNTR, variable number of tandem repeats.

tion geneticists. If the arguments put forth in one peer-reviewed paper (5) are to receive such weight it is difficult to see how the Kelly–Frye test could ever be satisfied. There will always be articles published that are critical of any scientific procedure.

ESTIMATING FREQUENCIES

What are the issues that have proven so contentious? At some crime scenes, samples of biological tissue such as blood, semen, hair follicles, bone, etc. are found, and these may be processed to yield DNA. One of the real advantages of the system is that DNA may be extracted from blood or semen stains that are several years old. Investigators seek to identify the person who left the biological sample by DNA matching. Generally the person sought is the perpetrator of the crime, and a match is thought to place a suspect at the crime scene. In other cases the material may be from the victim, and a match is thought to place the victim at a certain location. In the simplest analyses, failure to match DNA types between a suspect and crime scene material means that, if just the biological evidence is considered, the suspect is exonerated. The possibility of such exonerations, which are quite common, is widely applauded, although more sophisticated analyses do not necessarily give such a definite conclusion (see below). If, however, the suspect has a DNA profile that matches that of the crime scene material, then that person is not excluded from having provided the material. Two questions arise immediately: what is the basis for declaring that the two DNA profiles match? and how frequent is that profile in the population? Could the match be a spurious consequence of experimental error, or could the match have arisen purely by chance if some person other than the suspect left the crime scene material?

Even if the frequency is calculated correctly, Thompson (34) has described the care needed in presenting the value to a jury. A frequency of 1 in 1000 for a VNTR profile does not translate into a probability of guilt of 0.999, nor does the fact that there are expected to be 1000 such people in a city of size 1,000,000 negate the probative value of the frequency of the matching profile.

Matching Criteria. The use of highly variable marker systems for identification faces a dilemma. The more variants there are at a locus, the less likely it is that two random members of a population will match. For VNTR systems in which each locus has 20 variants, a profile based on 4 loci has 1.9×10^9 variants, while one with 8 loci would have 3.8×10^{18} variants, and this is likely to allow almost certain identification. Numerical analysis (29) on forensic databases has confirmed the very large number of DNA profiles in U.S. populations. Such figures have led Evett (35) to say, "The DNA profiling is so highly discriminating that in the event of a match the accuracy with which one estimates the population frequency is of minor importance." Be that as it may, the dilemma arises because of the difficulty of distinguishing between all the variants at each locus. There are many experimental reasons why it is not possible to infer the number of repeat units with certainty. Fragments of lengths 2000 base pairs (bp) and 2020 bp may not be separated on an electrophoretic gel, even if they may represent 200 and 202 copies of a repeat of length 10 bp.

The simplest analyses rest on "match/binning," which supposes that the true length of a fragment of estimated length x is enclosed by the interval $x \pm \delta$. Two fragments that have overlapping intervals are said to match; otherwise they do not match. It is generally accepted (16) that the standard deviations of measurement errors are approximately proportional to fragment length: the Federal Bureau of Investigation (FBI) uses a δ value of 2.5% of x, based on empirical studies involving repeated measurements of the same material, while Lifecodes Corporation uses a figure of $\delta = 1.8\%$ of x, and Cellmark Diagnostics chooses δ values corresponding to a gel migration distance of 1 mm. Once the existence of experimental error is recognized, it ceases to be possible to speak in terms of certainty. Two fragments can be said not to differ in length with a specified level of confidence if their appropriate confidence intervals overlap. An acceptable confidence level is needed, and the distribution of errors must be determined empirically. Measurement errors are correlated (16), so that determinations of matches for different pairs of fragments are not independent events. Once a match has been declared, the frequency of the matching fragments is found from all entries in a database that fall into a "bin" surrounding that fragment length. Care is needed in setting up criteria with appropriate correspondence for the matching and the binning procedures. Matching, on DNA connected with a particular crime, with the attendant problems of possible contamination and degradation, uses intragel comparisons. Binning, using DNA collected from blood under controlled conditions for establishing databases, is based on intergel comparisons.

Profile Frequencies. Determining the frequency of a profile found to match between crime scene material and the person of interest is also a statistical issue, since the DNA profiles of all relevant people are not generally available. An estimate of the frequency in the population is needed. The naive "counting" estimate is not very helpful, since it is very unlikely that any particular profile based on several loci will ever be seen in any sample. Although there is some value in simply presenting the fact that a profile has not been seen in, say, any of 1000 people examined to date, a more informative way (36) of presenting the same information is to say that nonoccurrence in a sample of 1000 implies that it is 99% likely that the true frequency is less than 1 in 218. This is preferable to the recommendation of the NRC report (14) and others (5) that a bound of the reciprocal of the sample size be used. The confidence limits are sometimes used by the defense as estimates of actual profile frequencies, which deflects attention from the fact the same answer follows from the nonappearance of a profile in a sample whether that profile is based on one or several loci. Counting estimates ignore the fact that a DNA profile consists of several components. Figures such as 1 in 218 are seen to be misleading when it is realized that, even if everyone in the world had the same two parents, who were heterozygous for different alleles at four independent loci, the frequency of any particular 4-locus profile would be 1 in 256. A better method of estimating frequencies is needed.

The match/binning procedure estimates profile frequencies as the product of the frequencies of the components of the profile. At one locus, this rests on the Hardy–Weinberg law of population genetics. Much time in court was wasted by a failure of both the prosecution to establish that Hardy–Weinberg, or independence of allelic frequencies at a locus, held in their databases and the defense to show that it did not. Early trials did not consider any empirical evidence on the issue, and in other cases analogies were drawn to analyses with other genetic markers. Matters were not helped by repeated reference (4, 25, 37, 38) to tests for consistency of total homozygosity (two equal alleles at a locus) with Hardy–Weinberg expectations. It was not total homozygosity that was at issue. The problem was that population geneticists had little experience in testing for Hardy–Weinberg for loci with many alleles, and that the databases generally contained few of the possible combinations of alleles even at single loci. It is only recently that demonstrations of Hardy–Weinberg in forensic databases have been published (22, 26, 30, 31). Other analyses, based on sampling approaches to exact tests (39) or on shuffling methods, are also showing overall consistency with the Hardy–Weinberg expectations. It is of comfort to realize that the greater demands being placed on statisticians by geneticists are being accompanied by

more powerful computer-intensive methods of analysis (40).

The difficulties of establishing independence between pairs of alleles at one locus are magnified when independence between all $2m$ alleles for m loci is needed. Analyses have been published—e.g., refs. 30 and 31—but numerical shuffling methods may be better. The prosecution has generally relied on an assumption of independence of frequencies of alleles at loci on different chromosomes.

Population Structure. One situation known to lead to dependence, or disequilibrium, between alleles within and between loci is known as the Wahlund effect. It results whenever the population sampled consists of a number of subpopulations with different allelic frequencies. Even if there is independence within each subpopulation, when frequencies are calculated at the population level, disequilibrium will result. Simply put, the product of averages is not the same as the average of products. For a while an unnecessary debate centered on whether or not substructuring could be detected by tests for disequilibrium, and this not die until it was shown empirically that these tests have very low power for detecting population substructure.

Opponents of current practice argue that substructuring invalidates use of the product rule, while proponents argue that the issue should not be whether there is substructuring in the population, but whether any substructuring has an appreciable effect on forensic calculations (41). Neither side has been blessed with an abundance of data, and indeed collecting such data may not be feasible. The argument for the effects of substructuring received most prominence in ref. 5 and was as follows: If two individuals match at one or two loci, this constitutes evidence that they belong to the same subpopulation and so will be likely to match at additional loci. The probability of a match is therefore higher than if the individuals (the perpetrator and the suspect) were drawn randomly from the entire population. There may be cases (42) in which there is prior knowledge that perpetrator and suspect do indeed come from the same subpopulation, and then population-wide estimates could be misleading. The counter to this argument (6) has been that the data do not support the notion that such effects will be of forensic significance. True probabilities may be over- or underestimated by two orders of magnitude, but this is unlikely to prevent the probability of a chance match at several loci remaining very small. Estimates of population subdivision parameters have been small within national populations (7).

Lewontin and Hartl (5) provided an argument of why the U.S. Caucasian population, for example, cannot be regarded as being genetically homogeneous. This population is derived from genetically diverse European subpopulations in recent generations and, at least initially, tended to maintain this historical separateness in marriage. Taking this argument to its logical conclusion, they say, "each particular person may require a different reference group composed of appropriate ethnic or geographic subpopulations." This appears to be heading toward the error made in a recent court decision (43) that said frequency calculations should be based on the distinctive ethnic background of the defendant, who was of mixed Italian, French, and American Indian descent. However, the calculations must apply to the unknown perpetrator, who is assumed different from a defendant who has pled not guilty, and for whom the ethnic details are unknown. Calculations are necessarily based on random people—and are essentially averages over all possible backgrounds. In their rebuttal to the Lewontin and Hartl paper, Chakraborty and Kidd (6) point out that " 'binned allele frequencies' are unbiased estimates, of the averages of all underlying ethnic or endogamous subgroups contained within the reference population." The claim that the ethnic background of the defendant is relevant is one of the most persistent fallacies in the debate (44), even though its fallaciousness has been documented [14 (p. 13), 42, 45].

As with most aspects of the debate, arguments raised by defense experts concerning the effects of population subdivision have generally been made without reference to forensic data. Thus Cohen (46) "attempts to quantify the strength of his effect by choosing extreme values for subpopulation allelic frequency variation" (47). Brookfield (47) expounds the pragmatic line used by prosecution experts, "It is impossible to find a case in which a likelihood ratio constituting strong evidence against a subject is converted [by allowing for population subdivision] to one in which the evidence has become weak."

Calculations of the frequencies of a DNA profile in a population are made to quantify the likelihood of a defendant's claim that some other person left the crime scene material. No calculations are performed if the defendant admits responsibility for the material. Given the defendant's claim, it is necessary to consider the collection of people who could have left the material. There may be evidence, such as from an eyewitness, pointing to a particular ethnic group, or it may be reasonable simply to consider a certain geographic location. The ideal data would result from a census of the population of possible perpetrators, as when blood samples were taken from 4583 men during an investigation of two rape/murders in Leicestershire (48) and the perpetrator was identified uniquely. The next-best data come from a sample that is representative of that population.

Current calculations are based on samples from paternity casework (e.g., Lifecodes Corporation) or from blood banks (e.g., FBI and Cellmark Diagnostics). Another source being investigated is umbilical cord blood from maternity hospitals. Paternity data tend to be from a wide geographic area, and blood bank or hospital data from a narrow region. There have been arguments about the ethnic composition of people becoming involved in paternity disputes or in donating blood, and these are unlikely to be resolved. A counter to the charge of biased samples is to increase the sampling frame, and this is being done by having samples collected from several states and coordinated by the Technical Working Group on DNA Analysis and Methods (TWGDAM). This increased sampling is geographically based, rather than being targeted on specific ethnic groups that might have distinct allele frequencies or, worse, unique alleles. Ethnic sampling is at the heart of the "ceiling principle" advocated by the NRC report (14).

Ceiling Principle. Suppose a matching profile has component alleles indexed by j and that the appropriate population for which to estimate the profile frequency is I. Then, if alleles j have frequencies p_{Ij} that are independent within and between loci in population I, an estimate of the profile frequency in that population is the product $\Pi_j p_{Ij}$, omitting factors of 2. The problem is that the precise population I is either unknown or not sampled. Instead current practice uses frequencies $\bar{p}_j$ of a larger population of which I is presumed to be a component. If the product $\Pi_j \bar{p}_j$ is less than $\Pi_j p_{Ij}$, this underestimation of the profile frequency would be prejudicial to the defendant. To overcome this problem, the NRC report notes that the true frequency is less than the product $\Pi_j \max_i(p_{ij})$, where $\max_i(p_{ij})$ is the maximum frequency of allele j over all populations i. This product of maxima serves as an upper bound for the product of unknown frequencies.

When the information is available, forensic calculations could be based on frequencies within each racial group and the proportion of the relevant population comprised of each racial group (49). The information required is usually not available, however.

To implement the ceiling principle in practice, the NRC report [14 (p. 13)] suggests sampling 100 individuals from 15 to 20 "relatively homogeneous genetically" populations. Agreement would be needed that this collection of data was an adequate basis for calculating upper bounds on profile frequencies applicable

to any population. There has been some hope, especially on the part of the courts (32), that the ceiling principle offers a compromise solution acceptable to both opponents and proponents of current methods, but there are several difficulties. The NRC report failed to stress the necessity of independence of allele frequencies (and on p. 83 implies independence is not necessary), whereas Cohen (50) has shown that, if there is disequilibrium between loci, the ceiling principle can underestimate the true frequency instead of providing an upper bound. The same result can, under certain conditions, happen when there is Hardy–Weinberg disequilibrium. Tests for independence of allelic frequencies at multiallelic loci will have low power in samples of 100 individuals and will not be possible at all in the populations *I* not identified or sampled. Almost no guidance was given to how genetically homogeneous populations are to be identified or what relevance they may have to, say, the current U.S. population. The NRC report does mention (p. 84) "the range of ethnic groups that are represented in the United States—e.g., English, Germans, Italians, Russians, Navahos, Puerto Ricans, Chinese, Japanese, Vietnamese and West Africans," although such groups are themselves likely to contain subgroups. No guidance was given to handle the situation, likely to be common, where independence of all alleles cannot be established at all loci in every database. A conservative procedure for a single database (30) is to discount one of the alleles of each pair in disequilibrium. The logical extension of this under the ceiling principle is to omit one allele at every locus for which there is evidence of disequilibrium in even one database, and this can lead to a halving of the number of alleles, and a substantial increase in the estimated profile frequency. A modified ceiling principle, which takes the maximum of the profile frequencies from each database, would remove this problem.

As an interim measure, the NRC report suggests the reporting of profile frequencies based on three major ethnic data bases, such as Black, Caucasian, and Hispanic—which is the strategy already in use. Critics of the NRC report have pointed out that it may be more reasonable to use the product $\max_i(\Pi_j p_{ij})$ as a ceiling, as this preserves the property that a VNTR profile belongs to one individual, not a collection of individuals with possibly different ethnic backgrounds. Empirical studies (23, 30) show that estimated profile frequencies, when based on several loci, differ very little when calculated from different geographic samples within a racial group, or even from different racial groups. The differences in individual allele frequencies between different subpopulations may not cause meaningful differences between profile frequencies (51): "the use of many loci means that one aberrant fragment size frequency may not be too important," and comparing *profile* frequencies over several databases may obviate the need for further sampling of vague and limited extent. "We conclude that mixture has little or no impact on the use of VNTR loci for forensics" (23).

The merits of the ceiling principle were somewhat diluted by the NRC report suggesting *ad hoc* frequencies of 5% or 10% if estimated frequencies lay below these values. Frequencies above these values are to be replaced by upper confidence limits. By these rules, an allele seen 20 times in a sample of 100 people would be assigned a frequency of 0.10, whereas one seen 21 times would be assigned a value of 0.14. When the 15 to 20 homogenous groups are sampled, the NRC report (p. 93) says it will be "unnecessary to take an upper 95% confidence limit for each allele frequency." Why limits are not needed for samples of size 100 was not explained. Setting confidence limits on estimated profile frequencies is the best way to convey the effects of the sizes of the databases on which the estimates are based. A confidence limit of a product, however, is not the product of the confidence limits.

Although there is an obvious appeal to the legal profession of being able to cite a report prepared under the aegis of the National Academy of Sciences, the fact that the report is perceived as having some flaws (66) means that ultimately the courts are going to have to rely on the scientific literature and expert witnesses. Even though there are many aspects of the NRC report that are not controversial, there is no move to implement its recommendations in any systematic way. There are no current plans, for example, to create the National Committee on Forensic DNA Typing called for in the report. Because of a perceived bias in the report against current practice, it appears that prosecution expert witnesses will challenge the report, while defense expert witnesses will seek out particular sets of data with high frequencies and/or disequilibrium for the alleles in the matching profile of interest. The effects of such selective choices of databases would be diminished by use of the ceiling principle modified to use the maximum over populations of products of allele frequencies.

Continuous Analyses. Many of the population and statistical genetic problems have arisen because of the attempt to apply discrete data techniques to data that are essentially continuous. Estimated fragment lengths do not fall into distinct classes, and the match/binning approach must therefore be approximate. Quite a literature has developed making this point—e.g., refs. 26 and 52—but it has had no impact on forensic practice in the U.S. Continuous analyses are based on the distribution of measurement errors for fragment lengths. Whether or not these errors are normally distributed has been debated, but that is a question that is easy to answer empirically. Joint distributions of errors for two or more fragments can also be found empirically and used in calculations that do not require assumptions of independence. As with the match/binning procedure, however, this still becomes problematical for sets of many alleles. Once the continuous viewpoint is adopted, it is no longer relevant to speak of matching or mismatching. Instead probability levels can be attached to the joint patterns and a jury could draw its conclusions.

Results from continuous analyses are often expressed as likelihood ratios, in statements such as "The evidence that the crime scene material and the suspect both have the particular DNA profile seen is one million times more likely to have arisen if the suspect provided the crime scene material than if some random unrelated person provided the material." The NRC report (p. 62) doubts that the logic underlying this simple statement could be conveyed to a jury, but this seems to be a challenge that the statistical profession could meet. One reason for the reluctance to adopt continuous analyses is that results are often presented in Bayesian language, and there continues to be disagreement within the statistical profession on the merits of Bayesian analyses. Bayesians use the likelihood ratio to convert the ratio of prior probabilities of guilt into the ratio of posterior probabilities, whereas frequencies feel uncomfortable with prior probabilities of guilt. Bayesian language is used in paternity suits, where prior values of 50% are routinely applied to the probability that the alleged father is the true father. It should be said, however, that use of VNTR information can result in very high posterior probabilities whether priors of 10% or 90% are used. The paternity index (i.e., likelihood ratio) is very high when the alleged father is the father, in the same way that the likelihood ratio in forensic calculations is very high when the suspect did leave the crime scene material. It is not often mentioned that the reciprocal of the population frequency calculated by the match/binning procedure is equivalent to the likelihood ratio.

Another use of the likelihood ratio has been proposed by Evett and Werrett (53) as a means of conveying the strength of the evidence to a jury. These authors would say that any likelihood ratio over 1000 constitutes "very strong" evidence, with lesser values having corresponding terms, down to "weak" for ratios be-

tween 1 and 33. This approach appears to avoid arguments over frequencies. Whether or not frequencies may be wrong by a factor of 100 does not matter beyond the "very strong" limit. The same point was made by Lewontin and Hartl (54), from a different perspective: "After all, 0.0001 is already a pretty small number. Why invoke unsupported assumptions to obtain a still smaller number that is exaggerated and unreliable?" In the end, however, there may be little difference between quoting frequencies or verbal descriptors if the defense asks the prosecution exactly what is meant by "very strong" or if the prosecution claims that there is little significance to a difference in frequencies that are either 10^{-8} or 10^{-10}.

In a comparison of the likelihood ratio and match/binning approaches, Berry *et al.* (16) considered that the chief defect of the match/binning method is that it has a large false exclusion rate. This is unlikely to be a criticism raised in court by the defense. It must be pointed out, however, that the match/binning procedure described by Berry *et al.* is not the same as practiced by the FBI (55).

The match/binning approach expresses the final result in terms of P, the population frequency of the matching profile, and $1/P$ is the ratio of the probability of the evidence of the matching profile given that the suspect left the crime scene material to the probability of the evidence if some unrelated person left that material. As Hagerman (56) explains, this ratio can be diminished if the typing laboratory may have falsely declared a match. If this false positive rate is α, then the likelihood ratio is diminished to $(1 - \alpha)/(P + \alpha - 2P\alpha)$. This statement may be more helpful than that of Lempert (57), who points out that the probability with which a laboratory declares a match when the suspect did not leave the crime scene material cannot be less than the false-positive rate for that laboratory regardless of the value of P. For $\alpha = 0.001$, the likelihood ratio when P is 1 in 1,000,000 would change from 1,000,000 to about 1000. There has been some debate (56, 58) on the actual error rates in proficiency tests, and the NRC report makes a strong case for the continuing need for such tests. Establishing rates of false positives or false negatives may not be easy.

DISCUSSION

There is widespread agreement that DNA profiles offer the possibility of human identification, "even with today's technology, which uses 3–5 loci, a match between two DNA patterns can be considered strong evidence that the two samples came from the same source," [14 (p. 74)]. There has been disagreement over the figures quoted for frequencies of such patterns, although it is difficult to escape the conclusion that the disagreement is more apparent than real. Just as the legal system is adversarial, so does the scientific press look for controversy. Both *Science* and *Nature* have carried news items with provocative headlines [e.g., "Forensic tests proved innocent" (59) or "FBI gives in on genetics" (60)], and the role of *Science*'s news stories in influencing court decisions has been documented above. The most disturbing example was the story in the *New York Times* of April 14, 1992, which stated that the NRC report had called for a moratorium on the use of DNA evidence in forensics. This story was retracted on April 15, 1992, and was refuted on page x of the report. As with all scientific issues, papers have been published (61) that have attracted strong rebuttals (35), or papers have not been published when they might have cleared up misunderstandings.

One way to avoid some of the population and statistical genetic issues is to use marker systems that avoid ambiguities in typing. Use of two-allele systems, particularly those of the presence/absence type, will be much easier to score (62) and to check for disequilibrium. Gaining sufficient markers to approach the number of types found with VNTR markers will require the use of PCR to provide enough DNA from very small crime scene samples, but such work is being reported (63). Using PCR and loci with more than two alleles, but with good possibilities for distinguishing variants, was described by Budowle *et al.* (64). Another possibility of obtaining discrete data is to look for patterns of variant repeat units along minisatellite alleles as reported by Jeffreys *et al.* (65).

With the present data, however, it is clear that the final word has not been said.

This paper was improved after comments were received from Wyatt Anderson, Bruce Budowle, Michael Clegg, Newton Morton, and Terence Speed. My opinions are not necessarily shared by these people, however. My work in this area is supported in part by National Institutes of Health Grant GM45344.

1. Jeffreys, A. J., Wilson, V. & Thein, S. L. (1985) *Nature (London)* **314**, 67–73.
2. Jeffreys, A. J., Wilson, V. & Thein, S. L. (1985) *Nature (London)* **316**, 76–79.
3. Jeffreys, A. J., Brookfield, J. F. Y. & Semeonoff, R. (1986) *Nature (London)* **317**, 818–819.
4. Lander, E. S. (1989) *Nature (London)* **339**, 501–505.
5. Lewontin, R. C. & Hartl, D. L. (1991) *Science* **254**, 1745–1750.
6. Chakraborty, R. & Kidd, K. K. (1991) *Science* **254**, 1735–1739.
7. Morton, N. E. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 2556–2560.
8. Neufeld, P. J. & Colman, N. (1990) *Sci. Am.* **262** (5), 46–53.
9. Thompson, W. C. & Ford, S. (1991) in *Forensic DNA Technology*, eds. Farley, M. A. & Harrington, J. J. (Lewis, Chelsea, MI), pp. 93–152.
10. Promega Corporation (1991) *Proceedings of the Second International Symposium on Human Identification* (Promega, Madison, WI).
11. Kirby, L. T. (1990) *DNA Fingerprinting* (Stockton, New York).
12. Farley, M. A. & Harrington, J. J., eds. (1991) *Forensic DNA Technology* (Lewis, Chelsea, MI).
13. Burke, T., Dolf, G., Jeffreys, A. J. & Wolff, R., eds. (1991) *DNA Fingerprinting: Approaches and Applications* (Birkäuser, Basel).
14. National Research Council (1992) *DNA Technology in Forensic Science* (Natl. Acad. Press, Washington).
15. Andrews v. State 533 So. 2d 841 (Fla. App. 5 Dist. 1988).
16. Berry, D. A., Evett, I. W. & Pinchin, R. (1992) *Appl. Stat.* **41**, 499–531.
17. Geisser, S. (1992) *J. Am. Stat. Assoc.* **87**, 607–614.
18. Geisser, S. (1990) *Chance* **3**, 8–9.
19. Chakraborty, R. (1992) *Hum. Biol.* **64**, 141–159.
20. Chakraborty, R., de Andrade, M., Daiger, S. P. & Budowle, B. (1992) *Ann. Hum. Genet.* **56**, 45–57.
21. Chakraborty, R. & Jin, L. (1992) *Hum. Genet.* **88**, 267–272.
22. Deka, R., Chakraborty, R. & Ferrell, R. E. (1991) *Genomics* **11**, 83–92.
23. Devlin, B. & Risch, N. (1992) *Am. J. Hum. Genet.* **51**, 534–548.
24. Devlin, B. & Risch, N. (1992) *Am. J. Hum. Genet.* **51**, 549–553.
25. Devlin, B., Risch, N. & Roeder, K. (1990) *Science* **249**, 1416–1420.
26. Devlin, B., Risch, N. & Roeder, K. (1991) *Am. J. Hum. Genet.* **48**, 662–676.
27. Devlin, B., Risch, N. & Roeder, K. (1991) *Science* **253**, 1039–1041.
28. Devlin, B., Risch, N. & Roeder, K. (1992) *J. Am. Stat. Assoc.* **87**, 337–350.
29. Risch, N. J. & Devlin, B. (1992) *Science* **255**, 717–720.
30. Weir, B. S. (1992) *Genetics* **130**, 873–887.
31. Weir, B. S. (1992) *Am. J. Hum. Genet.* **51**, 992–997.
32. People v. Barney 8 Cal. App. 4th 798 (1992).
33. Roberts, L. (1991) *Science* **254**, 1721–1723.
34. Thompson, W. C. (1989) *Law Contemp. Probl.* **52**, 9–41.
35. Evett, I. W. (1991) *Nature (London)* **354**, 114.
36. Nelson, L. S. (1978) *J. Qual. Technol.* **10**, 42–43.
37. Odelberg, S. J., Plaetke, R., Eldridge, J. R., Ballard, L., O'Connell, P., Nakamura, Y., Leppert, M., Lalouel, J.-M. & White, R. (1989) *Genomics* **5**, 915–924.
38. Cohen, J. E., Lynch, M. & Taylor, C. E. (1991) *Science* **253**, 1037–1038.
39. Guo, S. W. & Thompson, E. A. (1992) *Biometrics* **48**, 361–372.

40. Efron, B. & Tibshirani, R. (1991) *Science* **253**, 390–395.
41. Brookfield, J. (1992) *Nature (London)* **356**, 483.
42. Lewontin, R. C. (1993) *Am. J. Hum. Genet.* **52**, in press.
43. Weir, B. S. & Evett, I. W. (1992) *Am. J. Hum. Genet.* **50**, 869.
44. Wooley, J. R. (1992) *Proceedings of the Third International Symposium on Human Identification* (Promega, Madison, WI), in press.
45. Evett, I. W. & Weir, B. S. (1992) *Chance* **4**, 19–21.
46. Cohen, J. E. (1990) *Am. J. Hum. Genet.* **46**, 358–368.
47. Brookfield, J. (1992) *Heredity* **69**, 97–100.
48. Wambaugh, J. (1989) *The Blooding* (William Morrow, New York).
49. Walsh, K. A. J. & Buckleton, J. S. (1991) *J. Forensic Sci. Soc.* **31**, 49–58.
50. Cohen, J. E. (1992) *Am. J. Hum. Genet.* **51**, 1165–1167.
51. Curnow, R. N. (1992) *Appl. Stat.* **41**, 499–531.
52. Berry, D. A. (1991) *Stat. Sci.* **6**, 175–205.
53. Evett, I. W. & Werrett, D. J. (1990) in *Proceedings of the International Symposium on Human Identification* (Promega, Madison, WI), pp. 77–91.
54. Lewontin, R. C. & Hartl, D. L. (1992) *Science* **255**, 1054–1055.
55. Budowle, B., Giusti, A. M., Waye, J. S., Baechtel, F. S., Fourney, R. M., Adams, D. E., Presley, L. A., Deadman, H. A. & Monson, K. L. (1991) *Am. J. Hum. Genet.* **48**, 841–855.
56. Hagerman, P. J. (1990) *Am. J. Hum. Genet.* **47**, 876–877.
57. Lempert, R. (1991) *Cardozo Law Rev.* **13**, 303–341.
58. Higuchi, R. (1991) *Am. J. Hum. Genet.* **48**, 1215–1216.
59. Anderson, A. (1990) *Nature (London)* **346**, 499.
60. Anderson, C. (1992) *Nature (London)* **355**, 663.
61. Shapiro, M. M. (1991) *Nature (London)* **353**, 121–122.
62. Green, P. (1992) *Am. J. Hum. Genet.* **50**, 440–441.
63. Helmuth, R., Fildes, N., Blake, E., Luce, M. C., Chimera, J., Madej, R., Gorodezky, C., Stoneking, M., Schmill, N., Klitz, W. & Erlich, H. A. (1990) *Am. J. Hum. Genet.* **47**, 515–523.
64. Budowle, B., Chakraborty, R., Giusti, A. M., Eisenberg, A. J. & Allen, R. C. (1991) *Am. J. Hum. Genet.* **48**, 137–144.
65. Jeffreys, A. J., MacLeod, A., Tamaki, K., Neil, D. L. & Monckton, D. G. (1991) *Nature (London)* **354**, 204–209.
66. Weir, B. S. (1993) *Am. J. Hum. Genet.* **52**, in press.