Statistical Evaluation of DNA Fingerprinting: A Critique of the NRC's Report

B. Devlin, Neil Risch, Kathryn Roeder

Since the National Research Council (NRC) report on DNA fingerprinting (*1*) was released (May 1992), it has been used by the California Appellate Court and the Massachusetts Supreme Court (*2*) to rule that DNA fingerprinting evidence (*3*) should not have been admitted in previously tried cases. Courts base the decision of admissibility of a scientific method on whether it is generally accepted in the relevant scientific community. In the above-mentioned cases, the courts noted that the NRC report and a related article by Lewontin and Hartl (*4*) were sufficient to demonstrate a lack of consensus regarding estimates of genotype frequencies in U.S. populations.

Although we understand why the courts were impressed by the claims of population geneticists, namely Lewontin and Hartl, and an NRC panel, we argue that these opinions are only a minority view and that there is indeed a consensus (*5*) supporting the reliability of estimates of genotype probability. Forensic DNA testing has been adopted not only throughout the United States but in Canada, Europe, and elsewhere. Similarly, paternity testing, which uses identical methodology, has been accepted for years. Although a few population geneticists have argued for extremely conservative interpretations of forensic DNA data on the basis of suspicions that standard assumptions used in human population genetics are strongly violated, other population geneticists, who have analyzed the genetic data [variations based on numbers of tandemly repeating (VNTR) sequences] and explored the problem theoretically, argue that there is no evidence that these assumptions are strongly violated and there is little reason to expect they would be.

We disagree, as do many others (*6–14*), with many statements in the NRC report regarding human population genetics and statistical inference. Moreover, we do not believe that the proposed research will resolve the population genetics debate.

B. Devlin is in the Department of Epidemiology and Public Health at Yale University School of Medicine, New Haven, CT 06510. N. Risch is in the Departments of Epidemiology and Public Health and Genetics at Yale University School of Medicine, New Haven, CT 06510. K. Roeder is in the Department of Statistics at Yale University, New Haven, CT 06520.

The Arguments

The population genetics debate (*4, 15* versus *9, 16–19*) has focused on the assumptions of independence of alleles within a locus (Hardy-Weinberg equilibrium) and at different loci (linkage equilibrium). Those who argue against independence claim that dependency is a result of population heterogeneity. By heterogeneity we mean that each population or ethnic group is composed of subpopulations having different allele frequencies; such heterogeneity would formally violate Hardy-Weinberg and linkage equilibrium. Those who argue that independence assumptions yield reasonable approximations to genotype probabilities do so for the following reasons: (i) The subpopulations that originally constituted U.S. populations (for example, the English, French, Germans, Irish, Italians, and other subpopulations that constitute the Caucasian population) are quite similar in their allele distributions for traditional genetic markers, and the VNTR loci are unlikely to be substantially different. Thus, the effects of heterogeneity are small, even for founding generations. (ii) Intermarriage has been common throughout the history of the United States (*20*), reducing the effects of heterogeneity. (iii) Results of tests of allele independence for VNTR loci and results based on traditional loci agree in that violations of assumptions are minimal or lacking (*16, 17, 21, 22*); hence, heterogeneity cannot have a substantial effect.

The authors of the NRC report, while noting these published analyses, base their recommendations on the position of Lewontin and Hartl (*4*), who argued that subpopulations of an ethnic group differ more from each other than do the major ethnic groups (races). In addition, the authors of the report make two related statements: "differences among subpopulations of an ethnic group cannot be determined by comparing [ethnic groups]" (*1*, section 3, p. 10); and "the validity of the multiplication rule depends on the absence of population substructure, because only in this special case are the different alleles statistically uncorrelated with one another" (*1*, section 3, p. 6).

Human Population Genetics

In 1972, Lewontin (*23*) stated that subpopulations of an ethnic group differ more from each other than do the major ethnic groups. For the genetic markers and populations he studied, most of the gene diversity was found among individuals within populations (85.4%). Differences among subpopulations of the same ethnic group accounted for 8.3% of the diversity, and the remaining 6.3% was attributed to differences among ethnic groups. Lewontin's research (*23*) on human population structure is the only study cited by the NRC report that makes use of traditional genetic markers. This is odd because, in the 20 years since 1972, standard methods have been defined, and a large body of consistent results has been generated for migration, genealogy, isonomy, traditional genetic markers, and DNA fingerprinting (*9*). These results and other recent investigations of human population genetics have led to a conclusion very different from Lewontin's concerning the apportionment of genetic diversity (*24, 25*; see also *26*).

For instance, when Smouse and colleagues (*25*) studied the genetic structure of Amerindian populations by means of a multivariate (multilocus) analysis, they found the average diversity of ethnic groups to be nearly twice that of tribes within an ethnic group. Even a study (*27*) that made use of Lewontin's own methods and similar populations and loci, but with larger samples, failed to replicate his finding that gene diversity of ethnic groups was less than that of subpopulations.

When applied to the gene diversity of forensically important U.S. populations (for example, Caucasians or African Americans), the argument for extensive subpopulation diversity is even less plausible. Nei and Roychoudhury (*28*) showed that the amount of Caucasian subpopulation diversity was much less than that among ethnic groups and that both of these sources of diversity were far smaller than individual diversity. Similar results were found for relevant African and Mongoloid subpopulations (*29*).

U.S. populations, because of extensive interbreeding among individuals of different subpopulations and ethnic groups, are much less heterogeneous than the European and African subpopulations from which they are derived. Intermarriage, while homogenizing populations, increases diversity among individuals. Consequently, contrary to the NRC panel's assertion, U.S. populations fit the forensic paradigm of reference populations rather well: (i) most genetic diversity is among individuals; (ii) as the number of markers that comprise the multilocus genotype increases, the probability

of randomly choosing two individuals with identical genotypes becomes remote; and (iii) genetic diversity among subpopulations of an ethnic group is less extreme than the differences among ethnic groups. The VNTR markers are not exceptions to these observations (*11, 12, 19, 21, 22*).

Therefore, contrary to the NRC panel's assertion, differences between the ethnic groups can tell us much about the potential differences among their subpopulations, arguably providing an upper bound. They should also provide substantial information concerning the potential errors incurred by forensic calculations, which are based on databases comprising a mixture of subpopulations (with substantial intermarriage disrupting this structure). For instance, although African Americans and Caucasians exhibit statistically significant differentiation at all VNTR loci examined to date, if these two populations are mixed into a single reference population, we observe only relatively small differences between the true genotype probabilities (allowing for mixture) and the probabilities estimated with the assumptions of Hardy-Weinberg and linkage equilibrium (*13, 19*). Hence, even if ethnic databases are composed of a mixture of subpopulations, the effects on estimates of genotype probability will be small. Such results also contradict the panel's argument that "the validity of the multiplication rule depends on the absence of population substructure" (*1*, section 3, p. 6). Evidently, even when there is substantial substructure, the multiplication rule still yields adequate approximations (*13, 19, 30, 31*).

The Ceiling Principle

On the basis of their assumptions regarding U.S. populations, the panel proposed a method to account for population substructure. Their suggested method, which they called the ceiling principle, is to study 100 individuals from each of 15 to 20 populations, such as English, Russians, Navajos, Puerto Ricans, and West Africans. Allele frequencies would be estimated from these populations and then, for any particular genotype, the maximum allele frequency found among the populations would be chosen. In other words, a forensic genotype "probability" could be the product of a Navajo allele probability, a few Russian allele probabilities, a West African allele probability, and so on. They also add the condition that no allele probability should be below 10% (or possibly 5%) because, although an allele's probability may be rare in all populations sampled, larger probabilities could occur in other unsampled populations.

Until the 15 to 20 populations have been studied, the panel proposes the use of the maximum allele probability that occurs in the existing databases for three or more major ethnic groups. They also propose that the 10% rule should be strictly applied. Furthermore, rather than use the maximum allele probability per se, the upper 95% confidence bound should be used (*33*). Once the population studies have been performed, the panel states, "Assuming that the population studies do not show significant substructure, the 5% lower bound recommended earlier should be used" (*1*, section 3, p. 22).

It is difficult, however, to justify the arbitrary values of 10 and 5%, especially considering that some alleles have been found to be rare in a wide range of populations (*12, 34*). It is also difficult to justify choosing the maximum allele frequency over all of the populations. For example, the fact that certain alleles occur more frequently in the Navajos than they do in European and West African populations is of little consequence for a crime committed in Boston, where few Navajos reside. Others have also criticized the panel's method of calculating genotype probabilities as lacking scientific justification (*8, 9, 11–14*).

The stated objective of the panel's recommended study concerning population genetics is to examine the amount of subpopulation differentiation; however, we believe there are serious flaws in the study design. The panel says little about assessing population differentiation beyond the note, quoted above, stating that if the populations do not show significant substructure, the 10% bound can be lowered to 5%. Taking this at face value, there is little need to analyze new data: existing data are sufficient to demonstrate differences among populations. Hence, the bound will not be lowered.

Thus, only two questions relevant to population genetics remain: are allele distributions of subpopulations very different, and how much of the genetic diversity is attributable to ethnic groups, to subpopulations, and to individuals within subpopulations? The proposed experimental design would not yield satisfactory answers to these questions. Too many ethnic groups would be sampled at the expense of subpopulations, and too few individuals would be sampled within subpopulations. The result would be a large sampling variance that would exaggerate variation among subpopulations (*6*).

The critical flaw in the study design is the number of individuals to be sampled in each subpopulation. For the most conservative definition of alleles, the fixed-bin method (*32*), the number of alleles per locus is frequently greater than 20, and those alleles are relatively equiprobable. Even for this distribution, a sample of 100 individuals is too small. Furthermore, underlying those fixed-bin alleles are actually hundreds of alleles defined by size (*19*) and potentially thousands of alleles defined by base pair composition (*35*). Therefore, any sensible partition of the genetic variance (*36*) will show that most variance is attributable to sampling error. In addition, it will be impossible to partition variances accurately.

The other goal of the proposed study is to obtain maximum allele frequencies. The impact of sampling error on these estimates may also be tremendous. In general, it will exaggerate maximum allele frequencies (*6*).

Interpreting the Evidence

Although it is common practice to present genotype probabilities in court, we argue that it would benefit the court to understand the statistical interpretation of the evidence. Such an interpretation is readily available by means of the likelihood ratio (LR) (*37–39*), although the panel does not recommend its use (*40*). The ratio of likelihoods is calculated on the basis of two competing hypotheses: the numerator is the likelihood of observing the VNTR patterns of the two forensic samples given that they are from the same person, and the denominator is the likelihood of observing the VNTR patterns of the two forensic samples assuming that they are from two different people. When a match is declared and the evidence consists of discrete or binned alleles, the LR is simply the inverse of the genotype probability. Thus, if the genotype probability is 0.0001, the LR suggests that the evidence is 10,000 times more likely to be observed when both samples are from the same person than when the samples are from two unrelated individuals (*41*).

This example underscores the advantage of the LR, namely, its interpretability. Presenting the jury with the probability of 0.0001 frequently leads to confusion. For instance, the jurors may grapple with the question of whether the probability is interpretable without knowledge of the size of the reference population. They can also be confused by the fallacious argument that a probability cannot be as small as 10^{-12} because the U.S. population is only 2.6×10^8. In fact, as the number of VNTR loci that match increases, the probability goes to zero and LR goes to infinity, as they should because the chance that the samples are not from the same person, or a twin with an identical VNTR genotype, is essentially zero.

(*Continued on page 837*)

(*Continued from page 749*)

Conclusions

Surprisingly, the NRC panel did not investigate what is currently known about the genetic structure of U.S. populations. They do not appear to have realized that their fundamental motivational assumption was incorrect. As a result of this error, there is little scientific basis for their method of forensic inference—the ceiling principle. We agree, however, that it is extremely conservative. Fortunately, the methods used in court are already conservative. Compared with a statistically based LR method, standard methods have usually been more conservative, frequently orders of magnitude more conservative (*37*). Of course, the appropriate degree of conservativeness remains the venue of legal scholars, not population geneticists or statisticians.

The NRC panel's proposed study of population substructure will be counterproductive. Sampling error would be so large, the study would exaggerate the differences among subpopulations of an ethnic group and distort maximum allele frequencies. This research, if brought to fruition, is likely to fuel the debate over the validity of the assumptions of population genetics rather than resolve it with useful data.

In conclusion, we have serious concerns that the erroneous assumptions and conclusions in the NRC report are receiving undue weight in judiciary decisions. It would be unfortunate if these errors were to influence decisions of the admissibility of a very powerful forensic tool.

REFERENCES AND NOTES

1 Committee on DNA Technology in Forensic Inference, *DNA Technology in Forensic Science* (National Academy Press, Washington, DC, 1992).

2 1992 WL 184530 (CA App. 1 Dist.); 1992 WL 171780 (MA).

3. Some people have argued that the term DNA fingerprinting is misleading because, unlike dermal fingerprints, DNA fingerprints are not unique To the contrary, it is the analogy that is misleading neither partial dermal nor partial DNA prints are unique The probability of a chance match between two unrelated individuals for five highly polymorphic VNTR loci is miniscule (*16*), and eight to ten such loci are apparently sufficient to ensure uniqueness of all but identical twins, some full siblings, and the most highly inbred family members [for example, the Karitiana tribe; see N Risch and B Devlin, *Science* **256**, 1744 (1992)].

4. R C Lewontin and D. L. Hartl, *ibid.* **254**, 1745 (1991).

5. The judgment arrived at by most of those concerned [*Webster's Ninth New Collegiate Dictionary* (Merriam-Webster, Springfield, MA, 1983)].

6. B Devlin, N Risch, K Roeder, in preparation

7. B. Weir, *Am. J Hum. Genet.*, in press

8. _____, *Proc. Natl Acad. Sci. U.S.A.* **89**, 11654 (1992).

9. N. E. Morton, *ibid*, p. 2556, *Eur. J. Hum Genet.*, in press.

10. B. S. Weir and I. W. Evett, *Am. J. Hum. Genet*, in press.

11. B. Budowle and K. L. Monson, in preparation

12. I. Balazs, in *Second International Conference on DNA Fingerprinting*, S. D. J. Pena *et al.*, Eds. (Birkhauser Verlag, Basel, Switzerland, in press).

13 I. W. Evett, J. Scrange, R Pinchin, *Am. J. Hum. Genet.*, in press.

14. B. Budowle and K. Monson, in *Proceedings of the 1992 Symposium on Human Identification*, in press; R Chakraborty *et al.*, in preparation; N Morton, A. Collins, I. Balazs, *Proc. Natl. Acad Sci. U S.A.*, in press

15. J. E. Cohen, *Am. J. Hum. Genet.* **46**, 358 (1990), *ibid.* **51**, 1165 (1992), _____, M. Lynch, C E. Taylor, *Science* **253**, 1037 (1991); E. Lander, *Nature* **339**, 501 (1989); *Am. J. Hum. Genet* **48**, 819 (1991); P. Green and E. S. Lander, *Science* **253**, 1038 (1991).

16. N. J. Risch and B. Devlin, *Science* **255**, 717 (1992).

17. R. Chakraborty and K. K. Kidd, *ibid.* **254**, 1735 (1991); _____, L Jin, *Hum. Genet* **88**, 267 (1992); B. Devlin, N Risch, K. Roeder, *Science* **249**, 1416 (1990); B. Devlin and N. Risch, *Am. J. Hum Genet.* **50**, 549 (1992).

18 B. Devlin, N Risch, K. Roeder, *Science* **253**, 1039 (1991)

19. B. Devlin and N. Risch, *Am. J. Hum. Genet.* **50**, 534 (1992).

20 R J. R. Kennedy, *Am. J. Sociol* **49**, 331 (1944), *ibid.* **58**, 56 (1952); J. N. Spuhler and P. J. Clark, *Hum. Biol.* **33**, 223 (1961)

21 B Weir, *Genetics* **130**, 873 (1992)

22 _____, *Am J. Hum. Genet.* **51**, 992 (1992).

23. R. C. Lewontin, *Evol. Biol.* **6**, 381 (1972).

24. J. B. Mitton, *Am. Nat.* **112**, 1142 (1978); J. V Neel, *ibid.* **117**, 83 (1981); M. Nei, *ibid.*, p. 88.

25 P E. Smouse, R. S. Speilman, M. H. Park, *ibid.* **119**, 445 (1982).

26 There are also statistical reasons to doubt Lewontin's original result because his method treats allele probabilities as known constants rather than estimates. For relatively small samples, his method will exaggerate differences among subpopulations of the same ethnic group more than it will the differences among the ethnic groups (*6*)

27 B. D. H. Latter, *Am. Nat.* **116**, 220 (1980)

28 M. Nei and A K Roychoudhury, *Evol. Biol* **14**, 1 (1982).

29. Almost identical results would have been obtained if only those populations relevant to the United States were considered—in other words, exclude such populations as African pygmies and the !Kung bushmen.

30. See also the research in (*11–13, 21, 22*) For instance, Weir (*21, 22*) showed that there was a strikingly high correlation of genotype probabilities across ethnic groups and even greater correlation within ethnic groups

31. See also the research by I. W. Evett and P. Gill [*Electrophoresis* **12**, 226 (1991)]. The small errors induced by subpopulation heterogeneity need not be ignored, however, because there are standard techniques of population genetics to adjust the values, making them even more conservative and applicable even if there are grounds to suppose a relation between suspect and culprit (*9*). In addition, we note that forensic scientists apply other conservative corrections to the data [(*32*); R. Chakraborty *et al.*, *Ann. Hum. Genet.* **56**, 45 (1992)].

32. B. Budowle *et al.*, *Am. J. Hum. Genet.* **48**, 841 (1991).

33. It is highly improbable that a large number of probability estimates all need to be adjusted upward to the bound of the 95% confidence interval. Statistical approaches based on results for the variance of products [L. A. Goodman, *J. Am. Stat. Assoc.* **55**, 708 (1960); *ibid.* **57**, 54 (1962)] or a resampling method [B. Efron and R. Tibshirani, *Science* **253**, 390 (1991)] are more appropriate. See Weir (*9*) for further discussion.

34. I. Balazs *et al.*, *Genetics* **131**, 191 (1992).

35. A J. Jeffries *et al.*, *Nature* **354**, 204 (1991).

36. C. C. Cockerham, *Evolution* **23**, 72 (1969); *Genetics* **76**, 679 (1972); B. S. Weir and C. C. Cockerham, *Evolution* **38**, 1358 (1984); J. C. Long, P. E. Smouse, J. W. Wood, *Genetics* **117**, 223 (1987): J C. Long, *ibid.* **112**, 629 (1986).

37. B Devlin, N. Risch, K. Roeder, *J. Am. Stat. Assoc.* **87**, 337 (1992); D. A. Berry, *Stat. Sci.* **6**, 175 (1991); _____, I. W. Evett, R. Pinchin, *Appl. Stat.* **41**, 499 (1992).

38. I. W. Evett and B. S. Weir, *Chance* **4**, 19 (1992); B. S. Weir and I. W. Evett, *Am. J. Hum. Genet.* **50**, 869 (1992).

39. The NRC report argues against LR-based approaches on the basis of their complicated nature, which would be difficult for jurors to understand. Although we agree with the NRC panel that LRs that account for measurement error (*37*) are complicated to calculate, we see no reason for jurors to be bothered by that complexity because they do not do the calculations. Jurors simply evaluate the results, which are no more complicated than the results with which they are presently provided. Moreover, such calculations and interpretations are not novel. The LR in this case is directly comparable to the paternity index, which is used frequently to evaluate civil cases involving disputed paternity [P. Smouse and R. Chakraborty, *Am. J. Hum. Genet.* **38**, 918 (1986); M. P. Baur *et al.*, *ibid.* **39**, 528 (1986)]

40 The LR interpretation is based on whether the same person or two different people were measured. Even when the odds are overwhelming, the interpretation need not imply guilt. it is conceivable that the defendant was framed or that the laboratory made a mistake by loading the same sample twice. The latter possibility is discussed elsewhere (*7*), in which it is argued that most discussions of laboratory error rates, including that by the NRC panel, are misleading at best. In our opinion, forensics should now progress from discussion of the NRC report to refinement and comparison of alternative LR calculations.

41. We thank I. Balazs, M. Baird, B Budowle, A. Guisti, R. Harmon, J. Hartmann, T. Holford, B. Lindsay, K. Monson, and B. Weir for comments on drafts of the manuscript. Supported by NIH grants HG00648 and CA45052 (to N.R.) and NSF grants DMS9201211 and DMS9257006 (to K.R.).