1993

The Suspect Population and DNA Identification

Richard O. Lempert
University of Michigan Law School, rlempert@umich.edu

THE SUSPECT POPULATION AND DNA IDENTIFICATION

Richard Lempert*

Forensic DNA analysis typically proceeds by first determining whether alleles[1] found in DNA apparently left by the perpetrator of a crime at a crime scene[2] (the "evidence sample") match alleles extracted from a sample of the suspected criminal's blood (the "suspect sample"). If alleles drawn from the two sources match, the next step is to provide information about the probative value of the match by estimating the probability that alleles extracted from the blood of some random individual would have matched the alleles in the evidence sample. This probability estimate is typically made by estimating the frequency with which specific alleles are found in some population and combining these frequencies according to the product rule.[3] Obviously, a match

*Richard Lempert is Francis A. Allen Professor of Law, University of Michigan Law School, and Professor and Acting Chair of the Department of Sociology at the University of Michigan. He was a member of the National Research Council's Committee on DNA Technology in Forensic Science.

[1] An allele is "one of two or more alternative forms of a gene." COMMITTEE ON DNA TECHNOLOGY IN FORENSIC SCIENCE, NATIONAL RESEARCH COUNCIL, DNA TECHNOLOGY IN FORENSIC SCIENCE 167 (1992) (hereinafter NRC REPORT). The "alleles" used in most forensic work, however, consist of DNA fragments that are not thought to include any genes. Indeed, the fact that these DNA fragments have no bearing on an individual's traits allows for much greater variability in the DNA. Thus, in the forensic context, the term "allele" merely refers to fragments of DNA that can be distinguished from one another in the laboratory.

[2] Less frequently the issue will be whether alleles extracted from organic material found on the suspect or his possessions match a victim's alleles.

[3] According to the product rule, the probability of two independent events equals the probability of the first event times the probability of the second; with *n* independent events the separate probabilities of each of the *n* events are multiplied together to give the probability of their joint occurrence. Thus if the probability that a person had allele A = 1/10 and the probability that he had allele B = 1/10 and the probability that he had allele C = 1/10, and if the probability

between a suspect's DNA and evidence DNA provides less reason to believe that the suspect is guilty if many people share the suspect's DNA profile than if few people do.

Although the validity of the techniques that forensic scientists use in matching suspect to evidence DNA is generally accepted,[4] the application of the product rule to determine the evidential weight of a DNA match has generated considerable dispute. In particular, analysts have argued that given population substructure, allele frequencies estimated on the basis of a laboratory's reference sample may understate the allele frequencies in the ethnic group to which the defendant belongs and configurations of alleles, as estimated by the product rule, may underestimate the probability that these configurations would be found in specific subpopulations.[5] Apparently for these reasons, forensic laboratories have estimated allele frequencies for Caucasian defendants by reference to the allele frequencies found in a Caucasian population data base, allele frequencies for blacks by reference to the allele frequencies found in a black population data base, and allele frequencies for Hispanics by reference to allele frequencies found in an Hispanic population data base.[6]

If population substructure is the reason why allele frequencies for defendants of different gross ethnicities are estimated on the basis of allele frequencies in same-ethnicity data bases, then the effort made is inadequate, for the racial categories used are so general that considerable population substructure continues to exist within racial categories. Population subgroups subsumed under these socially-defined ethnic categories may be characterized by very different allele frequencies for alleles of different lengths at the loci tested.

Yet the population substructure problem is often not serious. The fact that a suspect is a member of a highly inbred group with allele frequencies quite different from those found in a reference data base ordinarily does not matter. Indeed, no problem a t all will arise unless the "suspect population," which is to say the group of people who plausibly might be suspected of having

that the person had one of these alleles was not affected by whether or not he had either or both of the others, the probability that the person would have alleles A, B, and C would be 1/10 x 1/10 x 1/10, or 1/1000.

[4] A report by the National Research Council recommends that courts should judicially notice the validity, in principle, of the currently most common laboratory procedure for detecting DNA variation, single-locus probes analyzed on Southern blots. For a description of this and other techniques of DNA analysis, see NRC REPORT, *supra* note 1, at ch. 2.

[5] *Compare* Richard C. Lewontin & Daniel L. Hartl, *Population Genetics in Forensic DNA Typing*, 254 SCIENCE 1745 (1991) *and* Joel E. Cohen *et al.*, *Forensic DNA Tests and Hardy-Weinberg Equilibrium*, 253 SCIENCE 1037 (1991) (comment on Devlin et al., 249 SCIENCE 1416 (1990)) *with* Ranajit Chakraborty & Kenneth K. Kidd, *The Utility of DNA Typing in Forensic Work*, 254 SCIENCE 1735 (1991) *or* Neil J. Risch & B. Devlin, *On the Probability of Matching DNA Fingerprints*, 255 SCIENCE 717 (1992).

[6] Today laboratories commonly provide juries frequencies from different ethnic data bases rather than just the frequencies among a defendant's coethnics. This information is apparently provided as a step toward conforming with a recommendation in the NRC Report, *supra* note 1.

committed the crime, contains members of the same inbred group.[7] The point is easy to intuit. Assume that a rape has occurred in a particular town and that the potential suspects (the only people the local police are likely to investigate and arrest) all live within twenty miles of the town. If the person who is arrested is a member of an inbred Indian tribe, and is the only member of that tribe living within a thousand miles of the scene of the crime, allele frequencies within his tribe play no role in determining the probability that another person left the evidence DNA. Because (with some very high probability) the only persons who might have left the evidence DNA are members of the suspect population, allele frequencies within that group are what determine the possibility that some other plausible suspect might have left the evidence DNA.[8] If,

[7] When I originally drafted this Article, this point had not been made in the literature, except imprecisely in an unelaborated allusion by a prosecutor. James Wooley, *A Response to Lander: The Courtroom Perspective*, 49 AM. J. HUM. GENET. 892 (1991). Moreover, several important articles at that time implicitly assumed that the proper population database consisted of a sample of defendant's coethnics. Lewontin & Hartl, *supra* note 5; Cohen, *DNA Fingerprinting for Forensic Identification: Potential Effects on Data Interpretation of Subpopulation Heterogeneity and Band Number Variability*, 46 AM. J. HUM. GENET. 358 (1990); E.S. Lander, *Population Genetic Considerations in the Forensic Use of DNA Typing, in* BANBURY REP. 32: DNA TECHNOLOGY AND FORENSIC SCIENCE 1436 (J. Ballantyne et al. eds., 1989). Additionally, the standard procedure then was to base allele estimates for Caucasians on a Caucasian database, for blacks on a black database, etc.

Such is the march of science and the delays in finding a publisher that the point has now been recognized by a number of authors. Chakraborty & Kidd, *supra* note 5; Bruce S. Weir, *Population Genetics in the Forensic DNA Debate* (review), 89 PROC. NAT'L ACA D. SCI. USA 11,654 (1992); Richard C. Lewontin, *Which Population?* (letter to the editor), 52 AM. J. HUM. GENET. 205 (1993). The justifications for again making the point are that most work recognizing this issue has appeared in technical scientific journals that lawyers and judges seldom read; that the scientists who have made the point have not at the same time considered, as I do below, the implications of the presence of the defendant's relatives in the suspect population or focused specifically on the attributes of that population; and because, in the words of a reviewer for this journal with whom I hope the reader will agree, "the exposition here is concise and incisive, and some reiteration is probably useful."

[8] Jonathan Koehler, in a thoughtful paper, argues that the appropriate reference population is the "source population" consisting of those who might have been sources of the evidence DNA rather than the "suspect population." He points out correctly that not all potential sources need be suspects, offering the example of DNA extracted from hair found in a bed at a murder scene that could have come from the victim's husband who died a week before the victim was killed and so is not a suspect in the case. Where, however, specific plausible sources are not suspects, as in Koehler's hair example, the possibility that they left matching DNA should be explicitly investigated and eliminated; little is to be gained by including people like them in some larger population data base. Thus, if eyewitness evidence indicates that the suspect in the murder described above is white and the victim's husband is black, the jury should learn the probability that a match would be found in the DNA of a white person sampled at random *and*, assuming the evidence DNA matched the suspect's DNA, whether the husband's DNA also matched.

From another perspective almost anyone could be a source. The killer in our example may have flown to New York City from Oslo and flown back unsuspected once his crime was committed. But to include all possible sources in the source population would usually yield far less appropriate frequency estimates than those based on the population of plausible suspects.

Koehler also argues that it may have been impossible for a suspect to have left the evidence DNA, but he can offer only an implausible example (a woman suspect who plants a man's semen in a woman she murdered to frame the man for the crime). Moreover, where a suspect before DNA testing is known to have been incapable of being the source of evidence DNA yet remains in our example, the suspect population was a mixed Caucasian population, a mixed Caucasian data base would provide the appropriate estimate of the defendant's uniqueness within the group of plausible suspects, however common his allele configuration was among his tribal peers.

The situation is similar when some small proportion of the suspect population shares the defendant's particular ethnic heritage. Even if alleles that match the defendant's are substantially more common among his ethnic peers than within the suspect population, there is little reason for concern if allele frequencies have been conservatively estimated (e.g., by binning criteria) in the first instance. If the bulk of the suspect population were Caucasians of mixed ancestry, use of a mixed Caucasian data base to estimate allele frequencies would be appropriate. The probability that a Caucasian would match the defendant's allele configuration might be much lower than the probability that a member of the defendant's ethnic group would match. However, the probability that a random member of the suspect population would match the defendant's allele configuration would be much like that probability among Caucasians since Caucasians dominate the suspect population. Although the presence of some persons who shared the defendant's ethnic heritage would make the match probability based on a Caucasian data base an underestimate of the true probability, the magnitude of the underestimate should not be so great that it would not be offset by the use of conservatively estimated allele frequencies in the first instance.

As the proportion of the suspect population that belongs to the defendant's particular ethnic group increases, the attention that must be paid to allele frequencies within that group also increases. When, for example, the suspect population consists of 50% Caucasians and 50% fellow ethnics, the sample used to generate estimates of the probability that a randomly selected member of the suspect population might have left the evidence DNA should be composed of half Caucasians and half fellow ethnics if allele frequencies within these two groups are meaningfully different. This composition is indicated whether the defendant is an ethnic group member or a Caucasian.

This analysis does not mean, however, that the use of a black data base is, for example, appropriate if a white man is arrested for rape in a black ghetto. The suspect population consists of those who are plausible suspects given those factors that condition suspicion. If a rape victim claims her assailant was white, the police are not going to arrest a black-appearing man for the crime no matter how many black men would have been potential suspects

a suspect, the legal system has a good handle on the problem. For these reasons I will continue to speak of the "suspect" rather than the "source" population as the appropriate reference group for estimating allele frequencies. As a practical matter little is likely to turn on this distinction, for in most cases the group of plausible DNA sources and the group of plausible suspects will be identical or there will be known plausible sources (*e.g.*, rape victims' husbands) who should be specifically eliminated as sources of evidence DNA. *See* Jonathan Koehler, *Error and Exaggeration in the Presentation of DNA Evidence*, 34 JURIMETRICS J. 21 (1993).

absent information about the defendant's race. The suspect population will consist of white-appearing males, and the data base used to estimate the uniqueness of a defendant's allele configuration should reflect that fact.

A suspect population can consist largely of one particular ethnic group when, for example, either living patterns or some other information about a criminal limits potential suspects to those of a particular ethnicity. Thus allele frequencies found in a mixed Caucasian reference sample might misestimate the uniqueness (within the group of plausible suspects) of the DNA of a Caucasian defendant who is a member of an ethnically homogenous isolated community within which a rape occurred. Similarly, allele frequencies found within an American black reference sample might misestimate the uniqueness (within the group of plausible suspects) of the DNA of a West Indian black defendant arrested in a black ghetto, if the victim's description of her assailant meant that only men with West Indian accents were vulnerable to arrest. Whether such misestimates will in fact occur, and their likely magnitudes if they do occur, are empirical questions. Perhaps the degree of misestimation is unlikely to be great, and conservative estimates of allele frequencies in the first instance may provide defendants with sufficient protection against the possibility that the different composition of reference samples and suspect populations would lead to underestimated allele frequencies. In some actual cases, however, the precise identification of the suspect population and the use of allele frequencies found in it may be important.[9]

[9] *See, e.g.*, United States v. Two Bulls, 918 F.2d 56 (8th Cir. 1990), *reh'g granted, vacated,* 925 F.2d 1127 (8th Cir. 1991) (en banc); State v. Passino, No. 185-1-90 Fcr (Dist. Ct. Franklin County May 13, 1991). Each case involved a Native American defendant and a crime in which a substantial proportion of the suspect population shared the defendant's tribal heritage. Bruce Weir and Ian Evett, two of the forensic world's most perspicacious commentators on the statistics of DNA identification, wrongly criticize the judge in *Passino* for suggesting the FBI's data bases may not have been appropriate for finding the probability of a coincidental match. They write,

> Once a match has been declared between the DNA profiles of crime-scene material and a suspect, the FBI calculates the frequency of that profile in the population. This calculation is to provide an indication of how likely it would be to find that profile in a random member of the population. In *State v. Passino*, the court apparently was not told that these calculations are of interest only under the hypothesis that the crime-scene material came from someone other than the suspect. If the suspect did not provide the crime-scene material, then his ethnic background is quite irrelevant. (If he did provide the crime-scene material, then there is no need to consider random members of the population.)

Bruce S. Weir & Ian W. Evett, *Whose DNA?* (letter to the editor), 52 AM. J. HUM. GENET. 869 (1992).

Their mistake, which echoes that of Wooley, *supra* note 7, whom they cite, is to fail to recognize that the ethnic composition of members of the suspect population is potentially important if it is likely to differ substantially from that found in the FBI' s available data bases. *See* Lewontin, *supra* note 7. Moreover, the suspect's ethnic background while technically irrelevant if allele frequencies representative of the suspect population are used, matters in practice because the harm of misidentifying the suspect population or of lacking access to a data base that is representative of allele frequencies in the suspect population is potentially greatest when the defendant shares the ethnicity of the suspect population. Thus, in evaluating the probability th at DNA was left by

The situation is different, however, if people within the suspect population have allele configurations across the loci tested with a relatively high probability of matching that of the defendant. Often the suspect population will include such people; they are the close relatives of the defendant. Not only are such people likely to live in the same vicinity as the defendant, but they are also likely to share some of the characteristics (e.g., general appearance, accent, mannerisms) that made the defendant a suspect in the case. Although the probability may be quite low that the DNA of a randomly selected member of the suspect population would match the defendant's DNA at the tested loci, the probability will be substantially higher that at least one member of the suspect population has matching DNA, because the alleles of relatives are not randomly distributed with respect to those of the defendant. Where, for example, a DNA profile consists entirely of very rare alleles so that the probability of a random match across several loci approaches zero, the probability that two brothers will match at two loci is about 6% and at four loci about .3%.[10] In these circumstances, to present a legal factfinder with match probabilities based on the assumption that no systematic association is present between the defendant's DNA and the DNA of other members of the suspect population is misleading.

Thinking in terms of suspect populations reveals that often there is little reason to be concerned that a defendant's ethnic group is poorly represented in a laboratory population data base. However, when a defendant's close relatives are members of the suspect population, statistics based on applying the product rule to allele frequencies taken from a laboratory's population data base will be misleadingly low, and extremely so, if they purport to reflect the likelihood based on the DNA evidence alone that some person other than the defendant might have left the evidence DNA. The statistics will not be as misleadingly low if they purport to represent the likelihood that a person randomly selected from the suspect population would have tested alleles matching those of the defendant, because in a large population a relative of the defendant would be unlikely to be chosen by chance.

The random selection probability is not, however, the statistic that is

a white man arrested on an Indian reservation, the proper allele data base would be one representative of reservation Indians. But, if a Caucasian data base were used, the allele frequency estimates probably would be as or more favorable to the defendant than they would have been had a proper suspect population data base been used, since a person's alleles are likely to be more common among his coethnics than among members of some other ethnic group. Thus, a Caucasian defendant in this case is unlikely to be prejudiced by the use of a data base that does not represent the suspect population. If, however, the defendant were a Native American living on the reservation, he would probably be disadvantaged by the use of a Caucasian rather than a suspect population data base to estimate allele frequencies. Weir and Evett's mistake is explainable because they focused only on the trial judge's concern with the defendant's Native American heritage. By itself this is unimportant; it becomes important when it is considered in conjunction with the ethnic composition of the suspect population.

[10] *See* NRC REPORT, *supra* note 1, at 87.

central to the factfinder's decision making. The key issue for the factfinder concerns the chance that other potential suspects have matching alleles. With DNA evidence as the only evidence in the case, this is a function of the number of potential suspects and the probability that a member of the suspect group possesses matching DNA. Random probabilities of matching allele configurations are often so low that even though the group of potential suspects may be much larger than the group of close relatives, the probability that at least one person has matching DNA will be larger, and sometimes considerably so, for the group of relatives than for the group of unrelated suspects. In these circumstances relatives in the suspect population must be treated separately from unrelated individuals.[11]

Finally, at least until DNA data bases come on line, a substantial amount of non-DNA evidence ordinarily will link a defendant whose DNA is tested to the alleged crime. This evidence, which is unlikely to similarly implicate other members of the suspect population, is often essential for concluding that a defendant with DNA matching the evidence DNA is uniquely linked to the crime.[12] When the non-DNA evidence is strong, neither population substructures nor the presence of relatives is likely to create a large risk of injustice. But where the non-DNA evidence is weak or where that evidence would as strongly implicate related others, as an identification based on appearance and accent might, the cautions mentioned above are essential if the DNA evidence is to be given its proper weight and justice to be done.

[11] A similar argument can be made in principle where the suspect population includes a small proportion of people who share the defendant's ethnicity and a larger group who do not. However, because the probability of matching alleles will be much lower for an unrelated member of the defendant's ethnic group than for a relative of the defendant, if the number of coethnics in the suspect population is small compared to the number of other ethnics, the probability that at least one other potential suspect has DNA matching the defendant's may be larger in the group of other ethnics than it is in the group of co-ethnics.

[12] For those accustomed to thinking in Bayesian terms, the certainty with which such evidence suggests the defendant's guilt may be thought of as a Bayesian prior probability that the DNA evidence modifies.