Volume 7, Fall Issue, 1993

DNA EVIDENCE: PROBABILITY, POPULATION GENETICS, AND THE COURTS

*David H. Kaye**

INTRODUCTION

Courts, attorneys, scientists, statisticians, journalists, and government agencies have been explaining,[1] examining,[2] promoting,[3] proselytizing,[4] denigrating,[5] and otherwise struggling with DNA identification evidence at least since 1985.[6] In the first wave of cases, expert testimony for the

---

* Regents' Professor, Arizona State University College of Law, Box 877906, Tempe, AZ 85287-7906 (602 965-2922, K@ASU.EDU). A version of this paper was presented at the 1992 Joint Statistical Meetings of the American Statistical Association, the Biometric Society, and the Institute of Mathematical Statistics. I am grateful to Herman Chernoff for comments on that paper and to Colin Aitken, Richard Lempert, Bruce Weir, and especially Bernard Devlin for comments on later drafts. The errors that remain despite this guidance are entirely my own.

1. *See, e.g.*, David H. Kaye, *DNA Paternity Probabilities*, 24 FAM. L. Q. 279 (1990); K.F. Kelly et al., *Method and Applications of DNA Fingerprinting: A Guide for the Non-Scientist*, 1987 CRIM. L. REV. 105; Miller, *DNA Fingerprints to Aid Sleuths*, 128 SCI. NEWS 390 (1985).

2. *See, e.g.*, OFFICE OF TECHNOLOGY ASSESSMENT, GENETIC WITNESS: FORENSIC USES OF DNA TESTS (1990); Alan Guisti et al., *Application of Deoxyribonucleic Acid (DNA) Polymorphisms to the Analysis of DNA Recovered from Sperm*, 31 J. FORENSIC SCI. 409 (1986).

3. *See, e.g.*, Andre A. Moenssens, *DNA Evidence and Its Critics—How Valid Are the Challenges?*, 31 JURIMETRICS J. 87 (1990).

4. *See, e.g.*, People v. Wesley, 533 N.Y.S.2d 643, 644 (Sup. Ct. 1988) ("[t]he single greatest advance in the 'search for truth,' and the goal of convicting the guilty and acquitting the innocent, since the advent of cross-examination."), *aff'd*, 589 N.Y.S.2d 197 (App. Div. 1992).

5. *See, e.g.*, Gina Kolata, N.Y. TIMES, Jan. 29, 1990, at A1 ("Leading molecular biologists say a technique promoted by the nation's top law-enforcement agency for identifying suspects in criminal trials through the analysis of genetic material is too unreliable to be used in court."); Janet C. Hoeffel, Note, *The Dark Side of DNA Profiling: Unreliable Scientific Evidence Meets the Criminal Defendant*, 42 STAN. L. REV. 465 (1990); Marjorie M. Shultz, *Reasons for Doubt: Legal Issues in the Use of DNA Identification Evidence*, DNA ON TRIAL: GENETIC IDENTIFICATION AND CRIMINAL JUSTICE 19 (Paul R. Billings ed., 1992), *reviewed by*, John F.Y. Brookfield, *Gene Justice*, 363 NATURE 122 (1993) (dismissing Professor Shultz's analysis as "parochial nonsense"). Several of the biologists referred to in the New York Times story have complained that their views were misrepresented. Moenssens, *supra* note 3, at 99-100.

6. The earliest instance of DNA analysis for legal purposes is Alec J. Jeffreys et al., *Positive Identification of an Immigration Test-Case Using Human DNA Fingerprints*, 317 NATURE 818 (1985) (applying the multilocus probes described in Alec J. Jeffreys et al., *Individual-Specific "Fingerprints" of Human DNA*, 316 NATURE 76 (1985), and Alec J. Jeffreys et al., *Hypervariable "Minisatellite" Regions in Human DNA*, 314 NATURE 67 (1985)). Soon after, this group applied the technique to a serial murder case described at length in JOSEPH WAMBAUGH, THE BLOODING (1989), excluding one suspect and

prosecution was rarely countered, and courts readily admitted the findings of commercial laboratories.[7]

In the wake of this early enthusiasm for DNA evidence, doubts emerged.[8] Diligent attorneys and enterprising defendants enlisted well-credentialed experts to scrutinize the work of commercial and crime laboratories. The resulting plethora of questions about laboratory procedures and analyses[9] convinced many courts, including the Supreme Courts of Georgia,[10] Massachusetts,[11] and Minnesota[12] to exclude at least

---

incriminating another.

7. *See* David H. Kaye, *The Admissibility of DNA Testing*, 13 CARDOZO L. REV. 353, 357 n.17 (1991). A case that is representative of this epoch is Cobey v. State, 559 A.2d 391 (Md. Ct. Spec. App. 1988). A man forced a woman jogging in a park into the woods, where, as the court of appeals put it, he "ravished" her and drove away in her car. A policeman issued a traffic citation to Kenneth Cobey, who was driving that car. Cellmark Diagnostics performed a "DNA fingerprint analysis" showing "a 'match' between the DNA in Cobey's blood sample and the DNA [extracted from] semen stains [on the woman's clothing]." *Id.* at 392. The state produced five experts "who testified that DNA fingerprinting was accepted in the scientific community," while Cobey "produced no expert evidence to the contrary." *Id.* at 392. To buttress the testimony for the state, the court of appeals relied on a news account in the *American Bar Association Journal* that "Cellmark Diagnostics of Germantown, Md., claims its 'DNA fingerprint' test can identify a suspect with 'virtual certainty,' and that the chances of any two people having the same DNA fingerprint are one in 30 billion." *Id.* at 392 n.7 (quoting D. Moss, *DNA—The New Fingerprints*, 74 A.B.A. J. 66 (1988)). Although the court cautioned that "we are not, at this juncture, holding that DNA fingerprinting is now admissible willy-nilly," but "are merely holding that, based upon this record, [the court below] did not err . . . since there was no evidence to the contrary," *id.* at 398, courts—even those confronted with expert testimony opposing a DNA identification—frequently cite *Cobey* for the proposition that all aspects of DNA analysis and all types of DNA probes are accepted among scientists, even though this "30 billion" figure pertains to a multilocus probe that is no longer used in this country for criminal identification.

8. *See* Kaye, *supra* note 7, at 357 n.18.

9. These included the possible effects of contaminants on forensic samples, the use of ethidium bromide, corrections for band shifting, the records of laboratories on proficiency tests, the size of data bases used to assess the significance of matching bands, and the procedure for calculating the frequency of matching DNA patterns within the general population.

10. Caldwell v. State, 393 S.E.2d 436 (Ga. 1990) (finding Lifecodes's "straight binning method satisfactory," but because laboratory's calculation that frequency of profile in population was 1/24,000,000 rested on assumption of Hardy-Weinberg equilibrium inconsistent with its data base, the more conservative figure of 1/250,000 derived from that data base would have to be used).

11. Commonwealth v. Curnin, 565 N.E.2d 440 (Mass. 1991) (holding Cellmark's DNA evidence in rape case erroneously admitted in absence of showing general acceptance of validity of process leading to conclusion that one Caucasian in 59 million would have incriminating profile).

12. State v. Schwartz, 447 N.W.2d 422, 428 (Minn. 1989) (responding to Cellmark's multilocus VNTR probe, said to produce a "banding pattern [whose frequency] in the Causcasian population is approximately 1 in 33 billion," the court concluded that "DNA typing has gained general acceptance in the scientific community," but "the laboratory in this case did not comport" with "appropriate standards," and further holding the statistical

some aspects of DNA evidence.[13] Nevertheless, in the majority of cases, the courts continued to hold DNA matches and probabilities admissible even in the face of conflicting expert testimony.[14]

With the publication of a long-awaited report of a twelve-member panel of the National Research Council,[15] a third wave of cases is crashing down upon this battered legal shoreline. Even before the National Academy of Sciences released this report for publication, unofficial announcements of an impending call for a moratorium on forensic DNA identification[16] produced consternation[17] and legal maneuvering.[18] Although the final report sought no such moratorium and strongly endorsed the theory behind forensic DNA analysis, it does question several aspects of current and past practice and does recommend improvements in the process. The pressure created by these pronounce-

---

conclusion to be inadmissible, because even if the computation is accurate, "we remain convinced that juries in criminal cases may give undue weight and deference to presented statistical evidence").

13. Other courts have also refused to admit certain forms of DNA evidence. *See, e.g.*, United States v. Two Bulls, 918 F.2d 56 (8th Cir. 1990), *vacated for reh'g en banc but appeal dismissed due to death of defendant*, 925 F.2d 1127 (8th Cir. 1991); People v. Castro, 545 N.Y.S.2d 985 (Sup. Ct. 1989); *cf.* Perry v. State, 586 So.2d 242 (Ala. 1991) (remanding for hearing on Lifecodes's adherence to proper procedures and acceptability of statistical methods).

14. *See, e.g.*, United States v. Jakobetz, 747 F. Supp. 250 (D. Vt. 1990) (applying relevance standard), *aff'd*, 955 F.2d 786 (2d Cir. 1992), *cert. denied*, 113 S.Ct. 104 (1992); United States v. Yee, 134 F.R.D. 161 (N.D. Ohio 1991) (applying general acceptance standard); *cf.* State v. Pierce, 597 N.E.2d 107 (Ohio 1992) (applying relevance standard, no defense experts); Satcher v. Commonwealth, 421 S.E.2d 821 (Va. 1992) (applying general acceptance standard and statute, no defense experts).

15. COMMITTEE ON DNA TECHNOLOGY IN FORENSIC SCIENCE, NATIONAL RESEARCH COUNCIL, DNA TECHNOLOGY IN FORENSIC SCIENCE (1992) [hereinafter NRC REPORT]. For a more comprehensive summary of the NRC Report and thoughts on its legal implications, see Kenneth R. Kreiling, Review-Comment, 33 JURIMETRICS J. 449 (1993).

16. Gina Kolata, *U.S. Panel Seeking Restriction on Use of DNA in Courts*, N.Y. TIMES, Apr. 14, 1992, at A1.

17. The panel's chair promptly repudiated the N.Y. Times concededly exaggerated account. Gina Kolata, *Chief Says Panel Backs Courts' Use of a Genetic Test*, N.Y. TIMES, Apr. 15, 1992, at A1.

18. FBI "interference" in the preparation of the report and a last-minute compromise in a crucial section of the report has been alleged. Leslie Roberts, *DNA Fingerprinting: Academy Reports*, 256 SCIENCE 300 (1992) (describing compromise within the National Academy of Sciences Commitee on a Statistical Standard); Rorie Sherman, *Genetic Testing Criticized*, NAT'L L.J., Apr. 20, 1992 (some courts have ordered production of the penultimate draft of the NRC report, which was leaked to and criticized by the FBI). Charges of FBI interference apparently come readily to the lips of some participants in the public debate. *See e.g.*, Rorie Sherman, *New Scrutiny for DNA Testing*, NAT'L L.J., Oct. 18, 1993, at 3 (quoting one defense attorney's reaction to a recent decision of the National Academy to impanel a new committee to update the population genetics chapter of the 1992 report as the "offensive" result of the "law enforcement [community's] dictating to the independent scientific community how they should examine problems").

ments[19] is shaping opinions across the nation.[20]

Of all the technological and scientific issues in this debate, the most difficult for the courts, and those that have generated the most disagreement within the scientific community, involve statistics. The disagreements revolve around one central challenge—presenting the degree of similarity between DNA in a crime sample and DNA in a defendant's sample so that a judge or jury can fairly assess the probative value of DNA evidence. The predominant procedure for criminal DNA testing in the United States involves two major steps: first, declaring a "match" between the two samples, and second, if a match is declared, estimating its relative frequency in a reference population. This frequency indicates, at least indirectly, the significance of a match. It reveals whether the match is as common as a polite smile or as rare as the enigmatic expression of the Mona Lisa.

In determining the admissibility of testimony on these points, courts have applied two competing standards. One is the general acceptance standard first applied to scientific evidence in *Frye v. United States*.[21] Under the *Frye* standard, courts do not inquire directly into scientific truth, but ascertain whether the scientific community has reached the consensus that the scientific procedure in question rests on a valid theory and generates reliable results when properly applied.[22] The other

---

19. A committee of defense lawyers is reviewing convictions involving DNA evidence, seeking to apply the report's recommendations retroactively, as it were. *See* Tim Beardley, *DNA Fingerprinting Reconsidered Again*, SCI. AM., July 1992, at 26. Prosecutors have begun to request calculations of the frequency of matching DNA types using the "ceiling method" advocated in the report. *See* Christopher Anderson, *Courts Reject DNA Fingerprinting, Citing Controversy After NAS Report*, 359 NATURE 349 (1992).

20. *See* People v. Barney, 10 Cal. Rptr. 2d 731 (Ct. App. 1992) (finding that product rule calculation method not prescribed by NRC panel for calculating frequency of DNA pattern is not generally accepted among population geneticists), *followed in* People v. Wallace, 17 Cal. Rptr. 2d 721 (Ct. App. 1993); Commonwealth v. Lanigan, 596 N.E.2d 311 (Mass. 1992) (same); State v. Vandebogart, 616 A.2d 483 (N.H. 1992) (same); State v. Cauthron, 846 P.2d 502 (Wash. 1993) (finding error in allowing expert to testify that defendant was the source of the incriminating DNA and yet excluding testimony of frequency of the DNA pattern given that the NRC panel had proposed a generally accepted method of calculation); *cf.* State v. Bible, 858 P.2d. 1152 (Ariz. 1993) (holding method as applied to 1988 data base not generally accepted); *Springfield v. State*, 860 P.2d 435 (Wyo. 1993) (holding frequency re-calculated with "the most conservative" NRC method admissible under relevance standard); People v. Atoigue, DCA No. CR 91-95A (Guam Dist. Ct. App. Div. 1992) (method not generally accepted among population geneticists). For discussion of these cases, *see infra* Part II(B)(4).

21. 293 F. 1013 (D.C. Cir. 1923) (holding inadmissible expert opinion of truthfulness formed from a primitive version of the polygraph).

22. *See generally, e.g.*, 1 MCCORMICK ON EVIDENCE § 203 (J. Strong ed., 4th ed.

approach treats general acceptance as but one factor that bears on the ultimate question of whether the scientific findings are sufficiently reliable to justify their admission in view of the dangers of uncritical acceptance by the jury and undue expense and consumption of time.[23] Although both standards are consistent with the wording and history of Rules 403[24] and 702[25] of the Federal and the Uniform Rules of Evidence, in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,[26] the Supreme Court unanimously held that the federal rules implicitly reject the *Frye* test.[27] Over some dissent,[28] the Court attempted to define "scientific knowledge,"[29] and it articulated four "general observations"[30] for use in determining whether

---

1992).

23. *See, E.g.*, State v. Pierce, 597 N.E.2d 107 (Ohio 1992) (applying relevance standard to uphold admission of DNA statistics despite NRC Report); cases cited, MCCORMICK, *supra* note 22, § 203 at 872 n. 31.

24. Rule 403 states: "Although relevant, evidence may be excluded if its probative value is substantially outweighed by the danger of unfair prejudice, confusion of the issues, or misleading the jury, or by considerations of undue delay, waste of time, or needless presentation of cumulative evidence."

25. Rule 702 provides: "If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise."

26. 113 S.Ct. 2786 (1993).

27. Although *Daubert* should accelerate the movement away from *Frye*, two factors may blunt the force of the decision. First, the Court continued to apply its wooden, "plain meaning" construction of the rules. *See, e.g.*, Glen Weissenberger, *The Supreme Court and the Interpretation of the Federal Rules of Evidence*, 53 OHIO ST. L.J. 1307 (1992). Second, *Daubert* concerned the general acceptance of a scientific conclusion about a putative teratogen. The general methodology for determining teratogenicity—the examination of data from toxiciologic and epidemiologic studies—was not controversial; only its application was in dispute, and the application of an accepted methodology plays no part in the normal *Frye* analysis. For cases hesitating or declining to follow *Daubert*, see, for example, State v. Bible, 858 P.2d 1152 (Ariz. 1993) and Fishback v. People, 851 P.2d 884 (Colo. 1993).

28. Chief Justice Rehnquist and Justice Stevens dissented from this portion of the opinion.

29. The Court treated Rule 702 as the "primary locus" of the trial court's obligation to screen out unacceptable scientific testimony. 113 S.Ct. at 2795. The rule speaks of "scientific . . . knowledge," and the Court propounded the tautology that "in order to qualify as 'scientific knowledge,' an inference or assertion must be derived by the scientific method." *Id.* "Proposed testimony must be supported by appropriate validation—i.e., 'good grounds,' based on what is known," recognizing, of course, that "it would be unreasonable to conclude that the subject of scientific testimony must be 'known' to a certainty." *Id.* However, it is not the inference or assertion itself that must be sufficiently "known." "The focus, of course, must be solely on the principles and methodology, not on the conclusions that they generate." *Id.* at 2797. This article contends that "good grounds" exist "based on what is known" to support the introduction of DNA evidence and a variety of statistics or probabilities that indicate how revealing such evidence is.

30. First, citing the positivist criterion that, in principle, a scientific hypothesis must

purportedly scientific testimony possesses sufficient validity and reliability to qualify as "scientific knowledge" and whether it would sufficiently "assist the trier of fact" within the meaning of Rule 702.[31]

To help meet the challenge of presenting properly performed DNA tests within this legal framework, this Article outlines the statistical procedures that have been employed or proposed to provide judges and juries with quantitative measures of such probative value, describes more fully how the courts have dealt with these procedures, and evaluates the opinions and the statistical analyses from the standpoint of the law of evidence. Part I outlines the procedure used to declare whether two samples of DNA "match." It explains how shrinking the size of the "match window," as some defendants have urged, will decrease the risk of false matches, but will also exclude highly probative evidence of identity. This section also demonstrates that a defendant's effort to show that a smaller match window would not permit the declaration of a match is irrelevant or misleading. Part II explains procedures for estimating the frequency of the incriminating genetic characteristics in various populations. These procedures have been the subject of an acrimonious debate, both in the courts and in the press, about the effect of "population structure." This section reveals that the population structure objection, which has proved so effective in court, applies most strongly to only a limited class of cases Thus, courts have erred in excluding DNA evidence on the theory that the scientific community advocates that the most "conservative" procedures must be used in all cases. Part III identifies more fundamental problems in the use of population frequency estimates. It advocates supplementary and alternative procedures that are essential if quantitative statements of the probative value of DNA

---

be subject to some empirical test that could falsify it, the Court observed that "a key question . . . will be whether [a theory or technique] can be (and has been) tested." 113 S. Ct. at 2797. Second, a "pertinent consideration is whether the theory or technique has been subjected to peer review and publication." *Id.* Third, ordinarily "the known or potential rate of error" should be assessed. *Id.* Finally, "general acceptance" within the scientific community "can yet have a bearing on the inquiry." *Id.* None of these factors, except presumably the first which excludes purely metaphysical theorizing, is " a *sine qua non* of admissibility," for "[t]he inquiry envisioned by Rule 702 is, we emphasize, a flexible one." *Id.*

31. To "assist the trier of fact to understand the evidence or to determine a fact in issue," as Rule 702 requires, the scientific testimony must be "relevant" and "fit" the circumstances of the case. 113 S.Ct. at 2795, 2796. Assuming that the identity of the actual criminal is a contested issue, properly conducted DNA tests of identity always will be relevant. The only arguable lack of "fit" might involve the selection of a data base for computing related statistics. *See infra* text accompanying note 155.

evidence are to be admissible.

I. DECIDING WHETHER DNA FRAGMENTS MATCH

The most common form of DNA analysis in criminal cases utilizes four or five so-called "single locus VNTR probes"[32] to produce a "multilocus genotype," or, more simply, a "DNA profile." DNA is a complicated but stable organic compound found in the cells of all organisms, from the humblest amoeba to the most arrogant human being. It is composed of two weakly connected strands of molecules that spiral around one another to form a double helix. Along the backbone of each strand are much smaller, relatively flat molecules known as nucleotide bases. There are four such bases, often referred to by their initials, C, T, A, and G. The C on one strand always pairs with the G on its complementary strand, and the A with the T. A little reflection reveals that there is an incredibly large number of possible orderings of these base pairs in a lengthy stretch of DNA.[33]

Using techniques of molecular biology,[34] fragments of chromosomes[35] that begin and end with certain sequences of DNA base pairs are excised from samples found in blood, semen, or other material containing sufficient DNA.[36] The beginning and ending sequences are chosen so that

---

32. *See infra* note 40 and accompanying text.

33. At each site, there are four possible pairs: AT, TA, CG, or GC. Two sites produce $4 \times 4 = 16$ possibilities, three produce $16 \times 4 = 64$, and so on, so that n sites can accommodate 4^n possibilities.

34. *See generally* MAXINE SINGER & PAUL BERG, GENES & GENOMES: A CHANGING PERSPECTIVE (1991).

35. A chromosome is essentially a tightly coiled molecule of DNA. Each parent supplies one each of 23 different chromosomes, so the human genome consists of 46 chromosomes arranged into 23 pairs. On the coiling of DNA, see, for example, Michael Grunstein, *Histones as Regulators of Genes*, SCI. AM., Oct. 1992, at 68.

36. Bacterial enzymes are used to cut the DNA into fragments. A given "restriction enzyme" binds to DNA when it encounters a certain short sequence of DNA base pairs and cleaves the DNA at a specific site. For example, the *Hae III* enzyme cleaves the strand ...GGCC... to yield ...GG and CC... . "Digesting" DNA with such an enzyme usually produces fragments ranging from several hundred to several thousand base pairs in length.

A technique for copying DNA permits minute quantities of DNA to be analyzed. *See, e.g.*, Henry A. Ehrlich et al., *Recent Advances in the Polymerase Chain Reaction*, 252 SCIENCE 1643 (1991). In most forensic applications to date, DNA that has been "amplified" in this way has been probed at less revealing loci that do not involve VNTRs. *See, e.g.*, State v. Williams, 599 A.2d 961 (N.J. 1991) (unopposed testimony of prosecution experts established general acceptance of PCR amplification followed by

the material they bracket tends to vary in size from person to person.

The lengths of the DNA fragments are measured by seeing how far they move through a slab of gelatinous material when attracted by an electric charge relative to DNA fragments of known lengths—a process known as electrophoresis. Just as a sleek panther can wend its way through a stretch of dense jungle more readily than a bulky elephant, in a given period of time shorter fragments (with low molecular weight) migrate farther in an electrophoretic gel than longer fragments (of high molecular weight).[37]

The variations in the lengths of the fragments from different people, referred to as "fragment length polymorphisms,"[38] result primarily from disparities in the number of repetitions of a short sequence of nucleotide base pairs.[39] The number of repetitions of this core or "consensus" sequence varies greatly among people—hence the phrase "variable number of tandem repeats," or VNTRs.[40] The fragments containing the tandem repeats can be detected by specially constructed molecular "probes" that bind to a specific consensus sequence.[41] By measuring the

---

dot-blot detection of HLA DQα polymorphism). Amplification coupled with more precise detection of VNTRs is, however, also possible and likely to dominate forensic applications in the near future. *See, e.g.*, Bruce Budowle et al., *Analysis of the VNTR Locus D1S80 by PCR Followed by High Resolution PAGE*, 48 AM. J. HUM. GENETICS 137 (1991). For a set of proposed safeguards in forensic PCR analysis, see NRC REPORT, *supra* note 15, at 63-73.

37. The molecular weight of a compound is equal to the total mass of its constituent atoms. Since DNA fragments all have pretty much the same mix of atoms, a fragment that has twice the length of another also has about twice the molecular weight.

38. They also are called "RFLPs" or "AmFLPs," depending on the procedure that yields the fragments.

39. *See* Yusuke Nakamura et al., *Variable Number of Tandem Repeat (VNTR) Markers for Human Gene Mapping*, 235 SCIENCE 1616 (1987). More than one core sequence may be repeated in some VNTRs. *See* Alec J. Jeffreys et al., *Minisatellite Repeat Coding as a Digital Approach to DNA Typing*, 354 NATURE 204 (1991) (proposing an analysis of the order in which two interspersed core sequences appear in the repetitive portion of fragments in order to provide greater discrimination).

40. The chromosomal locations that give rise to VNTR RFLPs or AmFLPS are known as VNTR or hypervariable loci. The number of repetitions of the core sequence can vary from a handful to a few hundred, depending on the particular locus, but, when the length of the core sequence is short, the differences between fragments from different subjects will be too small to detect on a typical electrophoretic gel. Resolution of fragments which differ by as little as two base pairs, however, has been reported with newer gels. *See* Rene Hubert et al., *A New Source of Polymorphic DNA Markers for Sperm Typing: Analysis of Microsatellite Repeats in Single Cells*, 51 AM. J. HUM. GENETICS 985 (1992).

41. These probes are short segments of single-stranded DNA with a radioactive or other readily identifiable component attached, like a sticker or tag on a suitcase. When the probe encounters a strand of DNA with the complementary sequence of bases, it pairs ("hybridizes") with the target DNA.

distances that the fragments tagged with these probes have migrated,[42] the approximate lengths of the VNTR fragments can be determined. Although VNTRs used in forensic work represent a minuscule portion of the full genome, the number of distinct combinations of them easily runs into the billions and trillions.[43]

flanking region	many repeats of a single consensus sequence	flanking region

----------> > > > > > > > > > ... > > > > > > > > > > ----------

Figure 1.

A schematic diagram of a VNTR fragment. Between two flanking regions of DNA (-) are many repeats of the same small sequence of base pairs (the consensus sequence >). The number of repeats often varies, as between the pair of chromosomes in an individual and as among the chromosomes from different people.

Because the prevailing method of agarose gel electrophoresis for

---

42. In one common procedure for "visualizing" the target DNA, the DNA is denatured to its single-stranded form and transferred from the electrophoretic gel to a nitrocellulose filter. The probe is applied to the filter, and any excess, unbound probe is washed away. X-ray film is placed next to the filter. Radioactivity from the probe exposes the film, producing a black band whose location reveals how far the restriction fragment migrated on the gel. *See generally* JAMES D. WATSON ET AL., RECOMBINANT DNA (2d ed. 1992). Many opinions liturgically recite all the steps of this "Southern blotting" procedure. *See, e.g.*, Springfield v. State, 860 P.2d 598 (Wyo. 1993) (quoting Fishback v. People, 851 P.2d 884 (Colo. 1993)).

43. A typical fragment from a given region of a chromosome (a "locus") easily can come in 20 or more discernibly different sizes ("alleles"). *See, e.g.*, S.J. Odelberg et al., *Characterization of Eight VNTR Loci in Agarose Gel Electrophoresis*, 5 GENOMICS 915 (1989). For the maternal and paternal pair of chromosomes, then, there are many possibilities: maternal "allele" 1 can pair with paternal "allele" 1, 2, 3, ..., or 20; maternal "allele" 2 can pair with paternal "allele" 1, 2, 3, ..., or 20; and so on. The result is $20 \times 20 = 400$ possible "allele" pairs. Without a study of family members to ascertain which "allele" is on which chromosome, however, a single-locus VNTR probe cannot distinguish a paternal-maternal "allele" pair from a maternal-paternal one. On a gel, the pair (1,2), for example, looks the same as the pair (2,1). The 400 possibilities therefore includes $(20 \times 19)/2 = 190$ duplicates, leaving 210 discernible single-locus "genotypes." At four such loci, $210 \times 210 \times 210 \times 210 = 1{,}944{,}810{,}000$ discernible "genotypes" are possible; likewise, $210^5 = 408{,}410{,}100{,}000$ distinguishable five-locus "genotypes" are possible. Obviously, not all of the mathematically possible combinations are realized in any human population, and some may be represented more frequently than others. The branch of biology that studies the distribution of genotypes across populations and within populations over time is known as population genetics.

measuring the lengths of the VNTR fragments is not sensitive enough to distinguish between fragments that are extremely close in size,[44] laboratories declare a match when two conditions are met. First, the examiner must feel that the crime sample fragments and the suspect's fragments have migrated the same distance on the gel. Second, computerized measurements must confirm that the difference in migration distances is less than plus-or-minus three (or some other number[45] of) standard deviations[46] of a set of independent, duplicate measurements.[47] The

---

44. *See supra* note 40. *Contra* State v. Vandebogart, 616 A.2d 483, 486 (N.H. 1992) ("A variation of even one nucleotide in the sequence of DNA is detectable.").

45. There is confusion as to what rule different laboratories actually use. The FBI requires that bands be separated by no more than ±2.5% of their mean molecular weight to declare a match. This window is slightly larger than the biggest difference observed in the FBI laboratory for the same sample measured twice. *See* Bruce Budowle et al., *Fixed Bin Analysis for Statistical Evaluation of Continuous Distributions of Allele Data from VNTR Loci, for Use in Forensic Comparisons*, 48 AM. J. HUM. GENETICS 841, 844 (1991). According to Eric S. Lander, *Invited Editorial: Research on DNA Typing Catching Up with Courtroom Application*, 48 AM. J. HUM. GENETICS 819, 820 (1991), this FBI study suggests that the Bureau's laboratory has a standard deviation for the difference between two measurements of about 1.5% of the molecular weight of their mean. If so, the ±2.5% match window corresponds to ±1.7 standard deviations. On the other hand, Neil J. Risch & Bernard Devlin, *On the Probability of Matching DNA Fingerprints*, 255 SCIENCE 717, 720 n.9 (1992), conclude from the same FBI study that the FBI "measurement error SD" is 0.625%, which implies a match window of four standard deviations. Likewise, Seymour Geisser, *Some Statistical Issues in Medicine and Forensics*, 87 J. AM. STAT. ASS'N 607, 609 (1992), comments that "the FBI . . . will declare a match between two samples if the bands are within 2.5% of the average of the two . . . a tolerance of about 4 standard deviations of the difference." Britain's Home Office Forensic Science Service estimates the standard deviation of the difference between two independent measurements to be 1.1%. *See* Donald A. Berry et al., *Statistical Inference in Crime Investigations Using Deoxyribonucleic Acid Profiling*, 41 APPLIED STAT. 499, 502 (1992). According to Lander, *supra*, and Risch & Devlin, *supra*, commercial laboratories report still smaller figures of about 0.6%. Lifecodes Corporation uses a match window of 1.8%, *See* Bruce S. Weir, *Review: Population Genetics in the Forensic DNA Debate*, 89 PROC. NAT'L ACAD. SCI. 11654, 11655 (1992), which amounts to ±3 standard deviations of the reported measurement error.

Some of these discrepancies may arise from differences in the materials being examined in the "calibration" studies. More consistent results may be expected from fresh DNA; evidentiary samples can be influenced by degradation and exhibit band shifting. The characterizations of the FBI's match window for forensic casework as ±4 standard deviations of the mean of the two fragments actually pertain to the standard deviation derived from their K12 cell line ("control DNA") measurements. It also should be noted that a match window of ±k standard deviations of the mean of the two fragments implies that the two fragments could be ±2k standard deviations apart and still "match." *See infra* note 47; Michael J. DiRusso, Note, *DNA "Profiles"—The Problems of Technology Transfer*, 8 N.Y.L. SCH. J. HUM. RTS. 183, 205 (1990) (criticizing Cellmark for reporting that it used a match window of ±3 standard deviations when it actually was using ±6).

46. The standard deviation is a statistic that measures the degree of variation in a set of numbers. If all the numbers are identical, their standard deviation is zero. If they vary greatly from their mean, then their standard deviation is large. For electrophoretic

observed differences seen in repeated measurements of DNA fragments of the same length thus define the "match window"—the range within which two bands can be declared to match.[48]

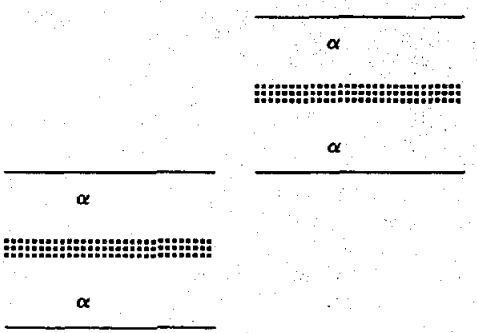
---

measurements, the standard deviation is greater for larger fragments (which migrate smaller distances on the gel).

47. In practice, reproducibility studies typically involve comparing the fragment lengths for VNTRs in DNA obtained from vaginal swabs (containing epithelial cells) and from blood taken from the same woman. Both samples contain the same DNA, and the two sources correspond to the situation in rape cases. An alternative would be to compare semen and blood samples from the same man. In such studies conducted by the South Carolina Law Enforcement Division DNA laboratory, corresponding bands never differed by more than 5.6% of their average length. B.S. Weir & B.S. Gaut, *Matching and Binning DNA Fragments in Forensic Science*, 34 JURIMETRICS J. 9 (1993). Such studies enable the laboratory to choose a window that is wide enough to be likely to result in a match when two samples come from the same source.

48. Weir & Gaut, *supra* note 47, lucidly describe the process:

> Suppose the vaginal sample length is denoted by e and the blood sample length by s. Then the relationship
>
> $$\frac{|s-e|}{(s+e)/2} \leq 2\alpha$$
>
> defines α (0.028 for the South Carolina laboratory). Alternatively, each band is no more than α from the average of the two lengths:
>
> $$\frac{|s-(s+e)/2|}{(s+e)/2} \leq \alpha$$
>
> $$\frac{|e-(s+e)/2|}{(s+e)/2} \leq \alpha$$
>
> This situation is shown below. In other words, there is uncertainty associated with an estimated band length. The true length of a band of estimated length e is thought to be contained in the interval $e \pm \alpha$, and two bands are said to match if they are no more than 2α apart:



Since the declaration of a match includes a subjective component (though it need not) which is inherently somewhat arbitrary, one would expect to see the process attacked in court. Indeed, it has been, but the challenges have met with little success. The Minnesota Supreme Court, in *State v. Jobe*,[49] deflected a challenge to the subjective phase of matching with the observation that "each sample is also examined by a second trained examiner and ultimately the 'match' is confirmed or rejected through computer analysis, using wholly objective criteria."[50] Likewise, courts in Arizona,[51] California,[52] and New York[53] have held that high standards of accuracy, such as the NRC panel's call for "a precise and objective matching rule"[54] do not require exclusion of results of the FBI's or Cellmark's match procedures.

These holdings are correct. As long as no visual match will be reported as a match unless confirmed by the quantitative matching rule, the imprecise, subjective phase serves as only a preliminary filter. It means that in some cases where the purely statistical rule would declare a match the laboratory will not report a match. When a sample from a defendant matches both objectively *and* subjectively, the defendant can hardly complain that the laboratory should not have bothered with the subjective phase of the procedure.[55]

Judicial discussions of the adequacy of the objective, statistical phase of matching have been less perspicacious. The issue can surface both when the prosecution offers proof of a match, and when a defendant offers evidence of a non-match. In the former, inculpatory situation, a defendant might argue that the match window is too wide,[56] and a more

---

> Once this numerical matching rule has been established for a particular laboratory, the evidence bands e can be compared to bands s from a suspect. If a visual match is declared, and if all pairs of corresponding bands in the two profiles differ by no more than 2α, then the two profiles are said to match.

49. 486 N.W.2d 407 (Minn. 1992).
50. *Id.* at 420.
51. State v. Bible, 858 P.2d 1152 (Ariz. 1993).
52. People v. Barney, 10 Cal. Rptr. 2d 731 (Ct. App. 1992).
53. People v. Wesley, 589 N.Y.S.2d 197 (App. Div. 1992).
54. NRC REPORT, *supra* note 15, at 72.
55. When the laboratory reports the proportion of people in the general population with DNA that would match the crime sample, it uses the purely statistical match rule. To the extent that the subjective component can only reduce the number of matches in the population, this frequency tends to overstate the degree to which the DNA test would incriminate innocent people. Of course, there may be separate reasons to question these estimates of population frequencies. These are analyzed *infra* in Part II.
56. *See, e.g.*, United States v. Yee, 134 F.R.D. 161, 207-08 (N.D. Ohio 1991); *see*

stringent rule that would preclude the declaration of a match should be used.[57] In many cases, this argument will be futile, for all the pairs of measurements will lie well within the match window.[58] Indeed, when this happens, experts may refer to the concordance of the two measurements within the match window as "an exact match,"[59] or "conclusive matches."[60] Conversely, when at least one pair of measurements spans nearly the full length of the window, and an analyst just speaks of "a match," a court may still admit the evidence, as did the magistrate judge in *United States v. Yee*.[61] That court justified its holding with the observation that "defendants who would be outside a smaller window but are within the F.B.I.'s larger window can make that point clear at trial."[62]

Although this may sound like a reasonable compromise, the *Yee* suggestion invites a potentially confusing exchange. The prosecution says to the defendant, "under our match rule, you match." The defendant replies, "That's your rule. Under a different rule, I don't match." What is the jury to make of this thrust and counterthrust? If all goes well, the exchange will make no difference because the jury also will be presented with the frequency with which the prosecution's procedure for declaring

---

also Geisser, *supra* note 45, at 609 (characterizing the ±2.5% window as "an extraordinarily wide net to declare a match").

57. In United States v. Jakobetz, 747 F. Supp. 250 (D. Vt. 1990), *aff'd*, 955 F.2d 786 (2d Cir. 1992), *cert. denied*, 113 S.Ct. 104 (1992), the defense attacked the FBI's match window of ±2.5% on the ground that "[t]he FBI derived this 5% window through an empirical analysis based upon the total variation of matches from known samples rather than a statistical approach that utilizes confidence intervals. Defense expert Dr. [Joseph] Nadeau testified that the distinction renders the FBI's mathematical approach scientifically unacceptable." *Id.* at 257. Since the statistical properties of a match window do not depend on how it was derived, the criticism that the court describes is misdirected.

58. This was the case in State v. Vandebogart, 616 A.2d 483, 488 (N.H. 1992) ("The FBI confirmed a visual match . . . because the degree of variation did not exceed plus or minus one percent."), State v. Alt, 504 N.W.2d 38, 47 (Minn. App. 1993) ("The greatest variance between Alt's DNA and any of the forensic DNA specimens on any of the probes is 1.3%, approximately half the size of the match window, and within the match windows suggested by the defense experts.") and *Jakobetz*, 747 F. Supp. at 257. After criticizing the ±2.5% rule, the defense expert in *Jakobetz* "conceded that if the autorad matches . . . were within plus or minus 1% of the number of base pairs, he would have more confidence in the conclusion that there was in fact a match." *Id.* This allowed the government nimbly to sidestep the criticism by pointing out that "all sixteen band matches (eight alleles from each the victim and the suspect on four different autorads) were within plus or minus 1%." *Id.* at 258.

59. *See, e.g.*, State v. Hammond, 604 A.2d 793, 802 (Conn. 1992) (noting that the "female portion of the DNA that came from the [semen] stain matched 'exactly' with that of the victim").

60. *See, e.g.*, *Jakobetz*, 747 F. Supp. at 257.

61. 134 F.R.D. 161 (N.D. Ohio 1991).

62. *Id.* at 208.

a match would produce reports of matching DNA in randomly chosen members of the reference population.[63] If the procedure almost always results in matches for samples from the same source and if the match would be rare among innocent people, then the evidence proves something, and its probative value is unaffected by the truism that the same measurements would not match under some even more stringent rule.[64] The defense testimony contemplated in *Yee* therefore proves very little.[65] Disputes over the strictness of particular windows[66] or *the* optimal match window—when there is no such thing[67]—may confuse and perplex the jury

---

63. *See infra* notes 149-208 and accompanying text.

64. Nevertheless, within the window that a laboratory uniformly applies to declare a match, some matches—those well within the window—are more probative than others. Thus, the defense (or the prosecution) should be permitted to argue that the smallest window that could produce a match between the crime scene DNA and the defendant's DNA is at least as pertinent as any broader window that also produces a match, and to introduce appropriate statistics about the narrower window. In particular, one could argue that the probative value of the closer match depends on the frequency with which the minimally matching window produces matches in reproducibility studies as compared to the frequency corresponding within the reference population. (I am indebted to William C. Thompson for this insight.) *Cf. infra* Part III(C) (likelihood ratio as a measure of probative value). But once the frequency of match with a given window is presented, merely introducing testimony that there exists another window that excludes the defendant is not particularly edifying. *See infra* note 65.

65. Once a jury knows that the match window is large enough to ensure that almost all duplicate measurements produce matches and that a defendant's VNTR fragments match the forensic sample in a way that would occur at a frequency of say, 1/100,000, in the relevant population, it gains little or no useful information from hearing that the fragments do not match in a smaller window that would produce a smaller frequency of, say, 1/200,000, if they did match.

66. In Perry v. State, 606 So. 2d 224, 225 (Ala. Crim. App. 1992), the appellate court reassured itself that a match declared by Lifecodes was acceptable because Lifecodes's match window "of 1.8%, was stricter than that used by the FBI . . . which [is] 2.5 percent." *Cf.* State v. Pierce, 597 N.E.2d 107, 113 (1992) (Cellmark's determination of a match was not unreliable just because "other laboratories and experts may use somewhat different criteria."). Such comparisons of these raw percentages, however, are misleading unless the standard errors of the laboratories are comparable. *See supra* note 45. Lifecodes's standard error is smaller than the FBI's, which makes Lifecodes's window more lenient than the percentages would suggest.

67. Neither statistical theory nor legal doctrine dictates the ideal size of the window. The former informs us that we can reduce the risk of falsely declaring a match only by increasing the risk of incorrectly failing to declare a match. Big match windows make for fewer false exclusions; small windows result in fewer false inclusions. With a match window of two standard deviations per independent comparison, the risk that at least one comparison out of ten for samples from the same person will not show a match is not .05, but $1 - .95^{10} = .40$. Increasing the window to three standard deviations obviously produces more matches when the samples being compared come from different people, but it reduces the risk of failing to declare a match when the samples being compared come from the same person to $1 - .99^{10} = .10$. For more sophisticated studies of real data, see Berry et al., *supra* note 45, at 520 (match-binning with a window of ± 2.5 standard deviations gives false exclusion rate of nearly 2% per probe) and Ian W. Evett

when it considers the probative value of a match. As a result, a court has discretion to exclude this testimony.[68]

The appropriateness of the rule for declaring matches also comes into play when the defendant seeks to prove that no match can be declared under the usual matching rules.[69] Again, it often will be the case that the exclusion results from measurements that place a pair of fragments well

---

et al., *An Illustration of the Advantages of Efficient Statistical Methods for RFLP Analysis in Forensic Science*, 52 AM. J. HUM. GENETICS 498, 502 (1993) (152 3-probe duplicate measurements produced 20% false exclusions for a window of ±1.2% and 2.6% for a window of ±2%.).

As for legal doctrine, one might think that some courts' reliance on a "two or three standard deviation rule" in discrimination litigation should dictate the use of an interval that spans the same number of standard deviations. This thought should be resisted. First, the rule itself does not mesh well with the more-probable-than-not or other evidentiary standards of proof. *See, e.g.*, David H. Kaye, *Hypothesis Testing in the Courtroom*, *in* CONTRIBUTIONS TO THE THEORY AND APPLICATION OF STATISTICS (A. Gelfand ed. 1987); David H. Kaye, *Is Proof of Statistical Significance Relevant?*, 61 WASH. L. REV. 1333 (1986). The two-standard-deviation rule for a normally distributed statistical measure of the difference between two groups merely means that a disparity of at least that magnitude will occur about 5% of the time that the rule is applied to cases where the disparity is a statistical fluctuation rather than a reflection of any real disparity. It says nothing about the frequency with which the rule will fail to identify true disparities when they are present, and it does not imply that one can be 95% "confident" that a disparity outside the 95% interval is due to an impermissible criterion. *Id.*; David H. Kaye, *Apples and Oranges: Confidence Coefficients Versus the Burden of Persuasion*, 73 CORNELL L. REV. 54 (1987). Second, whatever may be the probability that a single measurement deemed "significant" under the two-standard-deviation rule is due to chance, declaring that two samples of DNA match on five probes requires ten comparisons. As shown above, multiple comparisons give the rule quite different statistical properties. Instead of devising rules that treat relevant evidence as either admissible or inadmissible, as totally revealing or utterly worthless, the law here should be concerned with conveying to the judge or jury sufficient information to gauge the probative value of the evidence—the extent to which the various pairs of fragment lengths match.

68. *See* Fed. R. Evid. 403. If all the defense can say is that a sufficiently small window would not include the defendant, it is arguable that anything less than exclusion of the proposed testimony would be error. However, even if the defense fails to undertake a more careful analysis of the precise degree of matching and its implications, the testimony about small enough windows could prompt the prosecution to do so. *See supra* note 64. And, since the prosecution should be able to demonstrate the tautological nature of the defense argument, the danger of prejudice is not overwhelming. Consequently, a strict exclusionary rule may not be needed even in this situation. It suffices to leave it to the trial court to inquire whether additional analysis of the minimally matching window gives a substantially different picture than the frequency associated with the laboratory's conventional match window. If it does, argument about the effect of smaller windows should be allowed; if it does not, the testimony invites a pointless digression and should be excluded.

69. It has been said that the exclusion rate for most laboratories is about 30%. Bernard Devlin et al., *Statistical Evaluation of DNA Fingerprinting: A Critique of the NRC's Report*, 259 SCIENCE 748 (1993). Some of the exclusions have been dramatic. *See, e.g.*, Jonathan Rabinovitz, *Rape Conviction Overturned on DNA Tests*, N. Y. TIMES, Dec. 2, 1992, at B6 (man convicted of rape released after 11 years in prison).

outside the match window, and not much will turn on the precise width of the window.[70] But what happens when the putative exclusion is a closer call? A narrow match window reduces the number of falsely included defendants at the cost of excluding a large number of guilty defendants whose DNA fragments no longer "match" the crime sample.[71] Because there can be little difference between a pair of bands that barely falls into a match window and a pair that barely falls outside the same window, to consider the former an inclusion and the latter an exclusion would be misleading. To avoid this outcome, analysts may be tempted to designate such weak exclusions as "inconclusive."[72] Thus, in evaluating a defendant's effort to introduce non-matches as exculpatory evidence, the judge or jury should attend to the degree of non-matching and not just the label.

The standard matching procedure, with fixed match windows, does not lend itself to this task, but other statistical procedures do. They replace the somewhat artificial match *vs.* no match dichotomy with an inquiry into (a) the probability of finding the observed degree of congruence in the crime fragments and the defendant's fragments when all the fragments come from the same person, and (b) the probability of finding this

---

70. Presumably, this was the case in State v. Hammond, 604 A.2d 793 (Conn. 1992). In this unusual case, an FBI analyst testified on behalf of a man accused of rape. The analyst stated that the tests had been run properly because the "female portion of the DNA that came from the [semen] stain matched 'exactly' with that of the victim," while "neither the defendant nor the victim's boyfriend could have contributed any part of the semen stain on the victim's underwear." Even when the exclusions are clear, however, there remains a non-zero probability that the samples came from the same source. Consequently, occasional dicta like that offered by the Arizona Supreme Court in *State v. Bible*, 858 P.2d 1152 (Ariz. 1993), that "if samples do not match, they *must have* come from different individuals," cannot be taken literally.

71. *See supra* note 67.

72. The term "inconclusive" is appropriate as applied to testing that fails to produce any measurements at all. *See, e.g.*, State v. Woodall, 385 S.E.2d 253, 260 (W. Va. 1989) (holding that failure to match defendant's DNA with the sample from a rape victim was irrelevant because the crime sample lacked sufficient high molecular weight DNA to make any comparison). It is more problematic when used to designate a close non-match, or worse, when used to dismiss selected pairs of length measurements that almost match, so that a frequency of matches for the remaining probes in the population can be computed. *Cf.* NRC REPORT, *supra* note 15, at 61 ("When samples fall outside the match criterion, they should be declared to be 'inconclusive' or 'nonmatching.'"). For cases that may violate this precept, see Polk v. State, 612 So. 2d 381, 392 (Miss. 1993) (Cellmark's expert "testified that seven of eight bands from Georgia Mae Thomas's DNA met the criteria to be determined a match with the DNA obtained from the blood on Polk's underwear") and State v. Quatrevingt, 617 So.2d 484, 492 (La. Ct. App. 1993) (determination of a match was "supported by the evidence" that "two of the three percentages for probe DXYS14" fell within Lifecodes's match window).

congruence when the fragments come from different people.[73] Before considering these resulting alternatives to categorical matching, however, we should investigate the second part of the current mode of presenting DNA test results in court—estimating the frequency of matching profiles in some relevant population.

II. ESTIMATING MATCH-BINNING FREQUENCIES

If a match is declared, the weight of the evidence depends on the probability of such a result if the suspect is the source of the sample (an event that we may denote as S), compared to the probability of a match if someone other than the suspect is the source (O).[74] Although some experts seem to say that there is no chance of a false inclusion,[75] there are scenarios that would produce such an error,[76] and there are reported instances of false positive identifications.[77] In addition, even if the DNA fragments really are within the match window, there is some probability that other people have fragments in this region. If the relative frequency of the incriminating fragment sizes is large, so that many people would match, then the finding of a match is not very probative. Estimating the frequency requires some analysis of population data, and the adequacy of such analyses is controversial. Furthermore, even if a correct population frequency can be found, there is a risk that it will be interpreted as the

---

73. *See infra* Part III.

74. *See, e.g.*, Richard Lempert, *Some Caveats Concerning DNA as Criminal Identification Evidence: With Thanks to the Reverend Bayes*, 13 CARDOZO L. REV. 303 (1991). *See generally* David H. Kaye, Comment, *Quantifying Probative Value*, 66 B.U. L. REV. 761 (1986).

75. *See, e.g.*, Fishback v. People, 829 P.2d 489, 492 (Colo. Ct. App. 1991), *aff'd*, 851 P.2d 884 (Colo. 1993); People v. Shi Fu Huang, 546 N.Y.S. 2d 920, 921 (Sup. Ct. 1989).

76. *See* William C. Thompson & Simon Ford, *in The Meaning of a Match: Sources of Ambiguity in the Interpretation of DNA Prints*, FORENSIC DNA TECHNOLOGY (M. Farley & J. Harrington eds., 1990).

77. *See* State v. Bible, 858 P.2d 1152, 1180 n.16 (Ariz. 1993) (referring to reports of errors in paternity determinations); Lempert, *supra* note 74, at 324-25; NRC REPORT, *supra* note 15, at 88. *But see* People v. Mehlberg, 618 N.E.2d 1168, 1180 (Ill. App. Ct. 1993) (testimony of Robin Cotton of Cellmark Diagnostics that reports of Cellmark's erroneous attribution of maternity to a woman in Maryland are mistaken). There also are instances of "clerical errors" in calculating the frequency of matching DNA patterns in the general population. *See* Perry v. State, 606 So. 2d 224, 226 (Ala. Crim. App. 1992) (original three-locus frequency estimated to be 1/209,100,000 instead of 1/23,000,000 due to "clerical error").

probability that someone other than the defendant is the source of the evidence sample.[78] As a result, the use of match frequencies or probabilities has proved susceptible to challenge in court.

Martinez v. State[79] illustrates the type of testimony as to frequencies or probabilities that has provoked objections. In *Martinez*, Lifecodes Corporation "explained the significance of the match of DNA patterns" in the following way:

> Q. And what would be the answer to that question as far as the likelihood of finding another individual whose bands would match up in the same fashion as this?
> A. The final number was that you would expect to find only one individual in 234 billion that would have the same banding pattern that we found in this case.
> Q. What is the total earth population, if you know?
> A. Five billion.
> Q. This is in excess of the number of people today?
> A. Yes. Basically that's what that number ultimately means is that that pattern is unique within the population of this planet.
> Q. Is that consistent with your opinion earlier that the semen involved in this case came from Fernando Martinez?
> A. That is correct.

The defendant, Martinez, argued that the introduction of this testimony was error simply because "a figure 47 times larger than the world's current total population was 'nonsensical'; and it was so overwhelming as to deprive the jury of its function in fairly appraising all of the evidence." The Florida district court of appeals rejected this broad-brush argument against small frequencies, but other courts have been more sympathetic, especially when more focused arguments have been advanced and supported by expert testimony for the defense.[80] Indeed, the procedure for computing frequencies like the one in *Martinez* also has inspired the sharpest debate about DNA evidence outside of the court-

---

78. *See, e.g.*, People v. Shi Fu Huang, 546 N.Y.S.2d 920, 921 (Sup. Ct. 1989) ("If there is an adequate and reliable data base, a forensic scientist can calculate that a match did not occur by chance."); Lempert, *supra* note 74, at 306; *infra* text accompanying note 233.

79. 549 So.2d 694 (Fla. App. 1989).

80. *Compare* Perry v. State, 586 So.2d 242 (Ala. 1991), *with* Snowden v. State, 574 So.2d 960 (Ala. Crim. App. 1990).

room, in the pages of scientific journals.

This section therefore considers several procedures for computing the frequency of an incriminating set of DNA fragment lengths and the statistical and legal objections that can be raised. Part III considers the problems in using even an accurately determined population frequency to gauge the significance of a match.

A. Direct Estimation

How can one estimate the proportion of people in the relevant population whose DNA fragments would be considered to match the set of measured lengths of the VNTR fragments derived from the crime sample? One procedure recommended by several commentators[81] is simply to sample people in the relevant population, analyze their DNA, and report the number who match the crime sample.[82] Thus, the laboratory might report that of the, say, $N = 1{,}000$ DNA samples it has analyzed, only the defendant's was found to match the crime sample. The National Research Council report recommends this approach, at least for the time being.[83]

---

81. *See, e.g.*, Richard C. Lewontin & Daniel L. Hartl, *Population Genetics in Forensic DNA Typing*, 254 SCIENCE 1745 (1991); David A. Stoney, *Reporting of Highly Individual Genetic Typing Results: A Practical Approach*, 37 J. FORENSIC SCI. 373 (1992) (recommending direct estimation supplemented by more theoretical methods).

82. In deciding whether a sample of DNA in the database matches, the laboratory should apply the same matching rule, with the standard error applicable to inter-gel comparisons, that it used to declare a match in the case at bar. However, a broader match window for counting matches in the database could only lead to an overestimate of the population proportion; it would not prejudice a defendant who objects to DNA evidence of a match.

83. NRC REPORT, *supra* note 15, at 91. For no apparent reason, the recommendation is limited to cases in which no multilocus matches in the database are observed. In discussing this "counting" method, the panel also suggests that "an upper confidence limit of the frequency should be used in court" because "estimates used in forensic science should avoid placing undue weight on incriminating evidence" and "any loss of power can be offset by studying additional loci." *Id.* at 75. The first reason, however, begs the question: How does unbiased estimation place "undue" weight on the evidence? One could argue against the panel's suggestion, with equal force, that estimates used in forensic science should avoid placing too little weight on incriminating evidence. As for the panel's reliance on testing additional loci to enhance statistical power, such testing would not increase the number of people in the database; consequently, it might have no effect on power (as indicated by the width of the confidence interval). Furthermore, the panel misconstrues the meaning of the most common confidence interval when it explains that "the traditional 95% confidence interval . . . implies that the true value has only a 5% chance of exceeding the upper bound." A 95% confidence interval is computed according to a procedure that, if applied to many random samples from the same population, would include the population proportion in about 95% of these

As the NRC panel observes, the advantage of direct estimation is that it requires no theoretical assumptions (except in defining the reference population of possible perpetrators) and no knowledge of the dependencies among the restriction fragments. Even so, its use in court is subject to at least three objections.[84] For one, it grossly understates the evidential value of the incriminating match. As explained below, there is every reason to believe that matches are far less frequent than $1/N$. Of course, this does not mean that the defendant should be able to exclude the figure of $1/N$, which errs in his or her favor, but it counsels against a rule that would make it the sole indication of the significance of the incriminating match.[85]

The second objection is that a random sample of the relevant population is essential to a valid estimate, but existing databases are convenience samples.[86] This point has been consistently rejected in court,[87] largely because it is felt that the distribution of VNTRs is no different in a convenience sample than in a random sample.[88] Some

---

samples. Each sample, and hence each interval, would be different, and one *cannot* say that there is a 95% chance that the population proportion lies within the one and only available 95% confidence interval. *See, e.g.*, DAVID S. MOORE & GEORGE P. MCCABE, INTRODUCTION TO THE PRACTICE OF STATISTICS § 7.1 (2d ed. 1993). Despite these flaws in the report, presenting a confidence interval along with the sample frequency of matches is desirable, since it conveys information about the uncertainty in the unbiased point estimate. *See* Kaye, *supra* note 67.

84. Another objection, having to do with the choice of the reference population, is discussed *infra* text accompanying notes 149-208.

85. *Cf.* C. Thomas Caskey, *Comments on DNA-based Forensic Analysis*, 49 AM. J. HUM. GENETICS 893, 894 (1991) (The use of direct estimation "would represent a loss" of potential information available from the field of population genetics). The direct count frequency within a database would be the same, of course, for a match at 20 loci as for a match with testing at only one locus. Yet, the probability of a random match at 20 separate VNTR loci is many orders of magnitude smaller than that of a single locus match.

86. *See, e.g.*, State v. Bible, 858 P.2d 1152, 1186 (Ariz. 1993); People v. Mohit, 579 N.Y.S.2d 990, 998 (Sup. Ct. 1992); Geisser, *supra* note 45. In "convenience sampling," individuals are included in the sample because they are easily accessible. Not being the result of a procedure that gives every individual in the population a known probability of being sampled, the statistical properties of convenience samples are not well-defined. Lifecodes's samples come from paternity cases, while the FBI and Cellmark Diagnostics rely on bloodbanks. *See* Weir, *supra* note 45. Efforts to broaden or supplement the databases are underway. *Id.*

87. *But see* State v. Bible, 858 P.2d 1152, 1186 n.23 (Ariz. 1993) (observing that "frequency figures . . . are valid and accurate only if they come from a truly random sample," but purporting not to rely on this consideration in holding that Cellmark's calculation was erroneously admitted); Commonwealth v. Curnin, 565 N.E.2d 440 (Mass. 1991) (evincing concern that "Cellmark compiled its Caucasian data base by testing 200 blood samples collected at a New York City blood bank").

88. United States v. Jakobetz, 747 F. Supp. 250, 261 (D. Vt. 1990) ("Dr. Kidd

research supports this intuition.[89]

Third, it could be argued, as defendants challenging other estimates of population proportions have done, that existing databases are too small to permit useful estimates. Courts have also rejected this argument on the strength of conclusory statements by geneticists that sample sizes of a few hundred are sufficient to permit reasonable estimates.[90] However, the notion that some minimum size exists above which all estimates are reliable and below which none are hardly is in keeping with statistical theory. Even a small sample can supply a foundation for validly estimating the frequency of a characteristic in a vastly larger population. The appropriate reaction to the sample size concern is neither to reject the sample statistic out of hand nor to accept it without qualms, but to press for a range of estimates indicating the extent to which the calculation might vary from one such small sample to another.[91]

In sum, direct counts of the frequency of the incriminating DNA profile in the appropriate database ordinarily should be admissible. Indeed, at least one court has excluded indirect estimates in favor of more direct counts.[92]

---

testified that the composition of the data base may be less rigorous when the targeted genes or VNTRs occur randomly."), *aff'd*, 955 F.2d 758 (2d Cir. 1992), *cert. denied*, 113 S.Ct. 104 (1992); People v. Shi Fu Huang, 546 N.Y.S.2d 920 (Sup. Ct. 1989) (testimony from Lifecodes that a database of under 200 samples of university students from mainland China "'seemed' to be from a random sampling").

89. Bernard Devlin & Neil Risch, *Ethnic Differentiation at VNTR Loci, with Special Reference to Forensic Applications*, 51 AM. J. HUM. GENETICS 534, 545-46 (1992).

90. *See, e.g.*, *Jakobetz*, 747 F. Supp. at 261 ("According to Dr. Kidd, once it is determined that the alleles are randomly distributed throughout a targeted population, sample size can be decreased to as little as 100 individuals."); *Shi Fu Huang*, 546 N.Y.S.2d at 921 (minimum size said to be 200). For a more careful treatment of the issue, see Ranajit Chakraborty, *Sample Size Requirements for Addressing the Population Genetics Issues of Forensic Use of DNA Typing*, 64 HUM. BIOLOGY 141 (1992) and Ranajit Chakraborty et al., *Evaluation of Standard Error and Confidence Interval of Estimated Multilocus Genotype Probabilities and Their Implications in DNA Forensics*, 52 AM. J. HUM. GENETICS 60 (1993) (method for reporting match-binning frequencies that accounts for sampling error).

91. On the appropriate statistical procedures for producing such interval estimates, see Bruce S. Weir, *Forensic Population Genetics and the NRC*, 52 AM. J. HUM. GENETICS 437 (1993).

92. In Caldwell v. State, 393 S.E.2d 436, 443 (1990), Lifecodes's calculation that frequency of genotype in population was 1/24,000,000 was replaced with the figure of 1/250,000 derived from the "more conservative approach [of using] the database itself, and not 'any population theory.'" Because existing databases are much smaller than 250,000, however, it not obvious how "the database itself" was used to produce the 1/250,000 figure.

B. Inferences from "Allele" Frequencies

1. The Independence Method With Match Windows Equal to Bin Widths: Basic Bins

Direct estimates of match frequencies that give vanishingly small numbers, like those in *Martinez*, have become prevalent despite their shortcomings. The basic method now employed for inferring the "genotype" frequency from "allele" frequencies presupposes independence[93] of certain genetic characteristics and is therefore referred to as the "independence method."[94] The method involves three steps: estimating "allele" frequencies, deducing "genotype" frequencies at each locus, and deducing a "genotype" frequency for all the loci.

In the first step, for each DNA fragment, one counts the number of indistinguishable (or similarly sized) fragments in the database. This counting procedure is often called "binning" because it piles fragments of slightly different sizes into distinct "bins." To estimate the relative frequencies of different sized fragments, forensic laboratories use either "floating"[95] or "fixed"[96] bins.[97] DNA fragments of similar lengths that

---

93. Two events are independent if the occurrence of one is not associated with the occurrence of the other. Cards, dice, roulette wheels, coins, and balls in urns provide classic illustrations. For example, if a coin is tossed vigorously twice, obtaining a head on the second toss is independent of a head on the first. When events are independent, the probability of their joint occurrence is the product of the probabilities of each event: if the coin is fair, the chance of two heads is (1/2)(1/2) = 1/4.

94. Courts sometimes speak of a "product" or "multiplication rule" for independent events, but this terminology is infelicitous, for there is another multiplication rule, involving conditional probabilities, for dependent events. *See, e.g.*, William Fairley & Frederick Mosteller, *A Conversation About Collins*, 41 U. CHI. L. REV. 242 (1974); David H. Kaye, *Statistics for Lawyers and Law for Statistics*, 89 MICH. L. REV. 1520 (1991).

95. A floating bin is just the bar of a histogram centered on the length of a fragment seen in the incriminating DNA. (A histogram is a bar chart in which the heights of the bars represent the proportions of items in the range, or "bin," covered by that bar.) As Weir & Gaut, *supra* note 47, explain:

> Since any band of length s satisfying [the first equation of note 48] would be said to match an evidence band of length e, a bin is constructed around length e to contain all such lengths. From this equation, all matching lengths d must satisfy
>
> $$\left(\frac{1-\alpha}{1+\alpha}\right)e \le d \le \left(\frac{1+\alpha}{1-\alpha}\right)e$$
>
> These floating bins have approximate width 4α centered on the evidence

are put into the same bin constitute an "allele."[98] For example, if p_i denotes the proportion of each "allele" generated from these counts, the first fragment might migrate a distance that puts it in a bin containing $p_1 = 1/5$ of all fragments for that locus. The only legal objections that can be raised to the above procedure relate to the source and size of the

---

band. Each band in the database is examined, and those satisfying [the first equation of note 48] are assigned to the bin for that evidence band. In this way the bin frequency is obtained.

96. A fixed bin is a bar of a histogram established before referring to any of the observed fragments:

> A set of fragments of known length are used as bin boundaries. These fragments are those produced by digesting viral DNA with restriction enzymes, and the lengths serve as "sizing ladders" on electrophoretic gels. For binning, however, the important thing is that a set of bins are pre-defined with fixed boundaries. Once a match has been declared, the evidence band is assigned to a fixed bin. Because there is uncertainty associated with the length e of the evidence band, a window of width 2α centered on e is constructed. If this window lies wholly within a fixed bin, the band is assigned to that bin. If the window includes a bin boundary, it is not known to which fixed bin the true band length belongs. It is known, however, that the true length belongs to only one bin, and a conservative procedure is to assign the band to the bin with highest frequency.

Weir & Gaut, *supra* note 47. Consequently, "there is no logical basis for the recommendation of a recent National Research Council (NRC) report that the band be assigned to a bin obtained by adding the two adjacent bins in cases of overlap." *Id.*

The advantages of fixed binning are that the laboratory can estimate allele frequencies by consulting a table instead of performing new counts for each fragment in the crime sample and that statistical tests can be applied to the predefined bins to establish independence. In practice, the FBI uses bins whose widths exceed the match window, thus producing overestimates of allele frequencies.

97. In either case, the width of a bin should correspond to the laboratory's matching rule, using the standard error for inter-gel comparisons—an obvious precept that Lifecodes failed to observe in People v. Castro, 545 N.Y.S.2d 985 (Sup. Ct. 1989), but now abides by. *See* People v. Golub, 601 N.Y.S.2d 502, 504 (App. Div. 1993). For an empirical comparison of fixed and floating bins, showing that, on average, fixed bins produce larger frequency estimates for DNA profiles, see Keith L. Monson & Bruce Budowle, *A Comparison of the Fixed Bin Method with the Floating Bin and Direct Count Methods: Effect of VNTR Profile Frequency Estimation and Reference Population*, 38 J. FORENSIC SCI. 1037 (1993).

98. The term "allele" is taken from other contexts where it refers to a form of a gene, that is, a sequence of DNA that codes for observable traits. Two VNTR fragments of the same length would be considered "alleles" even though their base sequences might differ and even though they do not code for any known traits. Moreover, since the VNTR fragments in a database are clumped by size into bins, the bins contain a range of differently sized fragments. Therefore, a better term for a set of comparable fragments might be "binelle." *Cf.* Bernard Devlin et al., *Estimation of Allele Frequencies for VNTR Loci*, 48 AM. J. HUM. GENETICS 662 (1991) (procedure for deducing the frequencies of fragments with the same numbers of tandem repeats from the distribution of measured sizes).

database.[99]

The second stage of the analysis generates the frequency of the "genotype" at each VNTR locus.[100] It is here that the first independence assumption comes into play. Every person inherits two chromosomes that contain a particular VNTR, usually giving rise to two distinct fragments[101] for each enzyme-probe system.[102] Having estimated the relative frequency of each fragment size, one now computes the frequency of the observed pair {i,j} of fragments. If the population is in what geneticists call Hardy-Weinberg equilibrium,[103] then[104] the proportions of the "alleles" remain constant from one generation to another, and the proportion of this "genotype" $g_l = \{i,j\}$ at a locus l is just the product

$$P_l = 2p_ip_j \qquad (1).^{105}$$

---

99. The congruence of the database with the population of plausible suspects, which is treated below, is also open to attack. Contrary to the opinion of the Nebraska Supreme Court in State v. Houser, 490 N.W.2d 168, 183 (1992), Hardy-Weinberg equilibrium, see *infra* note 103, plays no role in the estimation of "allele" frequencies.

100. Since no genes are involved and the operationally defined "alleles" include a hodgepodge of true alleles, a more apt term would be "binotype." Devlin et al., *supra* note 98. In the usual terminology, however, a "single locus genotype" is the set of "alleles" detected by a single probe.

101. A single band will appear if a person's mother and father both transmitted the same allele (the person is homozygous) or if one band has not been detected.

102. As explained in Part I, each restriction enzyme cuts a long DNA molecule into much shorter fragments by cleaving a specific sequence of bases, and a probe binds to those fragments that contain varying numbers of the consensus sequence within these restriction sites. Consequently, the distribution of fragment sizes in the population depends on the enzyme and probe. After "digesting" DNA with one enzyme and applying a probe after electrophoresis and blotting, the probe can be washed from the DNA, and a probe that recognizes a different consensus sequence then can be applied. This probe identifies a length variation that starts at another location, or "locus," along the DNA molecule.

103. Hardy-Weinberg equilibrium follows rigorously under three conditions: (1) a Mendelian pattern of inheritance (no mutation and alleles segregate independently); (2) no selection (the expected number of fertile progeny from a mating that reaches maturity does not depend on the genotype of the mates); and (3) an infinite, unstructured population (i.e., matings and genotypes are uncorrelated in an infinite population). *See, e.g.*, L.L. CAVILLA-SFORZA & W.F. BODMER, THE GENETICS OF HUMAN POPULATIONS (1971).

104. The converse is not true. Hardy-Weinberg equilibrium is not a necessary condition for independence of alleles at a locus. Independence can exist in the presence of selection or non-random mating. *See* Richard C. Lewontin & C.C. Cockerham, *The Goodness-of-Fit Test for Detecting Natural Selection in Random Mating Populations*, 13 EVOLUTION 561 (1959); C.C. Li, *Pseudo-random Mating Populations. In Celebration of the 80th Anniversary of the Hardy-Weinberg Law*, 119 GENETICS 731 (1988).

105. The factor of 2 reflects the fact that "allele" i could lie on the chromosome inherited from the mother and j on the one from the father, or vice versa. In other words, the "genotype" could be written (i,j) or (j,i), where the first "allele" is from the maternal chromosome and the second is from the paternal one. *See supra* note 43. In a

Thus, if the first fragment at a locus is in a size range that contains 1/5 of all the fragments from this locus, and the second fragment at this locus falls in a bin that contains 1/10 of the fragments seen for people in the population, then the relative frequency of the "genotype" {1,2} at this locus 1 would be $P_1 = 2(1/5)(1/10) = 1/25$.

Initially, various experts argued that the number of homozygotes—individuals with apparently equal fragment lengths at a locus—exceeds the expected value under the assumptions of a Hardy-Weinberg equilibrium.[106] The argument swayed several courts,[107] and, indeed, most opinions that question the population frequencies do so because of express doubts about the Hardy-Weinberg equilibrium assumptions.[108] To some courts this is the sole debatable link in the chain of reasoning that produces "genotype" frequency estimates.[109] In reply, the FBI and other

---

population in Hardy-Weinberg equilibrium, the first situation occurs in a fraction $p_i p_j$ of the population, as does the second. Therefore, the proportion which is *either* (i,j) or (j,i) is $2p_i p_j$. If a DNA sample gives rise to only one "allele," i, it may be because the person inherited the same "allele" from both parents. Homozygosity, as this is called, can happen only one way—(i,i)—so its relative frequency is p_i^2. However, it is also possible that the person is heterozygous but has a "null allele" that escapes detection (*e.g.*, because it is very small and runs off the gel) or has alleles that "coalesce" on the autoradiograph because the two alleles are very close together. *See* E.M. Steinberger et al., *On the Use of Excess Homozygosity for Subpopulation Detection*, 52 AM. J. HUM. GENETICS 1275 (1993). When loci show only one "allele," the FBI uses the figure $2p_i$, which overstates the "genotype" frequency under either scenario; Lifecodes uses $2p_i^2$ for enzymes that rarely if ever produce null alleles and $2p_i$ otherwise.

106. *See* Caldwell v. State, 393 S.E.2d 436, 443 (Ga. 1990) (testimony of Jung Choi that Lifecodes's database indicated a population that was not in Hardy-Weinberg equilibrium); Joel E. Cohen, *DNA Fingerprinting for Forensic Identification: Potential Effects on Data Interpretation of Subpopulation Heterogeneity and Band Number Variability*, 46 AM. J. HUM. GENETICS 358 (1990); Eric S. Lander, *DNA Fingerprinting on Trial*, 339 NATURE 501, 504 (1989) (editorial noting "spectacular deviations from Hardy-Weinberg equilibrium" in Lifecodes's data, indicating "genetically distinct subgroups within the Hispanic sample").

107. *See, e.g.*, State v. Bible, 858 P.2d 1152 (Ariz. 1993) (misreading an expert's explanation of excess homozygosity as a concession that her calculation of the population frequency was not based on a generally accepted method); State v. Pennell, 584 A.2d 513 (Del. Super. Ct. 1989) (fact of Cellmark's match in serial murder case admissible under reasonable reliance test, but match probability of 1/180,000,000,000 inadmissible due to excess homozygosity indicating lack of Hardy-Weinberg equilibrium, and questionable binning procedures). Oddly, the claim of "statistically significant deviations from Hardy-Weinberg equilibrium" continues to impress courts even though, as indicated below, the scientific debate has ceased. *See, e.g.*, State v. Cauthron, 846 P.2d 502, 515 (Wash. 1993).

108. *See, e.g.*, *Bible*, 858 P.2d 1152; Prater v. State, 820 S.W.2d 429, 438-39 (Ark. 1991); *Caldwell*, 393 S.E.2d 436.

109. *See, e.g.*, State v. Wilson, 817 P.2d 1136 (Kan. App. 1991) (opinion designated not for publication) (where FBI agent "testified at trial the formula used in calculating the frequency of a particular DNA band was based on standard probability theory and derived from the Hardy-Weinberg equilibrium, which has been modified to compensate

scientists argued that excess homozygosity resulted from an imperfect measuring process rather than a disequilibrium.[110] In the ensuing technical debate,[111] critics retreated from the claim of excess homozygosity to the weaker claim that the statistical tests are not powerful enough to disprove the hypothesized lack of independence.[112] This fallback position has grown increasingly untenable as more studies report substantial equilibrium at most loci.[113]

The real issue, however, is not "statistical significance"[114] but rather

---

for limitations in forensic DNA profiling," and defense expert, a professor of biology at Kansas State University, testified that sample population was not in Hardy-Weinberg equilibrium, the conflicting testimony did "not conclusively show those results are unreliable, and disagreement goes "only to the weight of the test results."); Mandujano v. State, 799 S.W.2d 318 (Tex. Ct. App. 1990). For a particularly garbled account, see Martinez v. State, 549 So.2d 694, 695 (Fla. App. 1989) (describing "the Hardy-Weinberg equilibria" as "an established statistical data base" and then as a "formula").

110. *See* Bernard Devlin, *Neil Risch & Kathryn Roeder, No Excess of Homozygosity at Loci Used for DNA Fingerprinting*, 249 SCIENCE 1416 (1990). The FBI attributes the excess of apparent homozygotes in its database to "technical problems which sometimes show only one band exhibited from a heterozygote [as when] a band is so small it runs right off the gel, or 2 bands occur so close to one another so as to appear as a single band." People v. Mohit, 579 N.Y.S.2d 990, 996 (Sup. Ct. 1992).

111. *See* Joel E. Cohen et al., *Forensic DNA Tests and Hardy-Weinberg Equilibrium*, 253 SCIENCE 1037 (1991); Philip Green & Eric S. Lander, *Forensic DNA Tests and Hardy-Weinberg Equilibrium*, 252 SCIENCE 1038 (1991); Bernard Devlin et al., *Forensic DNA Tests and Hardy-Weinberg Equilibrium*, 252 SCIENCE 1039 (1991). *See also* Ranajit Chakraborty et al., *Apparent Heterozygote Deficiencies Observed in DNA Typing Data and Their Implications in Forensic Applications*, 56 ANN. HUM. GENETICS 45 (1992); Ranajit Chakraborty, *Statistical Interpretation of DNA Typing Data*, 49 AM. J. HUM. GENETICS 895 (1991); Ranajit Chakraborty & L. Jin, *Heterozygote Deficiency, Population Substructure and Their Implications in DNA Fingerprinting*, 88 HUM. GENETICS 267 (1992); Bernard Devlin & Neil Risch, *A Note on Hardy-Weinberg Equilibrium of VNTR Data by Using the Federal Bureau of Investigation's Fixed-Bin Method*, 51 AM. J. HUM. GENETICS 549 (1992); Bruce S. Weir, *Independence of VNTR Alleles Defined as Floating Bins*, 51 AM. J. HUM. GENETICS 992 (1992) (all defending reliance on Hardy-Weinberg equilibrium).

112. *See* Eric S. Lander, *Reply*, 49 AM. J. HUM. GENETICS 899, 900 (1991) ("Critics . . . reply that such tests have insufficient power to detect deviations [from Hardy-Weinberg equilibrium] if present"); *cf.* Seymour Geisser & Wesley Johnson, *Testing Hardy-Weinberg Equilibrium on Allelic Data from VNTR Loci*, 51 AM. J. HUM. GENETICS 1084 (1992) (proposing other statistical procedures for assessing independence).

113. For the view that statistical tests with substantial power demonstrate the independence of VNTR alleles, see, for example, Bruce Budowle et al., *Reliability of Statistical Estimates in Forensic DNA Typing*, *in* DNA ON TRIAL: GENETIC IDENTIFICATION AND CRIMINAL JUSTICE 79, 87-88 (Paul R. Billings ed., 1992) (summarizing studies); Chakraborty, *supra* note 90, at 155-56; Bruce S. Weir, *Independence of VNTR Alleles Defined as Fixed Bins*, 130 GENETICS 873 (1992); Devlin et al., *supra* note 69; Kathryn Roeder, *DNA Fingerprinting: A Review of the Controversy* § 3.2.1 (1993) (unpublished manuscript, on file with the author) (summarizing studies).

114. An observed discrepancy would be "statistically significant" if the probability of observing so large a departure from equilibrium when, in reality, the population is in

practical or substantive significance. While a small or moderate departure from equilibrium may not be detectable in the existing data, it may, at the same time, make no meaningful difference to the match-binning frequencies.[115] In this regard, simulation studies indicating that any departure from independence has minor effects on match-binning frequency estimates support simple multiplication of "allele" frequencies to find the "genotype" frequencies at each locus.[116]

After estimating "allele" and "genotype" frequencies, denoted be P_l at each locus l, the final step in the independence method requires combining the various P_l's. This procedure generates the relative frequency of the total "genotype" $G = \{g_1, g_2, g_3, g_4, g_5\}$ for a match at five loci. If "linkage equilibrium," a situation in which there is no correlation between "genotypes" at different loci, arises, the frequency of the pattern for all the loci resembles the outcome of a series of coin flips. It is the product of the frequencies at each locus:

$$P = P_1 P_2 \ldots P_5 \qquad (2).$$

2. *The Population Structure Objection*

The most powerful criticism of this simple calculation concerns the population structure[117]—the presence of subgroups with varying DNA

---

equilibrium, falls below some threshold, like 0.05.

115. For this reason, it can be misleading to insist that "the product rule . . . can only be applied when the pairs of alleles are statistically independent," Geisser & Johnson, *supra* note 112, at 1084; or that "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." NRC REPORT, *supra* note 15, at 79.

116. *See* Devlin & Risch, *supra* note 89; Evett et al., *supra* note 67, at 502 (simulations of 1.2 million false accusations for all pairs of three probes in Caucasian database, like previous experiments with other databases and probes, showed that "the assumption of pairwise independence between probes has no unacceptable practical effects"); Ian W. Evett, *DNA Statistics: Putting the Problem into Perspective*, 33 JURIMETRICS J. 139 (1992) (exhaustive pairing of 1500 Caucasians tested at three loci to generate over a million between-person comparisons to simulate cases of false accusations and estimating frequencies of matching profiles assuming independence gave nine false matches, most of which had match frequencies larger than 1/100); Ian W. Evett & R. Pinchin, *DNA Single-Locus Profiles: Tests for Robustness of Statistical Procedures Within the Context of Forensic Science*, 104 INT'L J. L. & MED. 267 (1991); Bruce S. Weir, *Independence of VNTR Alleles Defined as Floating Bins*, 51 AM. J. HUM. GENETICS 992 (1992); *cf.* Berry et al., *supra* note 45 (bands are independent at one locus, but measurement errors are correlated). *But see* Weir, *supra* note 113, at 886 (disequilibrium found for some bins at some loci for single-fragment measurements in FBI Hispanic and Black databases).

117. *See* United States v. Jakobetz, 747 F. Supp. 250, 263 (D. Vt. 1990)

patterns that tend to mate among themselves.[118] This structure contradicts the assumptions that guarantee independence of alleles at a specific locus (Hardy-Weinberg equilibrium),[119] and casts a doubt on the validity of multiplying "genotype" frequencies across loci. Depending on the intercorrelations of "alleles" within subgroups, the full "genotype" frequency P in the broad population may be higher, lower, or even the same as $P_1P_2P_3P_4P_5$. Likewise, depending on the details of the population structure, the multilocus "genotype" frequency in a particular subpopulation may be higher, lower, or the same as $P_1P_2P_3P_4P_5$.

But how much higher or lower? At present, it is doubtful that population structure makes much of a difference. Of course, the independence assumptions do not hold *rigorously*. Few assumptions in applied mathematics or statistics do. Since almost all scientific work proceeds on the basis of simplifying assumptions, the question is whether the simplification produces satisfactory approximations. Thus, some medical geneticists argue, often on the basis of impression, that the degree of population structure is modest[120] and that overestimates at some loci are likely to be countered by underestimates at others, so that the use of the final product will not systematically disadvantage defendants.[121]

("[S]ubstructure is arguably the weakest link of the DNA profiling chain."), *aff'd*, 955 F.2d 786 (2d Cir. 1992), *cert. denied*, 113 U.S. 104 (1992); *see also* NRC REPORT, *supra* note 15, at 79 ("[W]hether actual populations have significant substructure for the loci used" is "the key question."). *But see* Peter Donnelly, *Discussion of Paper by Berry, Evett & Pinchin*, 41 APPLIED STAT. 521, 524-25 (1992) (presenting a theoretical basis other than population structure to suspect difficult to detect correlations across loci).

118. Lewontin & Hartl, *supra* note 81; Cohen, *supra* note 111.

119. Despite the representations of some experts testifying in support of FBI findings of matches, the mere fact that people do not consider the VNTRs of their sexual partners does not satisfy the "random mating" assumptions behind Hardy-Weinberg equilibrium. *See supra* note 103. In *Jakobetz*, 747 F. Supp.at 260, for instance, the district court observed that "Dr. Lewontin did discredit the government's experts who casually concluded that VNTRs must randomly occur throughout the population because individuals do not consciously consider VNTRs when they choose their mates." Even after this rebuke, however, the FBI appears to have continued to advance this simplistic argument. *See, e.g.*, People v. Mohit, 579 N.Y.S.2d 990, 996 (Sup. Ct. 1992) ("The FBI, in supporting its claim of Hardy-Weinberg, argues that mating is random since individuals are not aware of their partner's VNTR patterns."). However, unless some other characteristic affecting mating within subgroups is correlated with VNTRs, the observation implies that each subpopulation is in equilibrium.

120. *See, e.g.*, United States v. Yee, 134 F.R.D. 161, 185-87 (N.D. Ohio 1991).

121. *See, Id.* at 187 (testimony of Stephen Diager and Kenneth Kidd); *Mohit*, 579 N.Y.S.2d at 997 (testimony of Michael Conneally). Support for these impressions may be found in Weir, *supra* note 113, which presents four-locus "genotype" frequencies for VNTR patterns computed according to "allele" frequencies in all possible pairs of Black, Caucasian, Florida Hispanic and Texas Hispanic databases, and concludes:

Moreover, studies of the frequency of matching "alleles" for large numbers of pairs of different people in laboratory databases seem to show no false matches across four or five loci and rates of matches on subsets of loci that do not depart markedly from the expected values given independence of "alleles" across loci.[122]

Accumulating evidence supports the independence of the VNTR loci.[123] Yet, because of the lack of dramatic differences in the frequencies of VNTR alleles across ethnic subpopulations,[124] and because of the small differences attributable to using an inapposite racial database in

---

> Although different bin frequencies lead to different four-locus estimated frequencies, the differences are rarely more than two orders of magnitude, and generally less than one order of magnitude. It is as though frequency differences tend to cancel each other—some fragments are more frequent in one database while others are less frequent.

Id. at 885.

122. *See* George Herrin, Jr., *Probability of Matching RFLP Patterns from Unrelated Individuals*, 52 AM. J. HUM. GENETICS 491 (1993); Neil J. Risch & Bernard Devlin, *DNA Fingerprint Matches*, 256 SCIENCE 1744 (1992); Risch & Devlin, *supra* note 45. From this analysis, these biostatisticians conclude that "[t]he observed independence of matching among loci, both in the FBI and Lifecodes data sets, provides no support for claims of linkage disequilibrium within ethnic groups. Indeed, if linkage disequilibrium among loci does exist, it has little effect on the probability of two random individuals having matching genotypes." *Id.* at 719. *See also* Weir, *supra* note 113, at 997 ("By randomly generating many profiles, however, this study has demonstrated that . . . whatever levels of dependence do exist are unlikely to have a meaningful impact on forensic calculations.").

123. *See, e.g.*, Devlin & Risch, *supra* note 89; Berry et al., *supra* note 45. *But cf.* Weir, *supra* note 113, at 886 (some two-locus associations found at .05 but not .01 significance level for some single-fragment patterns in FBI database for Blacks, but even these associations disappeared when only double heterozygotes were considered).

124. *See, e.g.*, Devlin & Risch, *supra* note 89. It is also suggested that genotype frequencies differ more among groups than within ethnic groups. *See* Ranajit Chakraborty, *NRC Report on DNA Typing*, 260 SCIENCE 1059 (1993) (letter insisting that "the extent of regional difference within a racial group is far less than that between races" and that "analysis of hypervariable DNA loci [demonstrate that] the mean kinship within race is 0.4%" which is "less by an order of magnitude . . . than for blood groups and isoenzymes"); Bernard Devlin et al., *NRC Report on DNA Typing*, 260 SCIENCE 1057 (1993) (letter asserting that "the estimate of diversity based on variance of allele frequencies among subpopulations is usually quite small—approximately 0.1%"); Bernard Devlin et al., *Statistical Evaluation of DNA Fingerprinting: A Critique of the NRC's Report*, 259 SCIENCE 748, 837 (1993). *Contra* Daniel L. Hartl & Richard C. Lewontin, *DNA Fingerprinting Report*, 260 SCIENCE 473, 474 (1993) (letter asserting that "there is approximately as much genetic variation among ethnic groups within major races as there is among the races"). This last statement from Hartl and Lewontin seems to recognize that their original claim of substantially *more* genetic variation among ethnic subgroups within races than across the races was overstated. Lewontin & Hartl, *supra* note 81, at 1745 (paper typically cited in opinions holding that population structure is so serious or controversial a problem that big bin computations are inadmissible).

simulation studies of false matches,[125] the controversy over the implications of population structure for the independence method lingers on.[126] As one might expect, the debate is not easy for the courts to untangle.[127] In *People v. Pizarro*,[128] for instance, the California court of appeals quoted at length from various scientific publications and submissions, and lamented:

> The difficulty is, where does this place us? It places us in the middle of the conflict as to whether or not the basic theory of population genetics involving broad racial and ethnic groups as opposed to the argument of substructure has any general acceptance in the relevant scientific community—a conflict which we cannot resolve on the present record.

Neither does the NRC study settle the issue. To the contrary, it pointedly avoids it. Unable to agree that the population structure objection is valid for VNTRs, the panel simply "decided to assume that population substructure might exist" and to propound one particularly "conservative" method to respond to this hypothetical problem.[129]

---

125. *See, e.g.*, Ian W. Evett, *DNA Statistics: Putting the Problems into Perspective*, 33 JURIMETRICS J. 139 (1992) (using Afro-Caribbean instead of Caucasian database to estimate three-locus profile frequencies in one million simulated cases of false accusations raised the false match rate from 9 to 16 per million). Such studies demonstrate that the "potential error rate" associated with the independence method weighs in favor of admitting such computations. *See supra* note 30 (*Daubert* "considerations" for discerning "scientific knowledge").

126. *See* John Brookfield, *Law and Probabilities*, 355 NATURE 207 (1992); Ranajit Chakraborty & Kenneth K. Kidd, *The Utility of DNA Typing in Forensic Work*, 254 SCIENCE 1735 (1991); Christopher Wills, *Forensic DNA Typing*, 255 SCIENCE 1050 (1992); Richard A. Nichols & David J. Balding, *Effects of Population Structure on DNA Fingerprint Analysis in Forensic Science*, 66 HEREDITY 297, 301 (1991); Bruce S. Weir, *Discussion of the Paper by Berry, Evett & Pinchin*, 41 APPLIED STAT. 521, 528 (1992); *cf.* Donnelly, *supra* note 117. *Compare* Daniel L. Hartl & Richard C. Lewontin, *DNA Fingerprinting Report*, 260 SCIENCE 473 (1993) *with* B. Devlin et al., *NRC Report on DNA Typing*, 260 SCIENCE 1057 (1993) (exchange of letters on genetic variability within ethnic groups of racial populations as opposed to the variability across populations, and on the interpretation of Dan E. Krane et al., *Genetic Differences at Four DNA Typing Loci in Finnish, Italian, and Mixed Caucasian Populations*, 89 PROC. NAT'L ACAD. SCI. 10583, 10585 (1992), discussed *infra* note 201, on allele frequencies within certain ethnic groups).

127. The tendency of courts to cite the opinions of other courts rather than scientists for scientific propositions and to lag behind the rapidly accumulating scientific literature combine to exascerbate the problem.

128. 12 Cal. Rptr. 2d 436, 456 (Ct. App. 1992).

129. NRC REPORT, *supra* note 15, at 94. *See also id.* at 80 ("the committee has

Even before the NRC committee spoke, "conservative"[130] alternatives for computing the match-binning "genotype" frequency *P* had been advanced—and implemented—to counter the population structure concern. With the NRC's recommendations for even greater caution, the pressure to overestimate population frequencies has increased. The opinion of the Supreme Judicial Court of Massachusetts, in *Commonwealth v. Lanigan*,[131] illustrates how compelling the calls for caution can be. In disposing of matches with population frequencies on the order of one in several million, the Massachusetts court explained that "[t]he national call for considered, conservative approaches to DNA testing, such as the use of ceiling frequencies, and the absence of such an approach in the present cases, underscore the wisdom of the motion judge in excluding the test evidence."[132]

However, it may not be so wise to compel the experts to bend over backwards in computing population frequencies. A more accurate estimate of the interval in which the true frequency lies may be of more assistance to the jury, or a somewhat different, but still conservative procedure, may be superior to the NRC committee proposal. For example, a series of recent papers show how to incorporate existing information on population substructure into estimates of population frequencies.[133] Before courts or legislatures decide on which methodology

---

chosen to assume for the sake of discussion that population substructure may exist and provide a method for estimating population frequencies in a manner that adequately accounts for it.").

130. A "conservative" estimate of an allele frequency is an estimate that is too large, and hence biased in favor of defendant. *See, e.g.*, Commonwealth v. Lanigan, 596 N.E.2d 311, 316 (Mass. 1992).

131. *Id.*

132. *Id.* at 316. *See also* State v. Cauthron, 846 P.2d 502, 517 (Wash. 1993) (remanding for a determination of whether "the empirical evidence utilized by Cellmark is valid under the criteria set forth by the [NRC] Committee").

133. *See, e.g.*, David Balding & Richard A. Nichols, *DNA Profile Match Probability Calculation: How to Allow for Population Stratification, Relatedness, Database Selection and Single Bands* (Mar. 24, 1993) (unpublished manuscript, on file with the author) (proposing a procedure that estimates genotype frequency within a subpopulation or among relatives using measures of interpopulation variation in allele frequency, said to be superior to the "complicated, ad hoc and overly-conservative" ceiling principle); A.W. Sudbury & J. Martinopoulos, *Assessing the Evidential Value of DNA Profiles Matching Without Using the Assumption of Independent Loci*, 33 J. FORENSIC SCI. SOC'Y 73 (1993); James F. Crow & Carter Denniston, *Population Genetics as It Relates to Human Identification*, PROC. FOR THE FOURTH INT'L SYMP. ON HUM. IDENTIFICATION (forthcoming 1994) (describing a procedure that incorporates existing data on population structure into computations of the reference population frequency); Bruce S. Weir, *Conditional Genotypic Frequencies in Forensic Analysis*, Paper presented at the Nat'l Institute of Statistical Sciences Forum on DNA Fingerprinting (Oct. 21, 1993) (describ-

to accept, they should consider the full gamut of conservative approaches and their relation to the basic independence method. With the essential features of these "overestimation" methods elucidated, we shall return to the population structure objection to match-binning frequencies.

3. *Overestimation Methods*

Independence with Big Bins. The FBI and other proponents of the independence method have responded to criticism of the equilibria assumptions with a form of confession and avoidance. They concede that, strictly speaking, the assumptions do not hold, but they argue that any plausible underestimate of the genotype frequency is avoided by the intentionally "conservative" aspects of match-binning as practiced by the FBI. Such practices include using big bins relative to match windows, combining bins so that none contain less than 5% of the alleles in the population, and treating a fragment near a fixed bin boundary as if it falls within the larger bin.[134] Many courts have accepted the assurances that these adjustments are more than generous and have held genotype frequencies obtained with the big bin variation of the independence method as admissible.[135]

Guessing. A few courts have demanded more. In *People v. Mohit*,[136] an Iranian-born physician was indicted for rape and sexual abuse of a patient during an office examination. FBI testing revealed a match between the crime sample (a vaginal swab) and a sample of Dr. Mohit's

---

ing another procedure to incorpate data on population structure and the possibility that the suspect and the perpetrator belong to the same subpopulation).

134. *See, e.g.*, Springfield v. State, 860 P.2d 435 (Wyo. 1993) (testimony of Michael Conneally that "the conservative approach of binning . . . more than makes up for any small differences there might be"); Bruce Budowle et al., *Fixed Bin Analysis for Statistical Evaluation of Continuous Distributions of Allelic Data from VNTR Loci, for Use in Forensic Comparisons*, 48 AM. J. HUM. GENETICS 841 (1991); Ranajit Chakraborty, *Statistical Interpretation of DNA Typing Data*, 49 AM. J. HUM. GENETICS 895, 896 (1991) (letter); *supra* note 105 (use of $2p$ rather than p^2 in cases of apparent homozygosity); Newton E. Morton, *Genetic Structure of Forensic Populations*, 89 PROC. NAT'L ACAD. SCI. 2256 (1992). The NRC Report favors combining the two adjacent, fixed bins when a fragment lies near the boundary between them. Why this is preferable to selecting the bin that has the larger frequency is mysterious. *See, e.g.*, Monson & Budowle, *supra* note 97, at 1043; *supra* note 96.

135. *See, e.g.*, United States v. Jakobetz, 747 F.Supp. 250, 259-61 (D. Vt. 1990), *aff'd*, 955 F.2d 786 (2d Cir. 1992), *cert. denied*, 113 S.Ct. 104 (1992); United States v. Yee, 134 F.R.D. 161, 182, 210 (N.D. Ohio 1991); State v. Futch, 860 P.2d 264 (Or. Ct. App. 1993). *Contra, e.g.*, Commonwealth v. Lanigan, 596 N.E. 2d 311 (Mass. 1992).

136. 579 N.Y.S.2d 990 (Sup. Ct. 1992).

blood. The FBI reported that "the probability of such a match occurring in the United States was 1 in 67,000,000 for Caucasians, 1 in 79,000,000 for Blacks, and 1 in 14,000,000 for Hispanics." While the conservative features of the FBI's binning did not satisfy the trial court, the court was not prepared to exclude a sufficiently conservative estimate of the genotype frequency. Thus the court proceeded to press the government's expert, a medical geneticist, for "the most conservative possible estimate conceivable."[137] The court then held the figure of 1/100,000 supplied by the geneticist[138] to be admissible instead of the 1/67,000,000 obtained via the independence method with big bins.

As a general matter, pressuring experts to raise their estimates until it is believed that the genotype frequency cannot go any lower lacks a certain elegance. Other seat-of-the-pants judgments by experts are not much better.[139]

Independence with Ceilings. Big bins and other ad hoc adjustments are disquieting. In the absence of direct studies of the variance of VNTR "alleles" and "genotypes" across subgroups of broad racial and ethnic populations and its effect on estimated match frequencies, it is impossible, a few scientists say,[140] to know whether the overestimation of "allele"

---

137. *Id.* at 999.

138. The precise basis for the estimate of 1/100,000 is not clear from the opinion. Michael Conneally, on whom the court relied, had opined that "the highest degree of dependence between genotypes on separate chromosomes could not possibly exceed 10%." *Id.* at 998. "Factoring the 10% correlation into the multiplication of the 4 genotype frequencies, Conneally came up with . . . 1 in 22 million." *Id.* He provided the 1/100,000 figure when the court demanded that he be still more conservative.

139. *Compare* People v. Shi Fu Huang, 546 N.Y.S.2d 920 (Crim. Ct. 1989) *with Mohit*, 579 N.Y.S.2d 990. In *Shi Fu Huang*, no objection to the independence method (with small bins) was raised, but defendant did question Lifecodes's use of a database of under 200 university students from mainland China. Dr. Michael Baird of Lifecodes reported that, given the size of the sample, "the range statistically could be between one in two and a half billion to one in several trillion." *Id.* at 922. The court ruled that it would admit "the lowest figure of probability, namely one billion to one." *Id.* How Lifecodes arrived at its interval estimate, and how the court managed to select a figure below that interval, are not disclosed. Similarly, in People v. Wesley, 533 N.Y.S.2d 643, 658 (Crim. Ct. 1988), *aff'd*, 589 N.Y.S.2d 197 (App. Div. 1992), Dr. Kenneth Kidd testified that "an examination of the data given by Lifecodes indicated that there was, in fact, no linkage disequilibrium" and that he "found no marked deviation from the expected" genotype frequencies at loci under Hardy-Weinberg equilibrium, but that slight deviations from Hardy-Weinberg equilibrium warranted reducing "mean power of identity" by "much less than a factor of 10." The court then limited the prosecution to estimates reduced by a factor of ten to 1 in 1.4 billion for American Blacks and 1 in 840,000,000 for Caucasoids.

140. *See, e.g.*, United States v. Yee, 134 F.R.D. 161, 183-84 (N.D. Ohio 1991) (testimony of Eric Lander); Eric S. Lander, 49 AM J. HUM GENETICS 899 (1991) (letter). Other scientists maintain that ample information already warrants the conclusion

frequencies introduced at the binning stage[141] or elsewhere[142] overcompensates or undercompensates for possible underestimation (with respect to a particular subpopulation) of the single-locus frequencies given by equation (1) or their product (2). Persuaded by this limitation in the ad hoc adjustments to the independence method, the NRC panel endorsed yet another variation on the independence method—the "ceiling principle."[143]

The ceiling principle, in its simplest form, capitalizes on studies of "allele" frequencies among subgroups. Once random samples of DNA from more or less homogeneous ethnic subgroups, such as "English, Germans, Italians, Russians, Navahos, Puerto Ricans, Chinese, Japanese, Vietnamese, and West Africans,"[144] are collected, the highest frequency, p_i^{max}, for each "allele" i in the crime sample, with respect to all of the subgroups, is selected. These frequencies are then multiplied as in the independence method to produce "genotype" frequencies. Since the largest frequency for *any* ethnic subgroup studied has been used at each locus, and no single ethnic subgroup has the maximum frequency at *every* locus, it would seem that the result must overstate the "genotype" frequency both within each ethnic subgroup and within each broader group composed of these subgroups. And, the estimated genotype frequency P that results from multiplying these ceiling frequencies p_i^{max} must be the same for every subgroup. Consequently, the committee insists that "the ceiling principle eliminates the need for investigating the perpetrator population because it yields an upper bound to the frequency that would be obtained by that approach."[145]

Unless every conceivable ethnic subgroup is studied, however, this simple formulation of the ceiling principle is not guaranteed to yield the upper bound.[146] It is possible (though at some point implausible) that

---

that underestimation is a remote possibility. *See infra* note 199.

141. *See supra* note 96.

142. *See supra* note 105.

143. NRC REPORT, *supra* note 15, at 82-85. For other commentary proposing this approach, see, for example, Lander, *supra* note 140; Lewontin & Hartl, *supra* note 81.

144. NRC REPORT, *supra* note 15, at 84.

145. *Id.* at 85. This upper bound is not the only, and certainly not the least upper bound, that avoids an inquiry into the population of plausible suspects. A lower ceiling, consisting of the maximum "genotype" frequency in any of the subpopulations, also could be chosen. *See* Weir, *supra* note 91. If computations with the NRC committee's "ceiling principle" are admissible, then, arguably, such refinements also should be.

146. In truth, the ceiling method is not guaranteed to produce an upper bound even when every subgroup for which there are frequency variations is considered. The method yields an upper bound only if the alleles occur independently at all loci in each subpopulation. If Hardy-Weinberg equilibrium does not exist at a locus, or if linkage

another, as yet unstudied subpopulation, has an "allele" frequency above the ceiling seen so far. To cope with this contingency, the NRC panel would require a minimum ceiling frequency of 5%, no matter how low all the subgroup frequencies are.[147]

The NRC committee's quest for a suitably conservative procedure does not stop at the imposition of the 5% lower upper bound. Until subgroup studies are complete, the panel calls for still higher ceilings. It recommends that each "allele" frequency be taken to be the higher of either 10% or the "upper 95% confidence limit" of the frequency seen in the major "race" with the largest frequency.[148]

---

equilibrium is absent across loci, then the ceiling method can understate the genotype frequency. *See* Joel E. Cohen, *The Ceiling Principle is not Always Conservative in Assigning Genotype Frequencies for Forensic DNA Testing*, 51 AM. J. HUM. GENETICS 1165 (1992). However, the deviations from equilibrium must be extreme to have this effect, and no reasons have been advanced for thinking that such deviations exist. *See supra* note 116.

147. NRC REPORT, *supra* note 15, at 84. This 5% minimum on the ceiling imposes a lower limit of 1/400 on the estimated frequency at each locus and a limit of $1/400^n$ on a match at n loci. The panel justifies this 5% lower bound on the upper bounds p_i^{max} as follows:

> Because only a limited number of populations can be sampled, it is necessary to make some allowance for unexamined populations. As usual, the problem is rare alleles. Genetic drift has the greatest proportional effect on rare alleles and may cause substantial variation in their frequency. Even if one sees allele frequencies of 1% in several ethnic populations, it is not safe to conclude that the frequency might not be five-fold higher in some subgroups.

Id. In some ways, the committee's reasoning here is remarkable. Ordinarily, the fact that "it is not safe to conclude that [something] might not" is not a reason to act as if it actually will happen. If this principle were applied generally, juries, businesses, governments, and all other decision-makers would be paralyzed, since actions so often rest on premises that are at best tentatively true. To say that any number less than 5% is "not safe" is to express a policy judgment about tolerable risk, and such a judgment can be defended only by specifying the risk in question and the dangers involved. Although the committee writes that its selection of a lowest upper bound of 5% "was based on population genetic theory and computational results . . . aimed at accounting for the effects of sampling error and for genetic drift," *id.*, its report omits any explanation or description of this theoretical and computational analysis. This omission is troublesome because more than one model of the introduction of new alleles, and hence, genetic drift, in a population can be proposed. *See, e.g.*, Bernard Devlin & Neil Risch, *Ethnic Differentiation at VNTR Loci, With Special Reference to Forensic Applications*, 51 AM. J. HUM. GENETICS 534, 545 (1992). The most specific statement the committee does make is that "[e]ven if one observed allele frequencies of about 1%, one would guard against the possibility that the frequency in a subpopulation had drifted higher by using the lower bound of 5%." *Id.* at 84. As a result, the panel simply fails to explain how it struck the desired "balance [between] rigor and practicality." *Id.* at 83.

148. NRC REPORT, *supra* note 15, at 93. If the NRC committee's recommendations are uncritically adopted, the interim ceilings may well be permanent. The panel would

4. Population Structure and Overestimation

We have discussed four methods of inferring the "genotype" frequency P from allele frequencies p_i: the independence method with bin widths that correspond to match windows, the independence method with big bins, the method of guessing at upper bounds, and the independence method with ceiling frequencies. The three modifications of the first form of the independence method all suffer from the same weakness as the direct estimation procedure in that they are likely to produce overestimates of the match frequency in a broad ethnic or racial population. In theory, the smallest "genotype" frequency that the interim 10% ceiling procedure can generate for four probes is $1/(200)^4 = 1/1,600,000,000$, and this "one in a billion" figure should be small enough to delight most prosecutors and to convince most jurors that the match is no accident. However, in practice, genotype frequencies computed with "allele" ceilings will be larger than in theory.[149] For example, the "genotype" in *United States v. Yee*[150] had a frequency of 1/35,000 when adjusted upward with big bins. According to some reports, the ceilings expand this figure by a factor of 2,000, to yield the frequency of 1/17.[151]

Does the hypothetical possibility of substantial population structure

---

allow the 5% lower upper bound to replace the 10% ceiling only when "the population studies do not reveal significant substructure." *Id.* From existing data on Caucasians, Navajos and West Africans, it is clear that statistically significant differences in allele frequencies are present (although they do not seem to be "significant" in the practical sense of producing a large proportion of markedly different multilocus genotype frequencies). The same is probably true for English, Navajos and West Africans. If "significant substructure" means statistically significant substructure in the enumerated subpopulations, then the subpopulation studies are pointless and the prospect of dropping from the ten percent ceiling illusory. *See* Devlin et al., *supra* note 69.

149. Large genotype ceiling frequencies could come to be most common during the transition from the interim to the final ceiling method. If the samples of the ethnic subgroups are small, the upper 95% confidence limit of some "allele" frequencies in some subgroup easily could exceed even the interim minimum "allele" ceiling of 10%.

150. 134 F.R.D. 161 (N.D. Ohio 1991).

151. *See* Rorie Sherman, *New Scrutiny for DNA Testing*, NAT'L L.J., Oct. 18, 1993, at 3, 52; Richard C. Lewontin, *DNA Evidence: Statistical and Biological Considerations*, Invited Papers on Statistical Issues in DNA Identification Evidence, JOINT STATISTICAL MEETINGS OF THE AMERICAN STATISTICAL ASSOCIATION, BIOMETRIC SOCIETY AND INSTITUTE OF MATHEMATICAL STATISTICS (Aug. 10, 1992). However, this figure may be overstated. *Cf.* Weir & Gaut, *supra* note 47 (elucidating errors in one expert's inflated computation of the ceiling frequency in another case). Less dramatic differences probably are typical. *See, e.g.*, Springfield v. State, 860 P.2d 435 (Wyo. 1993) (using ceiling frequencies of 1/17,000,000 for blacks and 1/221,000 for Native Americans, as opposed to big bin frequencies of 1/150,000,000 and 1/250,000, respectively).

warrant requiring such overestimation? At least in jurisdictions that consider the scientific merit of scientific evidence,[152] the answer depends on how hypothetical the population structure argument is and whether the independence method allows an expert to present a reasonable estimate of the population frequency and to quantify the uncertainty in the figure. I shall argue that speculation about the extent of population structure notwithstanding, in many cases a suitable population frequency estimate is obtainable without resort to extreme overestimation. Furthermore, I will demonstrate that this proposition has not been widely nor directly disputed in the scientific literature.[153]

My analysis builds on a fundamental distinction between what I denote as a general population case and a subpopulation case. A general population case arises when the appropriate reference population is a broad ethnic or racial population, and a representative sample of "allele" frequencies for this general population is available. A subpopulation case arises when the appropriate reference population is itself a subpopulation (or a population or set of subpopulations not represented in the database). The distinction is important because the presence of substructure in general population cases can be expected to cause predominantly one type of error—an overestimate of the population "genotype" frequency—and only relatively small errors in most instances. As a result, the population structure objection does not justify a rule of law that demands drastic overestimation in these cases.

In applying this distinction, it is critical to understand the limited role that the defendant's ethnic or racial status plays in evaluating the evidence of a match. The choice of the reference population for any frequency estimate should be appropriate to the facts of the case. Is the pertinent frequency to be found from a sample drawn from the general population? From a particular geographic area? From people resembling or related to the defendant? These questions are neither new nor special to DNA evidence.[154] One simple principle supplies the answers: The relevant population consists of all people who might have been the source of the

152. *See supra* text accompanying notes 21-26.

153. This inquiry is particularly important in jurisdictions that treat controversy over a scientific procedure as an absolute barrier to admissibility, regardless of the validity of the procedure. *See supra* text and accompanying notes 21-26.

154. *See, e.g.*, David H. Kaye, *The Admissibility of "Probability Evidence" in Criminal Trials—Part II*, 27 JURIMETRICS J. 160 (1987); Bruce S. Weir & Ian W. Evett, *Reply to Lewontin*, 52 AM. J. HUM. GENETICS 206 (1993).

evidence sample.[155] In most cases, this will not be people with a defendant's peculiar ancestry, but people of many ethnic groups.

Yet, some courts have been impressed with arguments that the appropriate reference population necessarily consists of people like the defendant.[156] In *People v. Mohit*,[157] for example, the court was concerned that the race and ethnicity of the dentist accused of raping his patient was not represented in the FBI's database. It noted that:

> The issue of inbreeding is of particular importance in this case. The defendant, Dr. Mohit, was born in the Iranian town of Shushtar. His ancestors over at least the past five generations were of Persian descent, all from the same town or a town close by. They are all of the Shiite Muslim religion. Dr. Mohit claimed that for religious reasons, and as a matter of tradition, inbreeding was very common in his family. He indicated that his maternal grandmother was the daughter of his father's great-grandparents. Marriage among first cousins was common in his town.[158]

The issue, however, is not the frequency of matching DNA patterns for inbred families of Shiite muslims from Shustar, Iran, but their frequency in the vicinity of Westchester County, New York, or, more precisely,

---

155. Among commentators, agreement on this point is now virtually unanimous. *See* Donald A. Berry, *Statistical Issues in DNA Identification*, *in* DNA ON TRIAL: GENETIC IDENTIFICATION AND CRIMINAL JUSTICE 91, 106 (Paul R. Billings ed., 1992) ("The standard is to use the race of the suspect [but this] makes no sense."); Budowle et al., *supra* note 113, at 81-83; Ian W. Evett & Bruce S. Weir, *Flawed Reasoning in Court*, CHANCE, Fall 1991, at 19; Richard Lempert, *The Suspect Population and DNA Identification*, 34 JURIMETRICS J. 1 (1993); Lempert, *supra* note 74, at 310; Richard C. Lewontin, *Which Population?* 52 AM. J. HUM. GENETICS 206 (1993) ("[T]he ethnicity of the *defendant* is not the directly relevant question, but rather the ethnic composition of the pool of possible alternative suspects.").

156. *See* Evett and Weir, *supra* note 155; Bruce S. Weir & Ian W. Evett, *Whose DNA?* 50 AM. J. HUM. GENETICS 869 (1992) (letter recounting exclusion of DNA evidence in State v. Passino, No. 185-1-90 (Franklin Co. Dist. Ct. Vt. May 13, 1991), because homicide defendant had mixed racial heritage). However, Lewontin, *supra* note 155, indicates that the trial judge in *Passino* may not have made this mistake. *See also* Richard C. Lewontin, *The Dream of the Human Genome*, N.Y. REV. BOOKS, May 28, 1992, at 31. Although the *Passino* court's conclusion may be defensible, it seems clear that the court was unduly impressed with "the uncontroverted testimony" that "the defendant is one half Italian, three eights native American Indian and one eighth French."

157. 579 N.Y.S.2d 990 (Sup. Ct. 1992).

158. *Id.* at 997. The court even cited a study finding that marriage among relatives (second cousins or closer) in Muslim countries is 20-55%.

their frequency among people other than Dr. Mohit who might have left their semen on the patient. Unless this group consists largely of Dr. Mohit's relatives, there is no need to estimate the frequency among people of his racial and ethnic background. The frequency among broadly defined racial and ethnic groups is the apposite figure.

On the other hand, cases do arise where the population of interest is, arguably, a genetically distinct subpopulation, and where little or no data specific to that subpopulation have been collected. *United States v. Two Bulls*[159] may be such a case. Accused of raping a girl on the Pine Ridge Indian reservation in South Dakota, Matthew Two Bulls moved to suppress testimony of a match between DNA extracted from semen on her underwear and his DNA.[160] The FBI estimated the frequency of the matching pattern in "a Native American population base."[161] However, the appropriate reference population is not all Native Americans, but only the Oglala Sioux. If the FBI's "Native American" database is an amalgam of distinct subpopulations,[162] while the suspect population is dominated by one subpopulation, the frequency of matches in the FBI's database might be beside the point.

Population cases. Although courts are coming to appreciate that a defendant's ancestry is, at best, tangentially relevant to the choice of a reference population, the relationship between the reference population and the estimation procedure has yet to be recognized in any reported opinion. Even the NRC Report, commendably lucid and comprehensive in other areas, overlooks the possibility of adapting the computational method to the circumstances of the case.[163] Yet, a simple, numerical

---

159. 918 F.2d 56 (8th Cir. 1990), *vacated for reh'g en banc but appeal dismissed due to death of defendant*, 925 F.2d 1127 (8th Cir. 1991).

160. *Id.* at 61. The district court had held the evidence admissible, and Two Bulls entered a conditional plea of guilty under a plea agreement. The court of appeals set the plea aside and remanded the case to the trial court for "an expanded pre-trial hearing" to determine whether the method of DNA typing was generally accepted and performed properly, and whether the statistical evidence was unfairly prejudicial.

161. *Id.* at 57 n.2. The FBI obtained a frequency of 1/177,000 using the independence method with big bins.

162. *See* State v. Passino, No. 185-1-90 (Franklin Co. Dist. Ct. Vt. May 13, 1991) ("The FBI's Indian Database is made up of a variety of different tribes. Approximately half of them are Sioux Indians from the Northern Great Plains. Other tribes include the Cherokee, Arapaho, Zuni and Menominee."); *supra* note 156.

163. The closest that the report comes to acknowledging this approach is the following passage:

> Some legal commentators have pointed out that frequencies should properly be based on the population of possible perpetrators, rather than

example illustrates how the force of the population structure objection depends on the nature of the reference population. To put the example in a forensic context, suppose that there has been a violent robbery and rape at a rest stop on an interstate highway and that the robber and rapist, identified as a Caucasian, left traces of his blood or semen.[164] Suspicion focuses on a particular man. Careful DNA testing demonstrates that he matches at each locus. If, however, this suspect is not the assailant, then we can say only that someone else is. We have no reason to expect the guilty party to be of the suspect's detailed ancestry or ethnicity. Therefore, we are interested in the frequency of the matching "genotype" among all Caucasians who use interstate highways—and not the proportion in the defendant's subpopulation. When the case comes to trial, the prosecution offers an estimate of the frequency of this "genotype" in Caucasians in order to gauge the probative value of the evidence of the match. The prosecution's expert computes the frequency using the independence method with bin widths equal to match windows in a large national database on Caucasian Americans. The defense objects that the estimate is prejudicial because the population may be structured, so the actual frequency could be dramatically larger[165] than the figures computed by the prosecution's expert using equations (1) and (2).

To test the validity of the defense's objection, let us start with the simplest possible case of population substructure—one locus with only two "alleles" and one population composed of two genetically isolated subpopulations. Subpopulation 1 represents 80% of the population and subpopulation 2 represents 20%. The "allele" frequencies are presented in Table 1.

---

on the population to which a particular suspect belongs. Although that argument is formally correct, practicalities often preclude use of that approach.

NRC REPORT, *supra* note 15, at 85. The report nowhere identifies its practical objections to defining the reference population in the only logically acceptable manner imaginable, and its implication that it is generally appropriate to seek some estimate of the frequency in the suspect's ethnic subpopulation is plainly mistaken.

164. *Cf.* United States v. Jakobetz, 955 F.2d 786 (2d Cir. 1992), *cert. denied*, 113 S.Ct. 104 (1992); Edwin McDowell, *Threat of Crime Rises on The Main Highway*, N.Y. TIMES, Oct. 28, 1992, at A14.

165. It also could be considerably smaller, but the defense is unlikely to note this possibility.

Allele	Freq. in subpop.1 (80%)	Freq. in subpop.2 (20%)	Freq. in total population
1	3/5	1/5	(3/5)(80%) + (1/5)(20%) = 13/25
2	2/5	4/5	(2/5)(80%) + (4/5)(20%) = 12/25

Table 1.

Frequencies of two hypothetical alleles in a structured population.

This population structure implies that equilibrium does not exist for the broad population, but it does not impeach the equilibrium assumptions within each subpopulation. In the two subpopulations, equations (1) and (2) hold and can be used to deduce the "genotype" frequencies within these subpopulations, and hence, in the total population.[166] Table 2 presents these frequencies.

Genotype	Freq. in subpop. 1 (80%)	Freq. in subpop. 2 (20%)	Freq. in total population
1,2	2(3/5)(2/5)	2(1/5)(4/5)	(12/25)(80%)+(8/25)(20%)=280/625

Table 2.

Frequencies of one genotype in a structured population.

Of course, the prosecution expert did not know the subpopulation frequencies. Thus, the expert could use only the population allele frequencies 13/25 and 12/25. Using these values in (1) and (2) gives a calculated population genotype frequency of 2(13/25)(12/25) = 312/625. In this example, the population structure objection is not well-taken. While, the simple independence method is slightly inaccurate, reporting 312/625 instead of the true frequency of 280/625, the error favors the defendant.

This result is the consequence of a general mathematical truth rather than the consequence of a clever choice of numbers. As long as a population is composed of two isolated subgroups, each of which is in equilibrium, the frequency for a diallelic locus estimated by ignoring the

166. *See supra* note 146.

population structure overstates the true frequency.[167] As a result, the independence procedure is already conservative, and resort to ceilings is unjustified.[168]

Unfortunately, this example is not representative of more complex systems. With more loci or subpopulations, the multilocus frequencies estimated without considering population structure can overstate the true frequencies for populations.[169] Since the number of possible alleles in VNTR systems is typically 20 or more, and considerably more than two subpopulations may be present, an inequality that applies only to the case of two alleles and two subpopulations is of little use. Nevertheless, even in the more realistic situation, on average, the error due to population structure inures to the defendant's benefit,[170] and the differences between the computed and the true single-locus genotype frequencies will rarely be large.[171]

Partly because this point has not been recognized in the legal literature, the population structure objection has proved remarkably powerful in court. In *People v. Barney*,[172] for instance, a California court of appeals concluded that the NRC Report, a paper and a reporter's observations in *Science*, and conflicting testimony of experts in various other cases demonstrated the existence of an unsettled scientific controversy over population frequencies.[173] Most recently, in *State v. Bible*,[174] the

---

167. For a proof, see David H. Kaye, *The Effect of Population Structure on Estimated Allele Frequencies* (1993) (unpublished manuscript, on file with author).

168. The frequency that independence with ceilings produces in our example is $2(3/5)(4/5) = 600/625$ as compared to the correct value of 280/625. However, the degree of excessive overestimation inherent in the ceiling method will vary with the numbers used in such examples.

169. *See* C.C. Li, *Population Subdivision with Respect to Multiple Alleles*, 33 ANNALS HUM. GENETICS 23 (1969).

170. *See* Ranajit Chakraborty et al., *Effects of Population Subdivision and Allele Frequency Differences on Interpretation of DNA Typing Data for Human Identification*, 1992 PROC. THIRD INT'L SYMP. ON HUM. IDENTIFICATION 205; Kaye, *supra* note 167 (about 60% of computations of allele frequencies in randomly structured populations are overestimates).

171. In simulations of randomly structured populations, the maximum ratio of the true single locus genotype proportions P to the computed proportions P^* seen in simulated populations was less than three. Although conditions can be created that make P/P^* arbitrarily large (meaning that the true value is many times larger than the estimated value), preliminary study suggests that both P and P^* must be very close to zero for this to occur. *See* Kaye, *supra* note 167. If the *worst* effect of population structure is to cause the estimated proportion to be one in a billion when the true proportion is one in a million, the objection seems not to justify the exclusion of DNA evidence in all cases.

172. 10 Cal. Rptr. 2d 731 (Ct. App. 1992).

173. Justice Chin, who wrote the *Barney* opinion, adhered to this conclusion in the opinion for another panel in People v. Wallace, 17 Cal. Rptr. 2d 721 (Ct. App. 1993),

Arizona Supreme Court relied on the *Science* articles, news accounts, and *Barney* and other cases to find a "lack of general acceptance of Cellmark's statistical probability calculations." These cases demonstrate that the judicial perception that population substructure is a problem in all DNA cases is widespread.[175]

Nevertheless, the concern is largely misplaced when the pertinent frequency is in the general population. In these cases the population structure objection is far less vexing than many opinions and a few articles suggest. There is a corollary to this conclusion. The NRC panel's influential call for more conservative methods in these cases is an unnecessary response even to a hypothetical problem. Post-NRC Report cases excluding genotype frequency estimates on the ground that computational methods less conservative than the NRC's version of independence with ceilings are inadmissible should not be followed. Indeed, under the analysis developed in this article, most of the cases should have found the frequency estimates to be admissible because the circumstances of the offenses pointed to no specific subpopulation of suspects. In these cases, the relevant population in which to consider the frequency of the incriminating match is a general population, and existing computational methods work reasonably well for such populations.[176]

In *Commonwealth v. Lanigan*, for instance, the Supreme Judicial Court of Massachusetts consolidated cases against two sets of defendants. Thomas Lanigan was indicted for the rape of a child and for sexual assault and battery of three minors.[177] Presumably, these victims identified their assailant as a Caucasian, and not as a member of some

---

and chastised the scientific community for questioning the need for or the desirability of the ceiling approach.

174. 858 P.2d. 1152 (Ariz. 1993).

175. *See also* People v. Atoigue, DCA No. CR 91-95A (Guam Dist. Ct. App. Div. 1992) (following *Barney*); Commonwealth v. Lanigan, 596 N.E.2d 311 (Mass. 1992); State v. Vandebogart, 616 A.2d 483 (N.H. 1992) (relying on NRC Report to establish that FBI's calculation of population frequency of 1/50,000 is too controversial among population geneticists, and remanding for a hearing on the general acceptance of the NRC ceiling frequency); State v. Cauthron, 846 P.2d 502 (Wash. 1993) (relying on early paper by Eric Lander questioning equilibria assumptions).

176. *Cf.* Richard Lempert, *DNA, Science and the Law: Two Cheers for the Ceiling Principle*, 34 JURIMETRICS J. 41 (1993) ("[I]n most forensic situations the problem the ceiling principle was designed to resolve—the possibility that forensic data bases would be ignoring population substructure substantially underestimate relevant allele frequencies—hardly ever exists because the proper reference population for estimating allele frequencies is typically a mixed population fairly represented by the data bases now in use.").

177. 596 N.E.2d at 312.

subpopulation of that race. The reference population consists of all people who might have committed the acts. Unless the victims were largely isolated from the general population, the class of plausible suspects is Caucasians in general, and not some subpopulation to which Lanigan belongs. For this broad population, the independence method without ceilings is appropriate.[178]

The other actions in *Lanigan* named Leo Breadmore Senior and Junior as defendants. They were accused of raping, assaulting, and having incest with granddaughters and nieces. One alleged victim who delivered a child testified before a grand jury that she had sexual intercourse only with the defendants. DNA analysis of blood samples from the victim, her child and the Breadmores proved that the younger Breadmore could not be the father, and that the elder Breadmore had alleles that were "2,500 times more likely . . . if he were the father of the child than if he were not the father." Once again, unless the mother was largely isolated from the multi-ethnic population in Massachusetts, the class of plausible suspects is Caucasians, and not just the subpopulation to which the defendants belong.[179] Similarly, the two cases consolidated in *People v. Barney* evinced no circumstances suggesting some special subpopulation.[180]

---

178. The FBI reported that the genotype frequency among Caucasians was 1/2,400,-000. *Id.* at 312-13.

179. However, this population does include a number of people closely related to the defendants, which poses a special problem. *See infra* note 230.

180. 10 Cal Rptr. 2d 731 (Ct. App. 1992). In *People v. Howard*, Octavia Matthews was found on the floor of her home, bleeding from multiple head wounds, with a rope wrapped around her neck. She soon died. *Id.* at 732. Kevin Howard was her tenant in another building. *Id.* at 733. Ample circumstantial evidence pointed to him. She had served him with an eviction notice. *Id.* His fingerprints were on a postcard in her upstairs bedroom. His wallet was on her bloodstained couch. *Id.* There also were bloodstains on a tile floor, a paper napkin in a cosmetics case and a tissue in a purse. *Id.* Howard had a fresh cut on a finger when arrested, and conventional blood analysis showed that the stains and Howard's blood "shared an unusual blood type found in approximately 1.2 persons out of 1,000 in the Black population (and not at all in the White population)." *Id.* at 73. In these circumstances, the reference population does not seem to be any special subpopulation of African Americans, and it is reasonable to consider the fact, as reported by the FBI, that "Howard's DNA pattern matched . . . and the frequency of such a pattern is 1 in 200 million in the Black population." *Id.*

In *People v. Barney*, a woman entered her car in the South- Hayward BART parking lot. *Id.* A man forced his way in, demanded money, and used a knife to force her to drive and park some blocks away. *Id.* There he molested and tried to rape her, ejaculated on her clothing, and took about two dollars in small change, her BART ticket with $3.80 credit on it, and her car keys. *Id.* The woman found Ralph Edward Barney's wallet on the floor of her car and recognized Barney as her assailant from his photo ID in the wallet. *Id.* When arrested, he had a knife, a BART ticket last used to enter the transit system from the South Hayward station with the same amount of credit remaining on it to match the missing ticket, and $1.82 in small change. *Id.* Again, none

This treatment of the post-NRC Report cases, however, might be criticized for ignoring the doctrinal basis of the opinions. I have faulted these opinions for not recognizing that the population structure objection is weak when the relevant population in which to estimate the match-binning frequency is a collection of subpopulations having different VNTR frequencies. But the cases to date come from jurisdictions that, in theory,[181] neither ask nor allow their courts to decide what is scientifically valid or invalid, but only to ascertain whether the scientific community has reached the consensus that a scientific procedure rests on a valid theory and generates reliable results when properly applied. If

---

of the circumstances suggests any special subpopulation, and it seems reasonable to consider the frequency of the incriminating DNA profile among African Americans generally—reported by Cellmark Diagnostics to be "1 in 7.8 million in the Black population." *Id.* at 734.

Most of the other cases giving population structure as the reason to exclude frequency calculations were general population cases. In both People v. Wallace, 17 Cal. Rptr. 2d 721 (Ct. App. 1993), and State v. Cauthron, 846 P.2d 502 (Wash. 1993), police apprehended a man hiding in the bushes with material or implements of a type used in a series of unsolved, relatively distinctive rapes in the area. Nothing in the opinions suggests that any of the accounts of the victims or other circumstances pointed to membership in some well-defined ethnic subgroup as a characteristic of the rapist.

In *State v. Bible* a nine-year-old girl bicycling to a ranch in Flagstaff disappeared, and her battered body was found hidden in the woods three weeks later. The defendant was apprehended the day she disappeared, driving a stolen car whose contents matched items found near her body. Cellmark reported that DNA from blood stains on his shirt matched the girl's DNA, and estimated the genotype frequency in the Caucasian population to be between 1/60 million and 1/14 billion. If the defendant was not the person who abducted, molested and killed the girl, then someone else in the Flagstaff area that day did, and it is reasonable to consider the frequencies of the incriminating genotype among broadly defined groups in assessing the probative value of the match.

Because the New Hampshire Supreme Court's opinion in *State v. Vandebogart* does not specify the circumstances that motivated the FBI to use Caucasians as the reference population, it is difficult to say whether this case falls into the general population category for which the debate over substructure is essentially irrelevant.

The most awkward case for the general population approach may be *People v. Atoigue* As in *Lanigan*, *Atoigue* involved sexual intercourse with a child leading to pregnancy, and the reference population under the hypothesis that the man identified by the twelve-year-old mother as the father was not responsible is not a unique ethnic or racial subpopulation. To the extent that the distribution of alleles in the population of Chamorro, Guam, is markedly different from that in the database actually used, however, the frequency produced from the database would be inappropriate. The Polynesian Chamorrons have been living on Guam for perhaps 1,000 years, and a relatively small number of people founded the population there. If the Chamorrons have remained genetically isolated from the other inhabitants of the island, then even a database derived from the island as a whole may be off the mark. Apparently, Cellmark tried to address this concern by testing 15 unrelated male police officers in Chamorro, but the opinion does not describe the results of this ad hoc testing and the steps taken to ensure that the officers were unrelated.

181. For cases that show how far courts in *Frye* jurisdictions have diverged from this theory, see MCCORMICK, *supra* note 22, § 203, at 871-72.

leading population geneticists cannot agree on the validity of the independence assumptions,[182] and a blue ribbon committee that includes scientists among its members recommends the most extreme forms of overestimation, how can it be said that, without resorting to ceiling frequencies, equations (1) and (2) are generally accepted?

This criticism, however, treats the scientific dispute at too high a level of generality. No population geneticist or statistician denies that the relevant population in which to estimate a match-binning frequency consists of all the people who might have committed the crime.[183] At most, as in the NRC Report,[184] there are occasional slips in phraseology that make it seem like the particular suspect's subpopulation is necessarily relevant. However, in reality, the defendant's subpopulation is only derivatively relevant, to the extent that it conforms to the reference population of plausible suspects.[185] Neither does any population geneticist or statistician dispute the mathematical truism that equations (1) and (2), when used with allele frequencies for a structured population in which the independence assumptions hold within each subpopulation,[186] tend to overstate the frequencies of VNTR genotypes in that structured population.[187]

Unfortunately, the leading scientific papers advancing population structure as a reason to avoid the independence method in estimating match-binning frequencies do not explicitly analyze the effect of such

---

182. *But see* Anderson, *supra* note 19 (characterizing the debate as "on the scientific fringe" and the creation of "a few scientists . . . who proclaim themselves to be extremists").

183. I have found an exception to this generalization in one conversation with one eminent population geneticist who had not previously thought through the issues in forensic DNA testing. Certainly, the publications of scientists endorsing the view that the relevant population consists of people who might have committed the crime have not come under attack. *See supra* notes 155 & 156. And, the publications of population geneticists illustrate the population structure objection primarily in cases where the population of potential suspects is very probably a subpopulation of the type of people represented in the broad racial and ethnic databases. *See* Lander, *supra* note 106, at 505 ("[T]he crime occurred in a small, inbred Texas town founded by a handful of families."); Lewontin, *supra* note 156, at 68-69.

184. NRC REPORT, *supra* note 15, at 94.

185. *See* NRC REPORT, *supra* note 15, at 85 (It is correct to say that "frequencies should properly be based on the population of possible perpetrators, rather than on the population to which a particular suspect belongs."). *But see* Balding & Nichols, *supra* note 133.

186. For a direct analysis of the validity of this assumption, see Dan E. Krane et al., *Genetic Differences at Four DNA Typing Loci in Finnish, Italian, and Mixed Caucasian Populations*, 89 PROC. NAT'L ACAD. SCI. 10583, 10585 (1992).

187. *See* Chakraborty et al., *supra* note 170.

structure in general population cases. Rather, they emphasize the possibility of error when the reference population consists of individuals of the defendant's subpopulation.[188] To this extent, the controversy over the extent and impact of population structure, as developed in the scientific literature itself, does not compel a conclusion that the theory underlying multilocus genotype frequency estimates for broadly defined reference populations is not generally accepted.[189] Even in the dwindling number of jurisdictions where general acceptance is essential to the admissibility of scientific evidence,[190] the *Frye* standard does not demand the exclusion of the evidence in general population cases. In fact, the *Lanigan* opinion makes this clear. *Lanigan*, recognizing that the population structure argument concerns "the possibility that using allele frequencies of larger population groups might produce an inaccurate frequency estimate *for members of substructure groups*," excluded the frequency evidence.[191] However, since there was no reason to estimate the frequency for such subgroups in *Lanigan*, "the lively and still very current dispute"[192] that the court identified did not justify exclusion of the evidence.[193] The same is true of *Barney*,[194] *Bible*,[195] and *Wallace*.[196]

---

188. The leading criticism of independence (with match windows equal to bin widths or with big bins) is Lewontin & Hartl, *supra* note 81. These geneticists contend that substantial structuring for VNTR alleles may be present in populations, and they discuss the ratios in multilocus genotype frequencies for different subpopulations—but not the ratio between the frequency computed in the population given a knowledge of its structure and that computed without this knowledge. *Id.* at 1748 (tables 1 & 2). They express concern that "if the wrong ethnic group is used as the reference population, then a very low probability, even zero, may be assigned to a particular VNTR type, when the true probability may actually be relatively high in the proper ethnic group." *Id.* They conclude that "to be scientifically reliable, the databases must be expanded to include detailed knowledge of the VNTR frequency distributions in a wide variety of ethnic subgroups that are likely to be relevant in forensic applications." *Id.* at 1749. Their paper demonstrates that at least two eminent population geneticists have doubts about genotype frequencies derived from broad population data but then applied to cases where the reference population is but one subpopulation within that broader population. Although the paper concludes with the more sweeping claim that "estimates of the probability of a matching DNA profile based on VNTR data, as currently calculated, are unjustified and generally unreliable," *id.* at 1750, the analysis does not focus on the distinct question of the magnitude of the errors in using allele frequencies in a broad population when the reference population is that same broad population. *Cf. supra* note 180.

189. Plainly, the NRC panel's desire for a single method of calculating an upper bound on genotype frequencies in any likely population or subpopulation is not a pronouncement about science, but a mere preference for one jurisprudential policy over another.

190. *See supra* text accompanying note 26.

191. 596 N.E.2d at 315 (emphasis added).

192. *Id.* at 316.

193. The court relied also on the NRC Report's statement that "whether actual

Subpopulation cases. The recognition that the population structure objection is attenuated in the general population cases does not mean that there can never be a legitimate concern about population structure. To the contrary, the objection has real bite when the group of people who might have left the crime sample are a narrow and possibly insular subpopulation. In these subpopulation cases, the scientific question is how much variation in genotype frequencies exists across subpopulations. Two views prevail on the subject. The skeptical camp contends that the variations at specific loci might be huge, or they might be minuscule, but that it is impossible to determine without studies of the distributions of particular VNTR alleles across subpopulations.[197] The other camp maintains that subpopulations within ethnic groups rarely differ substantially as compared to variations across ethnic groups,[198] so that the disparities among subpopulations are not matters of pure speculation. According to this view, the state of scientific knowledge, including computations of match-binning frequencies in various populations and subpopulations,[199] suggest that "differences among subpopulations are of

---

populations have significant substructure for the [alleles] used for forensic typing . . . has provoked considerable debate among population geneticists." *Id.* As we have seen, however, the relevant debate for a general population case in a *Frye* jurisdiction is not just over the degree of population structuring, but over the impact of substructure on genotype frequency estimates in the broad reference population. Finally, the court thought that the NRC panel's willingness to proceed "on the assumption that population structure may exist" established a lack of acceptance in the scientific community of the independence method for estimating genotype frequencies in general population cases. *Id.* This is an obvious *non sequitur.*

194. 10 Cal. Rptr. 2d 731 (Ct. App. 1992). Partly on the basis of statements of one science journalist, *Barney* detects a "change in the attitude of the scientific community" occurring with the publication in *Science* of the Lewontin-Hartl paper and the publication of the NRC Report. *Id.* at 744. These authors, the court observes, "conclude that because the frequency of a given VNTR allele may differ among subgroups, reference to a broad data base may produce an inaccurate frequency estimate *for a defendant's subgroup.*" *Id.* at 740 (emphasis added). Since the defendant's subgroup was not the appropriate reference population in the cases in *Barney* (see *supra* text accompanying note 180), the opinion does not explain why this controversy made it improper to admit the frequency estimates in the general population. In addition, it seems odd—and extremely risky—to resolve questions of general acceptance by what science journalists say scientists have said to them rather than what scientists have written in professional journals or said to courts. *See supra* notes 5, 17 & 182.

195. 858 P.2d 1152 (Ariz. 1993).

196. People v. Wallace, 17 Cal. Rptr. 2d. 721 (Ct. App. 1993).

197. *See, e.g.*, Lewontin & Hartl, *supra* note 81.

198. The literature supporting this view is summarized in Devlin et al., *supra* note 69. *See also supra* note 124.

199. One indication that the true "genotype" frequencies are much smaller than the "corrections" for putative population structure made by the overestimation methods lies in studies of subgroup frequencies. If data on two subpopulations are available, the

questionable importance"[200] and have little impact on multilocus "genotype" frequencies.[201]

---

singlelocus and multilocus genotype frequencies validly can be estimated under the independence assumptions, since any possible structure in the general population is eliminated or reduced by focusing on the subpopulations. Next, a database mixing these subgroups can be constructed, simulating a highly structured population. Using this simulated database, one can estimate allele frequencies and compute genotype frequencies as if the population were homogeneous. If the artificial population frequencies are close to the true frequencies in the simulated population, one must conclude that even the exaggerated substructuring does not produce much error.

Although detailed data on fully homogeneous subpopulations are not yet available, analyses along these lines have been performed mixing southeastern and southwestern Hispanic-Americans, African-Americans, and mixing Caucasian-Americans, and Afro-Caribbeans, Asians, and Caucasians living in England. *See* Devlin & Risch, *supra* note 89, at 546; Berry et al., *supra* note 45; Evett & Pinchin, *supra* note 116, at 271 ("Even in the extreme case of using an Afro-Caribbean instead of a Caucasian database, the consequences are not serious . . . . It is now clear that the precise shapes of the bandweight frequency distributions are not particularly important."); Monson & Budowle, *supra* note 97, at 1044-49 (four-locus genotype frequencies derived by crossing African-American, Caucasian, southeastern Hispanic and southwestern Hispanic databases rarely differ by more than a factor of ten). Comparisons between Caucasian-Americans, African-Americans, Hispanic-Americans and Chinese-Singaporans, Malay-Singaporans, and Indian-Singaporans tell much the same story. Shui Tse Chow et al., *The Development of a DNA Profiling Database in a HAE III Based RFLP System for Chinese, Malays, and Indians in Singapore*, 38 J. FORENSIC SCI. 874 (1993).

200. Devlin & Risch, *supra* note 89, at 546.

201. *See* Newton E. Morton, *Genetic Structure of Forensic Populations*, 89 PROC. NAT'L ACAD. SCI. U.S. 2556, 2560 (1992) (Kinship studies show human populations to have little structure, making the ceiling approach "absurdly conservative."). Krane et al., *supra* note 126, defend the ceiling principle as having "a sufficient margin of safety." *Id.* at 10586. Krane and his coauthors analyze blood samples from 73 Finns in Helsinki, 79 Italians in Milan, and 1,354 Caucasians in St. Louis to find that allele frequencies do vary among these groups. To judge the impact on forensic calculations, they examine discrepancies obtained by switching the databases for Finns and Italians (i.e., computing three-locus frequencies for Finns using allele frequencies for Italians and vice versa). Although intriguing, this analysis does not fully simulate the forensic practice. In court, more loci are used, reducing the probability that the frequencies estimated at every locus will be too low. In addition, the forensic databases reflect more heterogeneous populations, like Caucasians, so that the divergence in allele frequencies between them and their subpopulations are likely to be less than the disparities in frequencies between the alleles in two subpopulations. Indeed, when the researchers computed three-locus profile frequencies for Finns and Italians with allele frequencies appropriate to the St. Louis Caucasians, the disparities were somewhat less dramatic. Most (78%) of these profile frequencies are off by less than a factor of ten, and virtually all are within a factor of 100 of the correct values for Finns and Italians (which typically are on the order of 10^{-6} or less). *Id.* at 10586 (Figure 2). These findings thus suggest that the independence assumption with big bins produces genotype frequencies that are roughly correct even when applied to a subpopulation. Indeed, these numbers probably understate the accuracy of the independence assumptions with big bins. When the databases are switched, the individual whose genotype frequency is to be estimated is left in the cognate database, which elevates the frequency of this genotype in that database. *See* Chakraborty, *supra* note 124 ("inherent statistical artifact"); Bernard Devlin & Neil Risch, *NRC Report on DNA Typing*, 260 SCIENCE 1057, 1058 (1993) (letter pointing out "large upward bias" in Krane et al. for samples that included only 29 Finns and 70

If the we-know-enough camp is correct, then the overestimation procedures should not be required. They fall short of providing numbers that convey the evidentiary value of a match. The uncertainty in estimates of the genotype frequency P might be exposed in any number of ways. The overestimates—from big bins to ceilings—might be presented along with less extravagant estimates of P. Rather than using the largest allele frequencies p_i^{max} to arrive at a single number for P in any population, values of P computed via (1) and (2) might be given across a range of subgroups.[202] In light of the mounting evidence that the independence assumptions are reasonable for the VNTR enzyme-probe systems in use, the straight-forward independence method, supplemented by reasonable indications of the uncertainty in the results of these computations, seems to produce the most appropriate estimates of "genotype" frequencies.[203]

In contrast, the NRC Report advocates one form of overestimation because it seeks a procedure that is "appropriately conservative"[204] rather than reasonably accurate.[205] By limiting the presentation to the highest possible range for P in both general population and subpopulation cases, the NRC hopes to sweep the debate about population genetics under the proverbial rug. After all, how can scientists and lawyers quarrel when all that the scientists will say is that the genotype frequency cannot exceed some "appropriately conservative" value? Although this approach is not without appeal in subpopulation cases, where the disparities between the true genotype frequencies and those computed with the basic independence method are potentially the most pronounced, the NRC committee's ad hoc determination of what is "appropriately conservative" is as much a determination based on social policy as a declaration of what is scientifically acceptable.[206] Therefore, it would be a mistake for courts

---

Italians with three-locus profiles).

202. If the databases are such that sampling error is a serious issue, interval estimates can be presented. On the computation of these intervals, *see* Chakraborty et al., *supra* note 90.

203. *But see supra* note 133 (papers proposing the use of parameters that characterize the extent of substructure).

204. NRC REPORT, *supra* note 15, at 94.

205. The panel also "sought to develop a recommendation . . . flexible enough to apply not only to markers now used, but also to markers that might be technically preferable in the future." *Id.* It does not explain why the same procedure must be applied to all markers or why population studies cannot show that the simple independence method will not work with such markers.

206. The consistent undervaluation of the evidence that may result does not trouble the panel because "[w]hatever power is sacrificed by requiring conservative estimates can be

to conclude that scientific practice or theory dictates the use of the one procedure that the panel deems to be "appropriately conservative."

Still, one must ask if it is bad law to allow scientifically defensible estimates to be admitted. The law of evidence does not normally dictate which of several scientifically acceptable methods of analysis an expert may present in court. However, if jurors would be so bemused by a sensitivity analysis of P, or if they would ignore the larger estimates in favor of the more impressively infinitesimal ones, then the overarching objective of achieving a fair assessment of the DNA test results may be difficult to attain with the usual approach. Regrettably, there is no research to date that can definitively resolve this psychological issue of how jurors respond to extreme statistics,[207] but when the risk that the jury will overvalue or be unable to assimilate a range of figures is not demonstrable, the law should allow a qualified expert to pursue the scientifically acceptable approach that the expert finds most congenial. At the very least, the law should permit the expert to present both the "conservative" estimate *and* the best available estimate. This approach is well-suited to match-binning frequencies in general populations, and, arguably, it is acceptable even in the more vexing subpopulation cases.[208]

III. TO BIN OR NOT TO BIN

In Part II, I considered match-binning and the procedures for determining the frequency of a match in a reference population. I argued that more than one approach to producing a match frequency or probability is within the bounds of acceptable scientific practice, and that insisting

---

regained by examining additional loci." *Id.* at 85. As a purely scientific matter, however, it is preferable to be as accurate as possible in estimating the frequency, and then produce a range that reflects the uncertainties in the estimate. Scientists do not normally present parameters of theoretical interest by looking only to one end of a confidence interval, and it would be most peculiar to find a statistician advocating an inconsistent and biased estimator—one that is expected to depart from the true value, even as more and more observations are made, and that tends to err in a particular direction across many samples—simply because it is possible to gather still more data. Worse still, resort to more and more probes raises the risk of false exclusions under a match-no match rule. *See supra* text accompanying note 45.

207. David H. Kaye & Jonathan J. Koehler, *Can Jurors Understand Probabilistic Evidence?* 154(A) J. ROYAL STAT. SOC'Y 75 (1991).

208. The argument against admissibility of any estimate reaches its zenith in subpopulation cases involving uncommon, isolated ethnic groups (such as, perhaps, Polynesian Chamorrons) rather than more common subgroups (such as Italian-Americans). *See supra* note 180.

on the production of the single most conservative figure is not the best legal policy. However, a basic question remains whether even an appropriately computed match frequency—be it a ceiling frequency, a big bin frequency, or a basic bin frequency—should be admissible. The issue, in other words, is no longer whether the evidentiary rules specific to expert testimony demand the exclusion of the match frequency. The analysis and review of the scientific literature in Part II establishes that at least some version of match-binning—be it basic bins, big bins or "ceilings" of one kind or another—satisfies both *Daubert* (or other cases that require a court to assure itself that there is a scientifically valid basis for the testimony)[209] and *Frye* (or other cases that require the court to find general scientific acceptance).[210]

The remaining question involves the familiar balancing test for virtually all evidence: Does the balance of probative value and prejudice favor excluding relevant and scientifically acceptable estimates of the match frequency P in the suitably chosen reference population or populations?[211] The form of prejudice that arguably infects match-binning estimates is that they will unfairly impress the jury and induce them to slight other important evidence. Thus, courts have been concerned that very small fractions, by virtue of their large denominators, are just too impressive for jurors to handle properly and that jurors are likely to misconstrue them as stating the probability of innocence. Furthermore, commentators have argued that jurors may not appreciate the limited

---

209. *See supra* text accompanying note 26.

210. *See supra* text accompanying note 21. One possible source of confusion should be put to rest. How can "conservative" procedures like the NRC's ceiling methods satisfy the general acceptance test when scientists remain divided over the appropriateness of these procedures? This question, however, invites us to confuse the policy question of whether extreme overestimates are necessary or desirable with the more scientifically tractable question of whether the overestimation methods work as advertised to overstate the matching proportion P. There is little, if any, dispute over the proposition that for a structured population with each subpopulation in equilibrium, the ceiling methods produce generous estimates of P. *See* Eric S. Lander, *DNA Fingerprinting: The NRC Report*, 260 SCIENCE 1221 (1993) ("The NRC committee simply concluded that the chosen upper bound sufficed to eliminate serious scientific objections . . . while still allowing odds of up to 6,250,000:1 for a match at four genetic loci."). In fact, the very perception that the methods are enormously generous evokes antagonism on the part of the scientists and statisticians who see the ceiling computations as inappropriate. Consequently, the warning of the *Wallace* court that "the key players in this dispute" over the excessiveness of the ceiling principle must "agree to a compromise on statistical calculation" or "risk preventing any general acceptance at all, thus precluding the admissibility of DNA analysis evidence," 17 Cal. Rptr. 2d 721, 725 (Ct. App. 1993), rests on a failure to recognize what the debate is about.

211. *See supra* note 24.

meaning of match-binning frequencies.

These arguments, and the admissibility of accurate match frequencies, cannot be evaluated in a vacuum. A satisfactory analysis must look to the costs and benefits of admitting an estimate of the match frequency P relative to other methods of informing the jury about the value of the DNA test in discriminating between the innocent and the guilty. There is a spectrum of modes of presentation that may be combined in various ways: the pure opinion format, the improbability format, the likelihood ratio format, and the posterior probability format. In what follows, I explain what I mean by these phrases, and argue that some combination of the second and third approaches should be preferred.

A. The Pure Opinion Format

One can imagine a world in which numbers are *verboten*, and experts are constrained to stating categorical opinions. In the legal universe, this world is more hypothetical than real.[212] For a time, Minnesota seemed to have such a rule, and estimates of genotype population frequencies are still inadmissible regardless of their accuracy.[213] The rule forbidding numerical estimates emerged in a 1978 case involving microscopic

---

212. Cases explicitly rejecting this rule with DNA evidence include United States v. Yee, 134 F.R.D. 161, 211-12 (N.D. Ohio 1991); Martinez v. State, 549 So.2d 694 (Fla. Dist. Ct. App. 1989) (*see supra* text accompanying note 79); People v. Mehlberg, 618 N.E.2d 1168 (Ill. App. Ct. 1993); People v. Lipscomb, 574 N.E.2d 1345 (Ill. App. Ct. 1991); State v. Brown, 470 N.W.2d 30 (Iowa 1991); People v. Adams, 489 N.W.2d 192 (Mich. Ct. App. 1992) ("[T]esting would be a matter of speculation without the statistical analysis."); State v. Cauthron, 846 P.2d 502, 516 (Wash. 1993); Springfield v. State, 860 P.2d 435 (Wyo 1993) (rejecting the observation in Rivera v. State, 840 P.2d 933 (Wyo. 1992), that "the better practice" is the Minnesota rule excluding "statistical probability" because it "could be perceived as an opinion by the expert that the accused is guilty"). *But see* Perry v. State, 586 So.2d 242 (Ala. 1991) (remanding for hearing on Lifecodes's procedures for single locus VNTR tests and computation of $P = 1/209{,}100{,}000$, and whether figure is unfairly prejudicial); State v. Pennell, 584 A.2d 513 (Del. Super. Ct. 1989); John J. Walsh, *Forensic DNA Typing: The Canadian Experience*, PROC. THIRD INT'L SYMP. ON HUM. IDENTIFICATION 85 (1992) (criticizing unreported Canadian cases).

213. In State v. Joon Kyu Kim, 398 N.W.2d 544 (Minn. 1987), the court departed slightly from the pure "no numbers" rule. It allowed testimony as to frequencies of each protein or enzyme marker in a semen stain. Under this variant of the rule, the jury may be told the frequency of each marker, but not the frequency of their combination. Applied to VNTR studies, it would allow the expert to give the frequency of each "allele" and perhaps of the single-locus "genotypes" obtained from equation (1), but not of the multilocus "genotype" derived from (2). Cases pursuing this exception include State v. Johnson, 498 N.W.2d 10 (Minn. 1993) and State v. Alt, 504 N.W.2d 38 (Minn. Ct. App. 1993).

comparisons of hair samples.[214] The Minnesota Supreme Court applied it to blood antigens and serum proteins in the mid-1980s.[215] Then, in *State v. Schwartz*,[216] the court held that it governed DNA evidence as well. In this last case, police investigating the stabbing death of Carrie Coonrod found and seized bloodstained blue jeans in Thomas Schwartz's residence. Cellmark Diagnostic Corporation's report concluded that "it is the opinion of the undersigned that the DNA banding patterns obtained from the stain removed from the blue jeans and the blood of Carrie Coonrod are from the same individual."[217] This opinion rested on a "banding pattern [whose frequency] in the Caucasian population is approximately 1 in 33 billion."[218] The state urged the supreme court to allow this statistic to be admitted "after an adequate opportunity for cross examination and limiting instructions."[219] The court declined this invitation. "In dealing with complex technology, like DNA testing," it wrote, "we remain convinced that juries in criminal cases may give undue weight and deference to presented statistical evidence and are reluctant to take that risk."[220]

The defect in the Minnesota rule is obvious. The complex technology of DNA testing can produce figures that are not only relevant, but highly probative. The jury needs some estimate of the population frequency or the probability of a match with another source to give a match the weight it deserves.[221] An expert may be needed to calculate *P*, but the expert's

---

214. State v. Carlson, 267 N.W.2d 170, 176 (Minn. 1978), *described infra* note 246.

215. State v. Boyd, 331 N.W.2d 480, 482 (Minn. 1983); *Joon Kyu Kim* (allowing testimony as to frequencies of each marker but not as to the frequency of the set of incriminating markers).

216. 447 N.W.2d 422, 428 (Minn. 1989).

217. *Id.* at 424.

218. *Id.* Cellmark used a multilocus probe, which produces profiles that are harder to interpret statistically than the series of single locus probes that have come to dominate criminal testing. Thus, the 1/33 billion figure is a binomial probability computed in a different fashion from (1) and (2), which apply only to single locus probes. *See* Kaye, *supra* note 1. Interestingly, the calculation is essentially identical to one used over a century ago to analyze an allegedly forged signature in Robinson v. Mandell, 20 F. 1027 (C.C.D. Mass. 1868), *described in* Paul Meier & Sandy Zabell, *Benjamin Pierce and the Howland Will*, 75 J. AM. STAT. ASS'N. 497 (1980).

219. 447 N.W.2d at 428.

220. *Id.* The Minnesota Supreme Court adhered to this reasoning and result in State v. Jobe, 486 N.W.2d 407 (Minn. 1992). In State v. Nielsen, 467 N.W.2d 615, 620 (Minn. 1991), it intimated that MINN. STAT. § 634.26 (1989), which was enacted to overturn the *Carlson* line of cases, is somehow unconstitutional.

221. *See, e.g.*, Nelson v. State, 628 A.2d 69, 76 (Del. 1993) (finding trial court's exclusion of match frequency "inherently inconsistent" with its admission of testimony of a match, because "without the necessary statistical calculations, the evidence of the match was 'meaningless' to the jury"); State v. Brown, 470 N.W.2d 30 (Iowa 1991)

qualitative opinion about the conclusion to be drawn from this statistic or an expert's verbal characterization of this number is not based on any expertise in laboratory chemistry, genetics or biostatistics.[222] Unless invited by the defendant, such testimony should not be allowed.[223]

Only if it were certain, or nearly so, that jurors would misuse any such number would it be desirable to leave them at sea and hope that they might make it to port on their own. But there is no clear indication that "undue weight and deference" to statistical evidence is any more likely than insensitivity and hostility to the evidence[224] or helpless capitulation to an inscrutable opinion. Consequently, a blanket rule against statistics or probabilities relating to DNA evidence is unjustified. Global doubts about jurors' abilities to handle statistics do not lead to the conclusion that the dangers of prejudice substantially outweigh the probative value of well-founded estimates of population frequencies.

(holding expert testimony that "the likelihood of a person matching in all four fragments . . . would be one in several billion" admissible, since "[w]ithout statistical evidence, the ultimate results of DNA testing would become a matter of speculation"); State v. Vandebogart, 616 A.2d 483, 494 (N.H. 1992) ("A match is virtually meaningless without a statistical probability expressing the frequency with which a match could occur."); NRC REPORT, *supra* note 15, at 74 ("To say that two patterns match, without providing any scientifically valid estimate (or, at least, an upper bound) of the frequency with which such matches might occur by chance, is meaningless.").

It would not, however, be "meaningless" to inform the jury that two samples match and that this match makes it more probable, in an amount that is not precisely known, that the DNA in the samples comes from the same person. Nor, when all estimates of the frequency are in the many millionths or billionths, would it be meaningless to inform the jury that there is a match that is known to be extremely rare, if not unique, in the general population.

At least one court has suggested that the NRC Report's "meaningless" remark demonstrates that *Frye* precludes presenting evidence of a match without an estimate of the genotype frequency. State v. Bible, 858 P.2d 1152 (Ariz. 1993), *construing* State v. Cauthron, 846 P.2d 502 (Wash. 1993). This view is plainly mistaken. The general acceptance standard addresses the validity and reliability of the methodology that produces evidence of identity. The fact of a match is scientifically valid evidence of identity as long as it can be shown from theory and data that the genotype is not ubiquitous in the relevant population. How valid scientific evidence of a match should be presented to a jury is a legal rather than a scientific issue falling far outside the domain of the *Frye* test.

222. Even in Minnesota, it may be that opinions beyond the bland statement of a match are inadmissible. *See* State v. Alt, 504 N.W.2d 38, 52 (Minn. Ct. App. 1993).

223. The outcome in State v. Cauthron, 846 P.2d at 515-16, is consistent with this suggestion. Cellmark's Dr. Robin Cotton testified that she had "no doubts" that the defendant was "the source of the semen sample in the five [rape] cases that we got the result on" and that "the DNA could not have come from anyone else on earth." The court held that because this opinion testimony was not supplemented or replaced with "background probability information," it should not have been allowed. *Id.* at 516.

224. *See* Kaye & Koehler, *supra* note 207.

B. The Improbability Format

Because most DNA testers have chosen to compute match-binning frequencies, DNA evidence usually comes in the form of a determination of a match accompanied by a small number that is said to show the improbability of a match in a population of innocent suspects. Although nearly all courts dismiss the broad-brush objection to these numbers, there are subtle—and troublesome—ways in which match-binning frequencies could be unfairly prejudicial. These possibilities do not, I think, dictate a flat ban on *P*, but they do require steps to avoid abuse or misuse of the figure.

1. P is not $P(M_D \mid O)$

A more sophisticated legal criticism than the global objection to numbers is that population frequencies may be mistaken for the frequency with which the laboratory will declare a match between the defendant's sample and the crime sample (M_D) when the samples are from different sources (O). Contrary to what some testifying experts have claimed or implied,[225] the frequency of a DNA profile in a given population only reveals how often an *error-free* DNA test will give false positives when applied to that population. If matches result both from people whose DNA truly satisfies the matching and from people whose DNA does not match, but appears to because of non-random error such as mislabeling,[226] then the rate of false positives will be larger than the proportion *P*. In practice, of course, DNA tests are not always free of all non-random errors,[227] and even a tiny probability of a false positive error typically will swamp the vanishingly small estimates of population frequencies associated with matches at four or five VNTR loci.

Three strategies to counter the danger that a jury will confuse a match frequency with the probability of a false positive have been proposed.

---

225. *See* Kelly v. State, 792 S.W.2d 579, 583 (Tex. Ct. App. 1990) ("He [Kevin McElfresh of Lifecodes] noted that the statistical probabilities of such a match being incorrect was one in thirteen million."). For more examples, see Jonathan J. Koehler, *Error and Exaggeration in the Presentation of DNA Evidence at Trial*, 34 JURIMETRICS J. 21 (1993).

226. Undetected degradation and band shifting are not likely to generate false positives. On the possible sources of false positive laboratory errors, see Thompson & Ford, *supra* note 76.

227. *See supra* text accompanying note 77.

One is to reduce the false positive risk by improving laboratory proficiency and submitting samples to three laboratories for independent analyses.[228] A second response is to withhold the population frequency estimate and present the jury with the probability of a false positive in the case at bar, considering both the laboratory's rate of false positives on blind proficiency tests and the population frequency.[229] The third solution is to provide the jurors with both the laboratory false positive error rate and the estimated population proportion P, thereby impressing on the jury that the latter cannot be equated to the probability of a false match.[230]

---

228. Lempert, *supra* note 74, at 327-28. Whether the costs, both in terms of resources and increased false negatives, justify multiple testing is unclear. It may be enough to give defendants the right to retest at different laboratories and to subsidize multiple tests for indigent defendants who demand them. *Cf.* James Wooley & Rockne P. Harmon, *The Forensic DNA Brouhaha: Science or Debate?*, 51 AM. J. HUM. GENETICS 1164 (1992) (letter urging defense experts to retest rather than theorize about the possible sources of laboratory error).

In any event, vigorous legislative or administrative action to reduce the risk of false positive and false negative errors alike is eminently desirable. Even with the unusual safeguard of imposing on the proponent of DNA evidence at a preliminary hearing the burden of proving that a match follows from properly applied laboratory procedures, see E. Imwinkelried, *The Debate in the DNA Cases Over the Foundation for the Admission of Scientific Evidence: The Importance of Human Error as a Cause of Forensic Misanalysis*, 69 WASH. U. L.Q. 19 (1991), it will be immensely difficult to detect possible errors in a particular case. Moreover, the judicial system is unlikely to produce sufficient incentives for quality control. If the DNA testing is done moderately well, but not as well as it could be, the court must decide whether to exclude generally probative evidence because of the possibility that the laboratory may have erred in the case at bar. Courts are rightly loathe to exclude such evidence without a specific indication of laboratory error. If all cases went to trial and all defendants had skilled and astute counsel with access to experts who could look over the shoulders of the laboratory technicians, so to speak, the state would feel strong pressure to invest in the laboratories up to the point at which marginal benefits flowing from the admission of the laboratory findings equals the marginal cost of improvements in laboratory procedures. However, the vast majority of cases never reach trial, and very few defense lawyers have the knowledge and resources required to identify the particular instances when laboratory imperfections actually cause a problem. Prosecutions will be instituted and most defendants will plead guilty when faced with infinitesimal match-binning probabilities. At some point, of course, demands for quality control become excessive, but there is every reason to get things right before trial. The optimal level of quality control therefore is farther in the direction of increased expenditures than might at first be imagined.

229. *See* Paul J. Hagerman, *DNA Typing in the Forensic Arena*, 47 AM. J. HUM. GENETICS 876 (1990); Lempert, *supra* note 74, at 325-26.

230. *See* Russell Higuchi, *Human Error in Forensic DNA Typing*, 48 AM. J. HUM. GENETICS 1215 (1991); NRC REPORT, *supra* note 15, at 88 & 94 ("A laboratory's overall rate of incorrect conclusions due to error should be reported, but separately from, the probability of coincidental matches in the population. Both should be weighted in evaluating evidence."); *id.* at 89 ("The jury should be told both results."). Presumably, expert testimony could assist by combining the error rate with P to arrive at $P(M_D \mid O)$, the probability that the laboratory will declare a match for defendant given

These proposals underscore the point that the estimated population proportion P is not the probability that the DNA analysis that incriminated the defendant would incriminate an innocent person. Although the point does not mandate a categorical exclusion of P, it does militate in favor of adopting, as a precondition to the admission of P, a procedure, such as those described above, that would emphasize the distinction to the jury. Requiring the expert to give an estimate of the rate of false matching on independently administered blind proficiency tests may be the simplest prophylactic.[231]

2. *P is not P(O | M_D)*

Presenting P also can produce prejudice if the jury misinterprets it as the probability that someone other than the defendant is the source of the crime sample. In a sense, this is a more fundamental objection, since it pertains even to an error-free test, for which P actually is the risk of a false positive. Assuming, for simplicity, that the test is error-free, the fallacy works something like this: (a) P, the frequency of a match in the reference population, is the probability that an innocent person would match the crime sample; (b) defendant does match; therefore, (c) P is the probability that defendant is innocent.

Where does the fallacy occur? The first two steps are correct. If the population proportion is, say, $P = 1/100{,}000$, then the probability that any randomly selected person D will match (an event we may designate M_D) given that someone other than D is the source of the crime sample

---

that someone else is the source. If no explanation is provided, the jury may "be helplessly confused about the weight to accord the testimony [of a match] because ordinary people are not very good at working with conditional probabilities." Lempert, *supra* note 74, at 325.

Another source of possible error in the interpretation of P is the presence of relatives, who have a greater chance of sharing alleles with the defendant and matching the crime sample, than the figure P suggests. This is really an aspect of the problem of defining the reference population. Lempert capably surveys the possible solutions and concludes that "until technology advances, the most honest approach is to present the jury with the probability that it was left by one of the group of defendant's relatives whom the state has not been able to exclude from the suspect population." *Id.* at 214; *cf.* NRC REPORT, *supra* note 15, at 87; Balding & Nichols, *supra* note 133. *But see* Lempert, *supra* note 176 (conceding that the ceiling procedure is difficult to justify on scientific grounds, but defending it as an indirect vehicle for accomodating the problems of laboratory error and "micropopulations").

231. Naturally, defense counsel would remain free to buttress this generalized information with arguments about the adequacy of the laboratory work in a particular case.

(an event that may be denoted as O) is $P(M_D \mid O) = 1/100{,}000$. The fallacy occurs at the last step, which speaks of the probability $P(O \mid M_D)$ that D is not the source given the match between D and the crime sample. The rules of probability reveal that $P(M_D \mid O)$ is not generally equal to $P(O \mid M_D)$. The probability that a card drawn from a well shuffled deck is an ace of diamonds given the fact that it is a red card is 1/26, but the probability that it is a red card given that it is the ace of diamonds is one.

Although it is an elementary mistake to conflate the conditional probability of the match given innocence with the conditional probability of innocence given the match,[232] more than one court has fallen prey to this "inversion fallacy."[233] For example, the California court of appeals, in *People v. Axell*,[234] thought that Cellmark's report that "the frequency of that DNA banding pattern in the Hispanic population is approximately 1 in 6 billion" meant "that the chance that any but appellant left the unknown hairs at the scene of the crime is 6 billion to 1." Courts in Arizona,[235] Colorado,[236] Georgia[237], Illinois,[238] Indiana,[239] Mississippi,[240]

---

232. *See, e.g.*, MCCORMICK, *supra* note 22, § 211; Jonathan J. Koehler, *DNA Matches and Statistics: Important Questions, Surprising Answers*, 78 JUDICATURE 222, 224 (1993). Another way to recognize that P is not $P(O \mid M_D)$ is to consider the impact of population size on this probability. *Cf.* Laurence H. Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329 (1971). Suppose, as in Kelley v. State, 792 S.W.2d 579, 582 (Tex. Ct. App. 1990), the reference population consists of "white males" and the incriminating profile occurs with an estimated frequency of $P = 1/13{,}500{,}000$. If the reference population numbers 27 million, then the expected number of matching DNA profiles is two. The defendant is one of these two, which suggests—in the absence of other information linking him as opposed to the other potential match to the crime—that the chance that the other man is the source of the crime sample is one-half. This is a far cry from the court's thought that "[t]he statistical probability that the semen came from another white male was 1 in 13.5 million." *Id.* at 582. Of course, there is no particular reason to think that the reference population in *Kelly* numbers 27 million. It probably is much less. But that does not diminish the logical force of the argument. The one in 13.5 million figure is just a population proportion. It neither grows nor shrinks according to the size of the reference population. Yet, the conditional probability $P(O \mid M_D)$ that someone other than the matching defendant is the source is related to the number of other people who could be the source. Hence, these two quantities are not identical; even though P can be interpreted as $P(M_D \mid O)$, $P(M_D \mid O)$ cannot be equated with $P(O \mid M_D)$. *See, e.g.*, Koehler, *supra* note 225.

233. Kaye & Koehler, *supra* note 207. It also is called the "prosecutor's fallacy." William Thompson, *Are Juries Competent to Evaluate Statistical Evidence?*, 52 LAW & CONTEMP. PROBS. 9 (1989). Commentators, as well as experts, courts, and jurors, also have been known to commit this error. *See, e.g.*, Joseph Liebeschuetz, *Statutory Control of DNA Fingerprinting in Indiana*, 25 IND. L. REV. 204, 208 (1991) ("The exclusion frequency is the relative probability that the defendant committed the crime compared with a person selected at random from the general population.").

234. 1 Cal. Rptr. 2d 411 (Ct. App. 1991).

235. State v. Bible, 858 P.2d 1152, 1165 (Ariz. 1993) ("Cellmark concluded that the

New York,[241] South Dakota,[242] Tennessee,[243] Texas,[244] and the United Kingdom[245] have made or been presented with similar inversions.[246]

---

chances were one in 14 billion . . . that the blood on Defendant's shirt was not the victim's."); *id.* at 1189 (referring to "the product rule and the resulting opinion of the odds against a random match"). The court criticized the state for "tacitly [attempting] to argue that these probability figures could be equated with the probability that someone other than Defendant committed the crime." *Id.* at 1185.

236. Fishback v. People, 851 P.2d 884, 888 (Colo. 1993) ("Once a match is determined, its statistical significance . . . is usually expressed in terms of the likelihood that the crime scene samples came from a third person who has the same DNA profile as the suspect.").

237. Hornsby v. State, No. A93A1270, 1993 WL 497094, at *6 (Ga. Ct. App. Oct. 18, 1993) ("[T]he chances that the semen recovered from the victim belonged to someone other than the defendant were one in 70 million . . . ."); Bradford v. State, 420 S.E.2d 4, 5 (Ga. Ct. App. 1992) (Apparently unchallenged FBI DNA tests in rape case said to show that "the odds someone other than defendant attacked the victim were 1 in 49 million.").

238. People v. Miles, 577 N.E.2d 477, 484 (Ill. App. Ct. 1991) ("The probability of an African-American other than the defendant leaving the semen stain on the bed sheet . . . was 1 in 300,000.").

239. McElroy v. State, 592 N.E.2d 726, 728 (Ind. Ct. App. 1992) ("The State presented DNA identification evidence which showed the odds were 20 million to one that he committed the rape.").

240. Polk v. State, 612 So2d 381, n. 1 (Miss. 1992) ("The probability that the blood . . . was from any person other than Georgia Mae Thomas was calculated to be 1 in 530,000,000.").

241. People v. Davis, 601 N.Y.S.2d 174, 175 (App. Div. 1993) ("A Lifecodes technician . . . declared at trial that the statistical probability of someone other than the perpetrator providing the alleged 'match' was 'one in ten million.'").

242. United States v. Martinez, 3 F.3d 1191, 1193 (8th Cir. 1993) ("The FBI concluded that there was a 1 in 2600 probability that the semen found on the panties came from someone other than Martinez."); United States v. Two Bulls, 918 F.2d 56, 57 n. 2 (8th Cir. 1990) ("[P]robability of someone other than Two Bulls providing a match was one in 177,000."), *vacated for reh'g en banc but appeal dismised due to death of defendant*, 925 F.2d 1127 (8th Cir. 1991).

243. State v. Myers, 1993 WL 1416512 (Tenn. Crim. App. May 4, 1993) (Unpublished opinion reporting that an FBI agent "concluded that a 1 in 50,000 chance existed that an individual unrelated to and other than the defendant produced the semen sample found on the victim's clothing.").

244. Kelly v. State, 792 S.W.2d 579, 582 (Tex. Ct. App. 1990) ("The statistical probability that the semen came from another white male was 1 in 13.5 million."); Transcript at 2327, State v. Bethune, 821 S.W.2d 222 (Tex. Ct. App. 1991) ("There would [be] a one in 5 billion chance that anybody else could have committed the crime.").

245. R. v. Cannan, 92 Crim. App. 16 (1991) ("So far as the DNA evidence was concerned it seems that the chances of anyone else having been responsible for the semen found on the knickers was something like 260 million to one against.").

246. For still more examples and a discussion of the forces that induce these errors, see Koehler, *supra* note 225. The error is hardly confined to DNA identifications. *See, e.g.*, United States *ex rel.* DiGiacomo v. Franzen, 680 F.2d 515, 516 (7th Cir. 1982) (State criminalist testified that "the chances of another person belonging to that hair would be 1/4,500."); State v. Carlson, 267 N.W.2d 170, 175 (Minn. 1978) (Expert testified that "the likelihood that the hair found in the rug . . . in Carlson's bedroom . . . did not come from the victim would be on the order of one chance in 4,500."). The

The Minnesota court in *Schwartz* perceived the inversion fallacy as a reason to exclude *P* altogether,[247] but this reaction seems precipitous if less drastic measures will reduce the danger. Such measures include cross examination and opposing expert testimony or jury argument about the meaning of *P*.[248] They also include a rule that would preclude prosecutors or experts from describing *P*, as some do,[249] in ways that encourage the commission of the fallacy. Broader awareness of the fallacy should go far toward retarding its influence in the courtroom.

C. The Likelihood Ratio Format

I have argued that suitably computed and presented match-binning frequencies and probabilities pass muster under the conventional rules of evidence. They pose some danger of misinterpretation, but the risk can be reduced to the point where the usefulness of the testimony justifies its admission. This does not mean, however, that *P* has to be introduced in court in preference to any alternative. Match-binning, as we have seen, has several drawbacks. The threshold for declaring a match is arbitrary and existing match rules may be producing a high rate of false non-matches. The need to fit all comparisons into two rigid categories obscures distinctions that are reintroduced in vague ways when experts speak of "exact" matches on the one hand, or "inconclusive" exclusions

---

hair cases are reviewed more fully in THE EVOLVING ROLE OF STATISTICAL ASSESSMENTS AS EVIDENCE IN THE COURTS 60-67 (Stephen E. Fienberg ed., 1988). For a new set of abuses, see Commonwealth v. Pandolfino, 596 N.E.2d 390 (Mass. App. Ct. 1992).

247. State v. Schwarz, 447 N.W.2d 422, 428 (Minn. 1989) ("There is a real danger that the jury will use the evidence as a measure of the probability of the defendant's guilt or innocence.") (quoting State v. Boyd, 331 N.W.2d 480, 483 (Minn. 1983)).

248. *See* MCCORMICK, *supra* note 22 § 211; Kaye & Koehler, *supra* note 207. Existing empirical research indicates that the inversion fallacy can be counteracted with an argument like that based on population size. *See supra* note 232; William C. Thompson & Edward L. Schumann, *Interpretation of Statistical Evidence in Criminal Trials: The Prosecutors' Fallacy and the Defense Attorney's Fallacy*, 11 LAW & HUM. BEHAV. 167 (1987). However, this work is based on much larger values for matching probabilities *P*, and the counterargument probably would be less effective when an enormous population size would be needed to generate many falsely incriminated people in the reference population.

249. *See, e.g.*, People v. Miles, 577 N.E.2d 477 (Ill. App. Ct. 1991) (Cellmark's expert testified that, using database of African-Americans in Detroit, "the probability of an African-American other than the defendant leaving the semen stain on the bed sheet . . . was 1 in 300,000."); *cf.* United States v. Massey, 594 F.2d 676 (8th Cir. 1979) (improper closing argument concerning probability of matching hair samples).

on the other.[250] Finally, the matching frequency often is described in words that make it seem like the likelihood of innocence.[251]

Recognizing these problems, some statisticians have devised other methods for conveying the implications of the similarities between DNA samples. These alternatives dispense with the classification of test results into "matches" and "nonmatches," and instead look to the degree of similarity in the DNA fragments by quantifying the hypotheses that the defendant is the source (S) as opposed to the alternative that someone else is (O). These quantities come together in the likelihood ratio for the test results, which expresses how many times more probable the results are under S than O, and, hence, the relative likelihood of S and O.[252] If, for example, the measured differences in lengths of the VNTR fragments from the crime sample and the suspect's sample would arise nine times out of ten when the suspect is indeed the source of the crime sample, but only one time in 100,000 when someone else in the relevant population is the source, then the likelihood ratio would be $L = (9/10)/(1/100,000) = 90,000$.[253]

I shall not dwell on the details of producing the likelihoods. There are competing suggestions.[254] All involve a statistical model of the measurement error and an analysis of the distribution of DNA fragment sizes in a reference population with sampling error.[255] None can be dismissed as

---

250. *See supra* Part I.

251. *See supra* Part III(B)(2).

252. *See generally* A.W.F. EDWARDS, LIKELIHOOD: AN ACCOUNT OF THE STATISTICAL CONCEPT OF LIKELIHOOD AND ITS APPLICATION TO SCIENTIFIC INFERENCE (1972).

253. An analogous ratio using match-binning can be computed. If there is a match in an error-free test at n loci using a match window of three standard deviations for the $2n$ (presumed) independent measurements, this ratio is $L_M = (.99)^{2n}/P$. The numerator is the probability of a match on the $2n$ fragments from a common source; the denominator is the probability of a match drawn at random from the reference population in which the frequency of the matching "genotype" is P. The likelihood ratio L in the text is not computed in this way, and there need be no "match" (according to some preset match rule) in the fragments. The numerator of L represents the probability density of the measured differences in the fragment lengths (whether or not they fall into some preordained match window) for a common source. The denominator is the probability density for these differences (without regard to any preset bins or binning rules) for the crime sample and one drawn at random from the reference population.

254. *See* Donald A. Berry, *Inferences Using DNA Profiling in Forensic Identification and Paternity Cases*, 6 STAT. SCI. 175 (1991); Bernard Devlin et al., *Forensic Inference from DNA Fingerprints*, 87 J. AM. STAT. ASS'N 337 (1992); Jeffrey Morris et al., *Biostatistical Evaluation of Evidence from Continuous Allele Frequency Distribution Deoxyribonucleic Acid (DNA) Probes in Reference to Disputed Paternity and Identity*, 34 J. FORENSIC. SCI. 1311 (1989); *cf.* D.W. Gjertson et al., *Calculation of Paternity Using DNA Sequences*, 43 AM. J. HUM. GENETICS 860 (1988) (likelihood ratio for paternity).

255. Sampling error refers to possible differences between the sample and the

unreasonable or based on principles not generally accepted among the statistical community. Therefore, as with match frequencies, unless likelihood ratios are so unintelligible as to provide no assistance to the jury or so misleading as to be unduly prejudicial, they should be admissible.

Prejudice seems the more serious of these possibilities. As with match frequencies, the proposed likelihood ratios do not account for laboratory error, and a jury might misconstrue even a modified version that did as a statement of the odds in favor of S.[256] Just as these objections track those made against the matching frequency, so do the rejoinders. Once again, admission of the likelihood ratio *L* should not be allowed unless the risk of a false positive is incorporated formally or placed along side it. As for the second possible misinterpretation of *L*, that too is a question of jury psychology, and the answer is too uncertain to warrant excluding a statistically acceptable calculation. An expert who desires to present a reasonably computed value of *L*, either as a substitute for or a supplement to *P*, should be allowed to do so.[257]

D. The Posterior Probability Format

The likelihood ratio, while an improvement over the match-binning frequency, is still one step removed from what the judge or jury truly seeks—an estimate of the probability P(S | X) that the crime sample is the suspect's DNA given the observed fragment lengths X in DNA extracted from the samples. Recognizing this, a number of statisticians have argued that the likelihood ratio should not be presented to the jury in its own right,[258] but should be used to estimate the probability that the

---

population from which it is drawn.

256. The possibility of misinterpretation is present in the use of the phrase "identity index" that Devlin et al., *supra* note 254, at 341, propose for the likelihood ratio in this context. It also is present with a proposal for "a verbal convention, which maps from ranges of the likelihood ratio to selected phrases" like "strong evidence" or "weak evidence." Ian W. Evett, Comment, 6 STAT. SCI. 200, 201 (1991). *Cf.* David H. Kaye, *The Probability of an Ultimate Issue: The Strange Cases of Paternity Testing*, 75 IOWA L. REV. 75, 99-100 (1989) (criticizing the comparable convention of "verbal predicates" used in paternity testing).

257. If anything, one might argue that inasmuch as *L* includes all the information in *P* and more besides, it should be required, and *P* should be excluded. Some jurors, however, may find the less statistically sophisticated *P* a more comprehensible figure. As long as both quantities are relevant and not unduly prejudicial, it should be left to the proponent of the evidence to decide whether to introduce *P*, *L*, or both.

258. *See, e.g.*, Evett, *supra* note 256, at 201 ("[J]ust leaving a court with a likelihood

suspect is the source of the crime sample.[259] And a few experts have been willing to speak to this probability. In *Smith v. Deppish*,[260] the state's "DNA experts informed the jury that . . . there was more than a 99 percent probability that Smith was a contributor of the semen found on the swab." Likewise, in *State v. Thomas*,[261] a geneticist testified that "the likelihood that the DNA found in Marion's panties came from the defendant was higher than 99.99%."

Before accepting such pronouncements as admissible—as these courts apparently have—one should ask how these probabilities are obtained and whether they are appropriately placed before a jury. Although the opinions are silent on these matters, only one mathematically valid procedure is known for arriving at a probability that a defendant is the source. From the DNA testing in question and data on the distribution of the VNTR fragments in the reference population, we can estimate the probability P(X | S) of the measurements X given the hypothesis S that the suspect is the source of the DNA. Likewise, we can estimate the probability P(X | O) under the alternative hypothesis O that the crime sample DNA comes from another source. For concreteness, suppose, as before, that these probabilities are 9/10 and 1/100,000, respectively. They are conditional probabilities in that they pertain to an outcome (X) on the condition that one or another hypothesis (S or O) is true. But the conditioning runs in the wrong direction. We seek P(S | X), the probability that the defendant is the source of the crime sample given the data X, or P(O | X), the probability that someone else is the source given this same information. We already have seen that P(O | X) is not 1/100,000—to switch the letters around like this is to commit the inversion fallacy.[262] To invert P(X | S) or P(X | O) correctly takes more

---

ratio does not seem enough."); *cf.* Stephen E. Fienberg, Comment, *The Increasing Sophistication of Statistical Assessments as Evidence in Discrimination Litigation*, 77 J. AM. STAT. ASS'N 784 (1982) (criticizing presentation of a relative likelihood function).

259. *See e.g.*, Berry, *supra* note 254. *But see* Donald A. Berry, *Rejoinder*, 6 STAT. SCI. 202, 203-04 (1991). The NRC panel pretermitted all proposals involving likelihood ratios or posterior probabilities on the curious ground that "no forensic laboratory in this country has, to our knowledge, used Bayesian methods to interpret the implications of DNA matches in criminal cases." NRC REPORT, *supra* note 15, at 62. Under this reasoning, the panel should not have urged external blind proficiency of laboratories by a federal committee as a prerequisite to admissibility and should not have proposed the ceiling method of computing match-binning probabilities.

260. 807 P.2d 144, 148 (Kan. 1991).

261. 830 S.W.2d 546, 550 (Mo. Ct. App. 1992).

262. *See supra* Part III(B)(2).

work. The correct answer, however, is well-known:[263]

$$\text{Odds}(S \mid X) = L \text{ Odds}(S) \quad (3).$$

In words, the posterior odds (considering the fragment lengths X) that the defendant is the source are just the likelihood ratio times the prior odds (those formed without knowing this information).[264] In our illustration L = (9/10)/(1/100,000) = 90,000. Starting from the (dubious) premise that the presumption of innocence should be interpreted to mean that the defendant has the same chance as anyone else in the United States of being the source of the crime sample,[265] it follows that the DNA evidence raises the odds of S to 90,000/300,000,000 = 3/10,000. Alternatively, starting with prior odds of one, the DNA evidence prompts the conclusion that the posterior odds are 90,000 to one.

Expressions like (3) have a rich history in statistics and law. Known as Bayes's rule because of their ancestry,[266] they have been the subject of a protracted debate among academically inclined lawyers and statisticians.[267] In courtroom practice, three procedures have been used. In the expert-prior-odds implementation, the scientist implicitly or explicitly selects a prior probability for the jurors, applies Bayes's rule, and informs the jury that the scientific evidence establishes a single probability for the event in question. The prosecution relied on a Bayesian analysis of this type in *State v. Klindt*,[268] a gruesome chainsaw murder case decided before the emergence of DNA testing, and the Supreme Court of Iowa affirmed the admission of a statistician's testimony as to a posterior

---

263. *See, e.g.*, MICHAEL O. FINKELSTEIN & BRUCE LEVIN, STATISTICS FOR LAWYERS 93 (1990).

264. The odds in favor of an event are the probability that it will occur divided by the probability that it will not occur. Hence, Odds(S) = P(S)/P(O) and Odds(S | X) = P(S | X)/P(O | X). For instance, if the probability of an event is 1/4, then the odds are (1/4)/(1 − 1/4) = 1/3, or 1:3.

265. This interpretation of the presumption of innocence is found in John Kaplan, *Decision Theory and the Factfinding Process*, 20 STAN. L. REV. 1065 (1968). If the population of the United States is 300,000,000, the prior probability is 1/300,000,000, and the prior odds are 1/299,999,999 ≈ 1/300,000,000.

266. They date back to a paper appearing in 1763 and attributed to the late Reverend Thomas Bayes.

267. *See generally* Symposium, 13 CARDOZO L. REV. Nos. 2-3 (1991); David H. Kaye, *Introduction: What is Bayesianism?, in* PROBABILITY AND INFERENCE IN THE LAW OF EVIDENCE: THE LIMITS AND USES OF BAYESIANISM 1 (P. Tillers & E.C. Green eds., 1988), *reprinted as What is Bayesianism? A Guide for the Perplexed*, 28 JURIMETRICS J. 161 (1988).

268. 389 N.W.2d 670 (Iowa 1986).

probability in excess of 99 percent that a torso found in the Mississippi River was what remained of the defendant's missing wife. It is doubtful, however, that the Iowa courts appreciated the basis of the calculation. For years, courts in civil paternity cases involving testing of antigens routinely admitted testimony of posterior probabilities computed under the ad hoc and often undisclosed selection of a prior probability of one-half.[269] These courts probably did not recognize the Bayesian nature of the "probability of paternity" laid before them, but courts unmistakably apprised of the foundations of these probabilities have continued to approve of them.[270] Nevertheless, the expert-prior-odds approach is clearly ill-advised. It does not permit or assist the jury in integrating the scientific proof with the other evidence in the case. Instead, it requires the jury to defer to the expert's choice of the prior odds, even though the scientist's special knowledge and skill merely extend to the production of the likelihood ratio for the scientific evidence.

A second approach—the jury-prior-odds implementation—overcomes this defect. It requires the jury to articulate prior odds, to use them as prescribed by (3), and to return a verdict of guilty if the posterior odds exceed some threshold that expresses the point at which the reasonable doubt standard is satisfied. But this procedure raises serious questions about the jury's ability to translate beliefs into numbers[271] and about the desirability of quantifying the vague concept of reasonable doubt.[272] It, too, is far from optimal.

---

269. This practice first was criticized in Ira Ellman & David Kaye, *Probabilities and Proof: Can HLA and Blood Group Testing Prove Paternity?*, 54 N.Y.U. L. REV. 1131 (1979).

270. A few have imposed restrictions on the practice. *See, e.g.*, Commonwealth v. Beausoleil, 490 N.E.2d 788 (Mass. 1986) (criticized in Kaye, *supra* note 256. In Plemel v. Walter, 735 P.2d 1209 (Or. 1987), the Oregon Supreme Court rejected the expert-prior-odds implementation in favor of the variable-prior-odds Bayesian procedure discussed, *supra* text accompanying note 263. *See* David H. Kaye, *Plemel as a Primer on Proving Paternity*, 24 WILLAMETTE L. J. 867 (1988). Some rape cases in which the prosecution relies on a "probability of paternity" using undisclosed prior odds of one have generated appellate opinions critical of that probability. *See, e.g.*, State v. Hartman, 426 N.W.2d 320 (Wis. 1988). However, the opinions are not well reasoned. *See* Kaye, *supra* 256.

271. *See* Tribe, *supra* note 232; David H. Kaye, Comment, *Uncertainty in DNA Profile Evidence*, 6 STAT. SCI. 196, 199 (1991).

272. *See* Charles R. Nesson, *Reasonable Doubt and Permissive Inferences: The Value of Complexity*, 92 HARV. L. REV. 1187 (1979); Tribe, *supra* note 232. *Compare generally* Charles R. Nesson, *The Evidence or the Event? On Judicial Proof and the Acceptability of Verdicts*, 98 HARV. L. REV. 1357 (1985); *with* Daniel Shapiro, *Statistical-Probability Evidence and the Appearance of Justice*, 103 HARV. L. REV. 530 (1989).

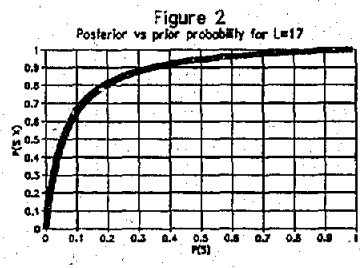
The third approach circumvents most of these problems.[273] In this variable-prior-odds implementation, the expert neither uses his own prior odds nor demands that the jurors articulate their prior odds for substitution into Bayes's rule. Rather, the expert presents the jury with a table or graph showing how the posterior probability changes as a function of the prior probability.[274] Bayes's rule merely acts as a heuristic device, displaying the force of the evidence across a wide range of prior probabilities. No juror need adopt Bayes's rule or any prior probability, but all jurors can see the distinction between $P(X \mid O)$, the probability of the evidence under the hypothesis that someone other than the defendant is the source, and $P(S \mid X)$, the probability that the defendant is the source given the evidence.

Thus, the variable-prior-odds implementation of Bayes's rule should be at least permissible.[275] It has the potential of preventing the judge and jury from misconstruing the match-binning probability $P(M_D \mid O)$ as the probability of innocence, from mistaking the probability $P(X \mid O)$ for the probability of innocence, and from misinterpreting the likelihood ratio L

---

273. *See* Ellman & Kaye, *supra* note 269; Kaye *supra* note 271.

274. Michael O. Finkelstein & William B. Fairley, *A Bayesian Approach to Identification Evidence*, 83 HARV. L. REV. 489 (1970). For $L = 90{,}000$, the posterior probability approaches one for all but invisible values of P(S). For example, the prior probability would have to be about 1/100,000 or less to keep the posterior to less than one-half. On the other hand, for smaller likelihood ratios the graph responds to P(S) over a broader range. Consider the match-binning frequency of 1/17 recomputed in *Yee* according to the ceiling method. *See supra* note 148 and 151. If this frequency were used to form the likelihood ratio $L = 17$, as described *supra* note 253 the graph would look like this:



275. The main drawback of the variable-prior-odds implementation is that it does not necessarily incorporate the risk of laboratory error. This is not wrong, as long as one make it clear what likelihood is being estimated. The concern is that the emphasis on the numbers that are available may lead the jury to overlook this consideration. As with the likelihood ratio or other probabilities, however, the most reasonable response is to insist that no DNA results be admitted without information on the rate of false positives as determined by external proficiency testing.

as the posterior odds of guilt. Of course, when L is immense enough to swamp any plausible prior probability, these inferential errors may be of no great moment, but when the most conservative procedures for computing probabilities are used to generate unduly modest values for L, the need for the jury to see how strongly even these underestimates affect a reasonably ascertained prior probability is greatest.[276] The expert should not be precluded from presenting a mathematically valid and possibly revealing explanation of the significance of the fragment lengths.

CONCLUSION

Analysis of DNA samples can produce revealing evidence of identity, but the search for a procedure to convey—intelligibly, accurately and fairly—the probative value of such evidence has proved challenging. The dominant method for assessing the evidential significance of DNA evidence in this country entails a declaration that two samples either do or do not match, followed, in the case of a match, by an estimate of the matching genotype frequency in some reference population. Some of the controversy that surrounds this methodology is specious. For example, with regard to the matching phase, once a match is declared, most arguments about the overbreadth of a match window are misleading.

Other arguments are less easily resolved. Of these, the most prominent and effective in recent litigation is the concern that populations could be structured in ways that seriously vitiate the population frequency estimates. This scientific issue warrants more refined and complete judicial scrutiny than it has received. The question is not whether there is absolutely no structuring.[277] It is not whether there are absolutely no departures from genetic equilibria.[278] It is whether the structuring and the deviations that it induces have an appreciable impact on VNTR genotype frequencies in the relevant population.

There is very little evidence, and certainly no scientific consensus, that the impact is substantial in any known population. But neither are population geneticists and statisticians unanimous in dismissing the concern. Where the reference population is a broad and probably

---

276. *See supra* note 274.

277. Likewise, in *Frye* jurisdictions, the question is not whether the scientific community agrees that there is absolutely no structuring.

278. In jurisdictions that have adopted the *Frye* standard, the question is not whether scientists agree that there are absolutely no such departures.

structured ethnic or racial population, as it is in the bulk of the litigated cases, the population structure objection amounts to the complaint that the basic procedure adds and then multiplies when it should multiply and then add allele frequencies. A striking series of studies show that the resulting differences in genotype frequencies rarely are dramatic and often favor defendants. Where the reference population is itself a subpopulation, however, requiring resort to extreme overestimation procedures, such as the one called for by a committee of the National Research Council, is more defensible. Still, to the extent that it is feasible to produce a series of estimates that show not only the best estimate for the subpopulation, but also how much that estimate could be in error, this solution may not be needed.

Beyond the debate over population structure is an issue that has lead one jurisdiction to eschew probability estimates altogether. Even when a suitable reference population frequency can be computed, it may be misinterpreted. To avoid prejudice, however, it suffices to bar the proponent of the evidence from mischaracterizing the match-binning frequency as the probability of a false positive or the probability of innocence and to apprise the jury of the probability of a false positive. In sum, given the current state of scientific knowledge, match-binning frequencies should be admissible, at least in general population cases, and perhaps in most subpopulation cases as well.

This is not to say, however, that such frequencies are the best way to express the evidential value of DNA testing. To the contrary, the match *vs.* no match decision produces a false dichotomy. There is little difference between samples that almost match and samples that do not quite match. The likelihood ratio, which states how much more probable it is to find the observed degree of similarity when the defendant is the source than when someone else is, overcomes this difficulty. This quantity should be admissible, either in lieu of or in addition to, a match-binning frequency.

Finally, testimony of the probability that the defendant (or someone else) is the source of the DNA in the crime sample should be admissible if the calculations conform to Bayes's rule and if they do not rest on a prior probability that the expert, rather than the jury, has produced. A Bayesian presentation should involve variable prior odds so that jurors can consider the other evidence in the case and are not compelled to accept an expert's prior probability or to force their own beliefs into a mathematical mold.

These conclusions rest on a particular philosophy concerning the role of expert witness and jury. The task of the expert is to assist the jury in evaluating evidence that it would be hard-pressed to understand fully on its own. The task of the jury is to decide what the evidence proves. With DNA evidence, the expertise of the laboratory technician or scientist is needed to explain what lies behind the pattern of dark bands on an autoradiogram. The expertise of the statistician is needed to explain how probable the patterns are under various hypotheses and how these probabilities affect the plausibility of these hypotheses. Unless more inferential errors are likely to arise with the expert testimony than without it, the rules of evidence should not bar experts from providing all or some of this information to a jury. And, where the range of uncertainty can be described, the law should not force an expert to present this information in a manner that always favors one party over another. Because it is far from obvious which method of presentation—match-binning with basic bins, match-binning with overestimation, likelihood ratio, or Bayesian—will prove most helpful to all jurors, the courts should permit the litigants to advance the combination of reasonably computed statistics or probabilities that they deem most suitable. A healthy pluralism is preferable to a rigid catechism.

APPENDIX: THE INTERIM CEILING PROPOSAL

As noted in the body of the article, the NRC committee's quest for a suitably conservative procedure does not stop at the imposition of the 5% lower upper bound. Until subgroup studies are complete, the panel calls for still higher ceilings. It recommends that each "allele" frequency be taken to be the higher of either 10% or the "upper 95% confidence limit" of the frequency seen in the major "race" with the largest frequency.

The upper end of a confidence interval. The proposal to use these vaulted ceilings from racial databases while awaiting the results of subpopulation studies does not flow inexorably from any generally accepted scientific or statistical theory. Indeed, the upper bound of the confidence interval is presented, candidly, as "a *pragmatic* approach to recognize uncertainties in current population sampling."[279] The enumerated "uncertainties" are the sampling method—"the current 'convenience sample' manner"[280]—and "sampling error."[281]

As a response to these concerns, using the upper end of a confidence interval is most peculiar. To see why, one must understand just what a confidence interval is. Contrary to the view expressed by some courts,[282] confidence intervals do not account for any and all errors in estimation. Confidence intervals are one way to address one kind of error. They express the likely range of sampling error in a probability sample—one in which every item sampled has a known probability of being selected. When, and only when, the probability structure of the sample is known can a 95% confidence interval be said to be the result of a procedure that, under repeated sampling, would generate intervals that capture the true value about 95% of the time.

When convenience samples are collected, the laws of probability cannot reveal how the statistics from the samples will behave. The variability and bias arising from convenience sampling are, quite simply, not addressed by confidence intervals, which are directed to the variability in unbiased, random sampling. Computing a confidence interval for a non-probability sample may be a "pragmatic" response to the concern

279. NRC REPORT, *supra* note 15, at 92 (emphasis added).

280. *Id.* at 92.

281. *See supra* text accompanying note 147.

282. *See* Brock v. Merrill Dow Pharmaceuticals, 874 F.2d 307, 312 (5th Cir. 1989), *cert. denied*, 494 U.S. 1046 (1990). *Contra* Michael D. Green, *Expert Witnesses and Sufficiency of Evidence in Toxic Substances Litigation: The Legacy of* Agent Orange *and Bendectin Litigation*, 86 NW. U. L. REV. 643, 666-68 (1992).

over the sampling method, but only in the same sense that using big bins is a pragmatic response to the concern about population structure. The committee embraces the former, while it shuns the latter.

The second justification for the upper end of the 95% confidence interval on each allele frequency fares no better. A confidence interval is a reasonable response to a concern about sample size, but the committee's proposed treatment of these intervals remains peculiar. In a small database, very rare alleles are likely to be underrepresented, and the estimate of any allele frequency is inherently uncertain in the sense that another sample could produce somewhat different estimates. With small databases, then, it might be appropriate to pick an *a priori* lower bound for the rarest alleles.[283] But this could not justify using only the upper end of a confidence interval, especially in broad racial and ethnic databases that are reaching appreciable sizes. Instead, the obvious way to cope with sampling error is to present an interval estimate of the match-binning frequency *P* computed after the best available estimates of each allele frequency are applied in (1) and (2).[284]

The 10% floor on the ceilings. The substitution of 10% for the 5% lower upper bound while awaiting direct studies of population structure "is designed to address a remaining concern that populations might be substructured in unknown ways with unknown effect and the concern that the suspect might belong to a population not represented by existing databanks or a subpopulation within a heterogeneous group."[285] But any concern about the suspect's racial and ethnic identity is misplaced. It bears repeating that the pertinent reference population is not the defendant, but all people, of whatever race and ethnicity, who plausibly might be suspected of leaving the trace evidence.[286] And, the choice of 10% is no more scientific than the earlier choice of 5%. Both rest on an unarticulated balancing of competing policies.

---

283. *See* Chakraborty et. al, *supra* note 90.

284. *See id.*; Weir, *supra* note 45 (emphasizing the fact that a "confidence limit of a product . . . is not the product of the confidence limits").

285. NRC REPORT, *supra* note 15, at 92.

286. *See supra* text accompanying note 155; NRC REPORT, *supra* note 15, at 85.