2008

The Role of Race in DNA Statistics: What Experts Say, What California Courts Allow

David H. Kaye
Penn State Law

Recommended Citation
David H. Kaye, *The Role of Race in DNA Statistics: What Experts Say, What California Courts Allow*, 67 *Sw. L. Rev.* 303 (2008).

THE ROLE OF RACE IN DNA STATISTICS: WHAT EXPERTS SAY, WHAT CALIFORNIA COURTS ALLOW*

D. H. Kaye**

I. INTRODUCTION 304
II. THE DEVELOPMENT OF CALIFORNIA LAW 304
A. The General Acceptance of Multiplying Allele Frequencies 304
B. The Relevance Objection 307
1. *Pizarro* 307
2. *Wilson* 308
3. *Prince* 309
4. Beyond *Pizarro* and *Prince* 310
III. WHY DNA STATISTICS ARE RELEVANT 311
IV. PROBLEMS WITH RACE AS A "PRELIMINARY FACT" 314
A. Foundation Evidence Unrelated to the Defendant 315
B. Other Evidence of Defendant's Guilt as Foundation Evidence 318
V. RACE-BLIND STATISTICS 319

* Editors' note: This Article was presented at the Southwestern University Law Review Rules of Evidence: FRE v. CEC Symposium in Los Angeles on February 2, 2007. It was inadvertently omitted from the issue dedicated to this symposium, 36 SW. U. L. REV. 561-909 (2008).

** Regents' Professor, ASU Sandra Day O'Connor College of Law; Professor, ASU School of Life Sciences; and Fellow, Center for the Study of Law, Science, and Technology. Part of this Article draws on *DNA Probabilities in* People v. Prince*: When Are Racial and Ethnic Statistics Relevant?*, *in* PROBABILITY AND STATISTICS: ESSAYS IN HONOR OF DAVID A. FREEDMAN 289 (Terrence Speed & Deborah Nolan, eds. 2007). I am grateful to David Balding, Jim Crow, George Sensabaugh, and Terry Speed for commenting on a draft of that paper.

I. INTRODUCTION

When a defendant's DNA matches a sample found at a crime scene, how compelling is the match? To answer this question, DNA analysts typically use numbers—"relative frequencies," "random match probabilities" or "likelihood ratios."[1] They compute and present these quantities for the major racial or ethnic groups in the United States, supplying prosecutors with such mind-boggling figures as "one in nine hundred and fifty sextillion African Americans, one in one hundred and thirty septillion Caucasians, and one in nine hundred and thirty sextillion Hispanics."[2]

A line of California cases rejects this established practice on relevance grounds.[3] The theory of these cases is that only the perpetrator's race is relevant to the crime; hence, it is impermissible to introduce statistics about other races.[4] This Article describes these cases—and their recent demise. Relying on the statistical concept of likelihood, it then presents a logical justification for referring to a range of races and identifies some problems with the one-race-only rule.[5] The Article concludes by noting some ways to express the probative value of a DNA match quantitatively without actually referring to variations in DNA profile frequencies among races.[6]

II. THE DEVELOPMENT OF CALIFORNIA LAW

A. *The General Acceptance of Multiplying Allele Frequencies*

Geneticists normally apply a theoretical population-genetics model to derive DNA profile frequencies from empirical estimates of "allele"

1. *See* DAVID H. KAYE, THE DOUBLE HELIX AND THE LAW OF EVIDENCE: CONTROVERSIES OVER THE ADMISSIBILITY OF GENETIC EVIDENCE OF IDENTITY (forthcoming); Jonathan J. Koehler, *When Do Courts Think Base Rate Statistics Are Relevant?*, 42 JURIMETRICS J. 373, 375-76, 388 (2002).

2. People v. Nelson, 48 Cal. Rptr. 3d 399, 404 n.2 (Ct. App. 2006), *rev. granted*, 147 P.3d 1011 (Cal. 2006). A septillion is 10^{24}, so 130 septillion is on the order of 10^{26}. The other figures quoted would have slightly fewer zeroes in them, but they still make Carl Sagan's famous "billions and billions" seem small. *See* CARL SAGAN, BILLIONS & BILLIONS: THOUGHTS ON LIFE AND DEATH AT THE BRINK OF THE MILLENNIUM (1997).

3. *See infra* Part II.B.

4. *See infra* Part II.B.

5. *See infra* Parts III & IV.

6. *See infra* Part V.

frequencies.[7] We need not get bogged down in the formulas to appreciate the key ideas. The simplest imaginable model lumps everybody into a large population in which all people choose their mates independently of the DNA variations (the *alleles*) used in identification testing.[8] In effect, the alleles are randomly shuffled throughout the population, and the expected proportions of people with the various possible combinations of alleles (the *profiles*) are, within various factors of two, the product of the proportions of all the alleles in the large population.[9]

Of course, the population is "structured" in the sense that different racial groups tend to mate preferentially among themselves.[10] This would not affect the profile frequencies if every racial population had the same initial proportions of alleles, but samples categorized by self-identified racial categories show variations in the proportions across races.[11] Geneticists basically assumed that mating was random *within each race* and computed expected proportions within each major race according to the "unmodified product rule"[12] described above. In ordinary "general population cases,"[13] in which the perpetrator might conceivably be from any of the major racial groups, it became conventional to present these profile frequencies for each major "race"—"Black," "Hispanic," and "Caucasian."[14]

For a time in the 1980s and 1990s, the accuracy of the within-race estimates was intensely controversial.[15] Some expert witnesses testified that there were scientific objections to the quality and quantity of the data on

---

7. COMM. ON DNA FORENSIC SCI.: AN UPDATE, NATIONAL RESEARCH COUNCIL, THE EVALUATION OF FORENSIC DNA EVIDENCE 25 (1996) [hereinafter COMM. ON DNA FORENSIC SCI.: AN UPDATE]; *see also* DAVID J. BALDING, WEIGHT-OF-EVIDENCE FOR FORENSIC DNA PROFILES (2005); JOHN S. BUCKLETON ET AL., DNA EVIDENCE INTERPRETATION (2005); IAN W. EVETT & BRUCE S. WEIR, INTERPRETING DNA EVIDENCE: STATISTICAL GENETICS FOR FORENSIC SCIENTISTS (1998).

8. COMM. ON DNA FORENSIC SCI.: AN UPDATE, *supra* note 7, at 90.

9. *Id.* at 90-91.

10. *See id.* at 99 ("Matings tend to occur between persons who are likely to share some common ancestry and thus to be somewhat related.").

11. *But see id.* at 57-58, 94 ("[D]ifferences among individuals within a race are much larger than the differences between races.").

12. People v. Venegas, 954 P.2d 525, 536, 542 (Cal. 1998).

13. *See* David H. Kaye, *DNA Evidence: Probability, Population Genetics, and the Courts*, 7 HARV. J.L. & TECH. 101, 137, 148 (1993) (distinguishing between "general population cases" in which more or less anyone (or anyone within a major racial group) might have committed the crime and "special population cases" in which "the group of people who might have left the crime sample are a narrow and possibly insular subpopulation").

14. *See, e.g.*, *Venegas*, 954 P.2d at 537.

15 *See* JAY D. ARONSON, GENETIC WITNESS: SCIENCE, LAW, AND CONTROVERSY IN THE MAKING OF DNA PROFILING (2007); KAYE, *supra* note 1; *see also infra* notes 16-20.

DNA allele frequencies and that the racial populations could well contain significant substructure themselves.[16] These difficulties could make the profile frequency calculated for the racial population smaller (or, more often, larger)[17] than its true value. Experts recruited by the prosecution conceded that the estimated profile frequencies were just estimates, but they were confident that the unmodified-product-rule computations were approximately correct.[18] Unable or unwilling to take sides in this debate, an independent panel of scientists and other experts appointed by the National Academy of Sciences proposed further research and temporary use of the "modified ceiling principle"[19]—a "conservative" variation on the unmodified product rule.[20]

California appellate courts were divided over the general acceptance of these methods for computing profile frequencies within the major races until relatively recently.[21] In 1998, in *People v. Venegas*,[22] the California Supreme Court determined that the "modified ceiling approach [was] generally accepted by the relevant scientific community as a forensically sound approach for calculating statistical probabilities of random matches of DNA profiles."[23] The next year, in *People v. Soto*,[24] the court concluded from "the clear weight of judicial authority, and the published scientific commentary, that the unmodified product rule . . . has gained general acceptance in the relevant scientific community and therefore meets the [California] standard for admissibility."[25] Yet, the admissibility of these DNA statistics in California was not assured.

16. *See, e.g.*, United States v. Jakobetz, 747 F. Supp. 250 (D. Vt. 1990), *aff'd*, 955 F.2d 786 (2d Cir. 1992); United States v. Yee, 134 F.R.D. 161 (N.D. Ohio 1991), *aff'd sub nom.* United States v. Bonds, 12 F.3d 540 (6th Cir. 1993).

17. *See* Kaye, *supra* note 13, at 142 (arguing that population structure generally favors the defendant in a general-population case).

18. *See, e.g.*, *Jakobetz*, 747 F. Supp. 250; *see also Yee*, 134 F.R.D. 161.

19. *Venegas*, 954 P.2d at 538.

20. The committee regarded the ceiling principle as "conservative" in the sense of being generous to the defendant. *See* COMM. ON DNA TECH. IN FORENSIC SCI., NATIONAL RESEARCH COUNCIL, DNA TECHNOLOGY IN FORENSIC SCIENCE 76-79 (discussing the "Multiplication Rule" ("Product Rule")), 82-83 (discussing the "Ceiling Pnnciple") (1992).

21. The schism began when *People v. Barney*, 10 Cal. Rptr. 2d 731 (Ct. App. 1992), diverged from *People v. Axell*, 1 Cal. Rptr. 2d 411 (Ct. App. 1991).

22. 954 P.2d 525.

23. *Id.* at 552.

24. 981 P.2d 958 (Cal. 1999).

25. *Id.* at 977.

B. *The Relevance Objection*

1. *Pizarro*

A line of California appellate cases continued to reject DNA statistics on distinct relevance grounds. The California Court of Appeal for the Fifth District initiated this misadventure in *People v. Pizarro.*[26] In 1990, Michael Pizarro was convicted of "murder, forcible lewd or lascivious act on a child under 14, and forcible rape" for raping and suffocating his thirteen-year-old half-sister.[27] Vaginal swabs from the girl's body revealed semen that matched Pizarro's DNA type.[28] At trial, an FBI analyst testified that "[t]he likelihood of finding another unrelated Hispanic individual with a similar profile as Mr. Pizarro is one in approximately 250,000."[29] Pizarro appealed, contending that the DNA evidence was inadmissible because the prosecution had failed to demonstrate that the DNA test procedure and the within-race multiplications were generally accepted in the scientific community.[30] The court of appeal remanded for a hearing on this issue.[31] After the hearing, the trial court again ruled that the evidence was admissible and reentered the judgment.[32] A second appeal and a rehearing followed.[33] In its final opinion in 2003, the court of appeal determined that it was improper for the prosecution to have offered an estimate of the frequency of a genotype in the Hispanic population — and no other group[34] — when there was no proof (other than that pointing to Pizarro) that the perpetrator of the crime was Hispanic.[35] For this reason, the court of appeal again reversed.[36]

---

26. 3 Cal. Rptr. 3d 21 (Ct. App. 2003), *rev. denied* (Oct 15, 2003). This case was disapproved of in *People v. Wilson*, 136 P.3d 864 (Cal. 2006). The description here is restricted to a single issue. Some factual complications also are omitted. For a more complete discussion of the case, see D. H. Kaye, *Logical Relevance: Problems with the Reference Population and DNA Mixtures in* People v. Pizarro, 3 LAW PROBABILITY & RISK 211 (2004).

27. 3 Cal.Rptr. 3d at 28-29, 38.

28. *Id.* at 38.

29. *Id.* at 97-98 (internal quotation marks omitted). This characterization of the statistic is ambiguous. *See* Kaye, *supra* note 26, at 213 n.20.

30. 3 Cal. Rptr. 3d at 29.

31. *Id.*

32. *Id.*

33. *Id.*

34. Actually, the prosecution presented frequency estimates in two population groups — Hispanic and Caucasian. But these were chosen because the defendant was said to be half Caucasian and half Hispanic. *Id.* at 97-98. For brevity, this Article refers only to the Hispanic database.

35. *Id.* at 100-05.

36. *Pizarro*, 3 Cal. Rptr. 3d at 106-07.

This result is justifiable. As noted above, the frequencies of DNA alleles vary across self-identified racial or ethnic groups.[37] If the perpetrator could have come from any of several groups, looking to only one such group for a random-match probability could be misleading. Moreover, hearing only one statistic, the jury might jump to the conclusion that the perpetrator must have come from that one racial group.

But the court did not stop here. It announced in dictum that the standard practice of giving a range of frequencies for *all* the major racial or ethnic groups in the United States also would have been unacceptable because the frequency of the genotype in any given racial group is totally irrelevant without proof of the perpetrator's race.[38] According to the court:

> [I]n the absence of sufficient evidence of the perpetrator's ethnicity, any particular ethnic frequency is irrelevant. The problem is . . . one of preliminary fact . . . . It does not matter how many Hispanics, Caucasians, Blacks, or Native Americans resemble the perpetrator if the perpetrator is actually Asian. If various ethnic frequencies are presented to the jury, each will have been admitted without adequate foundation.[39]

The *Pizarro* court insisted that the race of the possible perpetrator was one of those "preliminary facts" that had to rest on "independent proof."[40]

2. *Wilson*

This relevance argument recurred in *People v. Wilson*.[41] "[T]he body of a 13-year-old white girl was discovered on the living room floor of her Vacaville home by her mother and sister" in the early evening.[42] "She had been strangled with a telephone cord" and had "bruises, scrapes, and scratches on her body."[43] Wilson, who was African-American, had scrapes and bruises and blood on his clothing when he was arrested at two a.m. the next morning.[44] DNA matches were found between the victim's DNA and stains on Wilson's clothing, and between stains on the victim and Wilson.[45]

As the Court of Appeal for the First District described the statistics:

> Wilson's genetic profile would be expected to occur in 1 of 96 billion

---

37. *See supra* note 11 and accompanying text.
38. 3 Cal. Rptr. 3d at 104.
39. *Id.* at 104.
40. *Id.* at 34.
41. 21 Cal. Rptr. 3d 102 (Ct. App. 2004), *aff'd*, 136 P.3d 864 (Cal. 2006).
42. *Id.* at 104.
43. *Id.*
44. *Id.* at 105.
45. *Id.*

> Caucasians, 1 of 180 billion Hispanics, and 1 of 340 billion African-Americans [and the victim's] genetic profile would be expected to occur in 1 of 110 trillion Hispanics, 1 of 140 trillion Caucasians, and 1 of 610 trillion African-Americans.[46]

The court deemed these numbers admissible, writing that "[w]hen the perpetrator's race is unknown, the frequencies with which the matched profile occurs in various racial groups to which the perpetrator *might* belong are relevant for the purpose of ascertaining the rarity of the profile."[47]

3. *Prince*

Although the Fifth District's preliminary-fact rule had not spread, it remained potent in its lair. The Fifth District applied it again in *People v. Prince*.[48] Patrick Paul Prince was charged with burglary, assault, and sexual crimes against five victims.[49] In two of the attacks, "DNA matching his was found on a mask that each girl identified as having been worn by her attacker."[50] Comparable evidence was not available in the other crimes, but the prosecution maintained that the modus operandi was so distinctive that all had to have been committed by the same individual.[51] A criminalist found DNA on the mask and discovered that it matched Prince's at nine STR loci.[52] She testified about "a likelihood ratio that compared two different alternative possibilities, i.e., either the individual contributing the known reference sample contributed the evidence DNA and that is why the profiles matched; or the evidence DNA was contributed by some unknown, unrelated individual who happened to have the same DNA profile."[53] Using data from the FBI on three samples of about 200 Caucasians, 200 Hispanics, and 200 African-Americans, she concluded that "for the Caucasian population, the evidence DNA profile was approximately 1.9 trillion times more likely to match appellant's DNA profile if he was the contributor of that DNA rather than some unknown, unrelated individual; for the Hispanic population, it was 2.6 trillion times more likely; and for the

---

46. *Id.* at 106.
47. 21 Cal. Rptr. 3d at 110.
48. 36 Cal. Rptr. 3d 300 (Ct. App. 2005), *rev granted*, 132 P.3d 210 (Cal. 2006), *rev. denied*, 142 P.3d 1184 (Cal. 2006) (in light of the court's decision in *People v. Wilson*, 136 P.3d 864 (Cal. 2006)).
49. *Id.* at 302.
50. *Id.* at 303.
51. *Id.* at 303-04.
52. *Id.* at 309.
53. *Id.*

African-American population[,] it was about 9.1 trillion times more likely."[54]

The jury returned a verdict of guilty on all counts.[55] Prince appealed, primarily on the ground that "this statistical evidence was irrelevant because the prosecution failed to present substantial evidence to prove that the perpetrator was Caucasian, Hispanic, or African-American."[56] The California Court of Appeal applied *Pizarro*, insisting that independent evidence had to establish the racial or ethnic identity of the perpetrator for any statistics on the frequency of a DNA genotype in that racial or ethnic group to be relevant.[57] However, it added an escape mechanism. It reasoned that because there was substantial non-DNA evidence linking Prince, a Caucasian, to the two crimes, the statistic for Caucasians was relevant after all.[58] The court regarded the admission of the statistics for Hispanics and African-Americans as erroneous, but harmless.[59] Despite the perceived errors at trial, the court of appeal thus affirmed the convictions.[60]

4. Beyond *Pizarro* and *Prince*

The relevance analysis in *Prince* and *Pizarro* is logically defective. It is wrong to say, as the *Prince* court does, that "[t]he probative value (hence, the relevancy) of a [DNA] profile's frequency in an ethnic population depends on proof that the perpetrator belongs to that ethnic group."[61] Part III of this Article explains why. It applies a likelihood-based theory of relevance to show that the statistics for various "races" or ethnic groups ordinarily are relevant without regard to the defendant's race or ethnicity. Part III then shows that even if it were true that frequencies in major subpopulations were only conditionally relevant, this condition normally could not be satisfied by other evidence that the defendant is guilty. Instead, evidence that makes it more probable that some *other* member of his racial or ethnic group (as opposed to any other such group) is the culprit would be necessary. Evidence that only singles out the defendant but otherwise

---

54 *Prince*, 36 Cal. Rptr. 3d at 310.

55. *Id.* at 303.

56. *Id.* at 310.

57. *Id.*

58. *Id.* at 325-26.

59. *See id.*

60. More precisely, it affirmed the convictions on ten counts and reversed the convictions on another two counts. *Prince*, 36 Cal. Rptr. 3d at 327. The published portions of the opinion do not explain the basis for the partial reversal.

61. *Id.* at 304.

provides no information about the race or ethnicity of the perpetrator does not fill in the putative logical lacuna. Hence, the court's thought that the other evidence against Prince made the likelihood ratio of 1.9 trillion relevant is at war with its own exclusionary rule.

In a sense, the problem is academic. As of late 2006, the California Supreme Court sided with the First District, and it chose *People v. Wilson*[62] as the vehicle for reviewing the relevance analysis of *Prince* and *Pizarro*. In *Wilson*, the Supreme Court recognized the logical difficulties with the dictum in *Pizarro*. It also granted review in *People v. Prince*,[63] but then dismissed this grant of review "[i]n light of our decision in *People v. Wilson* . . . ."[64] Thus, *Pizarro* and *Prince* are no longer good precedent in California.

But why, precisely, was the Supreme Court correct in rejecting *Pizarro* and *Prince*? The analysis that follows reveals the power of some simple statistical concepts to clarify issues of relevance. The exposition is valuable not only because one never knows whether the Fifth District's relevance rule might appeal to some other courts across the country, but also because any analytical tool that might sharpen the thinking of lawyers and courts is well worth considering.

III. WHY DNA STATISTICS ARE RELEVANT

"'Relevant evidence' means evidence . . . having any tendency in reason to prove or disprove any disputed fact that is of consequence to the determination of the action."[65] Consequently, to the extent that a DNA match tends to prove that the biological trace evidence that apparently originated from the perpetrator of the crime contained the defendant's cells, it is relevant. A DNA match has this tendency when the probability of observing all the matching trace evidence is greater *if* the defendant is the source than *if* someone else is.[66] If E stands for the observations of the DNA profiles, and if D represents the hypothesis that the defendant (D) is the

62. 136 P.3d 864.

63. 132 P.3d 210 (Cal. 2006).

64. 142 P.3d 1184 (Cal. 2006).

65. CAL. EVID. CODE § 210 (West 1995 & Supp. 2008); *cf.* FED. R. EVID. 401 ("'Relevant evidence' means evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence.").

66. *See* D. H. Kaye, *The Relevance of "Matching" DNA: Is the Window Half Open or Half Shut?*, 85 J. CRIM. L. & CRIMINOLOGY 676, 683-84 (1995); Richard Lempert, *Some Caveats Concerning DNA as Criminal Identification Evidence: With Thanks to the Reverend Bayes*, 13 CARDOZO L. REV. 303, 320 (1991).

source, while *I* represents the hypothesis another unrelated individual is the source, then the evidence *E* is relevant to prove *D* if and only if the conditional probability of the evidence *E* on the hypothesis *D* (that is, $P(E|D)$) is greater than the corresponding probability on the alternative hypothesis (that is, $P(E|I)$).[67] In other words, *E* is relevant (and tends to prove *D*) if and only if the ratio of these two quantities (the likelihood ratio *LR*) is greater than one:[68]

$$LR = \frac{P(E \mid D)}{P(E \mid I)} > 1 \qquad (1)$$

But how is a court to know whether a likelihood ratio exceeds 1? This is where the racial frequencies come in. If the evidence is very likely to arise when D is the source and very unlikely when someone else is, then the numerator in (1) is much greater than the denominator, and the likelihood ratio is much greater than 1. *E* is then not only relevant but also highly probative.[69] Putting to the side such complications as the possibility of cross-contamination that would produce a false-positive match, as I do throughout this article, if the matching DNA profile is very rare among individuals other than D, then the denominator is small, and condition (1) holds.

This analysis reveals when and why a DNA match is relevant, but it does not yet explain why any statistics are also relevant. They are not relevant as explicitly defined in Rule 210, but they are relevant nonetheless as an aid to the understanding of evidence that meets the Rule's definition.[70] DNA statistics possess this "explanatory" or "secondary" relevance because they enable the jury to appreciate the probative value of the primary evidence, the DNA match. That is, like a chart or map that is admitted to help the jury understand a witness's testimony about a material fact, these

67. *See* Richard O. Lempert, *Modeling Relevance*, 75 MICH. L. REV. 1021, 1025 (1977).

68. *Id.* at 1025-26.

69. Arguments for using likelihood ratios to measure probative value are presented in D. H. Kaye & Jonathan J. Koehler, *The Misquantification of Probative Value*, 27 L. & HUM. BEHAV. 645 (2003). The likelihood ratio is commonly employed as a component of a Bayesian analysis. *See, e.g.*, Peter Donnelly & Richard D. Friedman, *DNA Database Searches and the Legal Consumption of Scientific Evidence*, 97 MICH. L. REV. 931 (1999). However, one can maintain that the likelihood ratio expresses the strength of evidence without endorsing or subscribing to Bayesian methods for statistical inference. *See, e.g.*, RICHARD M. ROYALL, STATISTICAL EVIDENCE: A LIKELIHOOD PARADIGM 172-73 (D.R. Cox et al. eds., 1997).

70. For the text of CAL. EVID. CODE § 210, *see supra* note 65 and accompanying text.

numbers are admissible because they help the jury comprehend the meaning of the primary fact that two samples of DNA match.

To establish that the usual DNA statistics tend to show how probative the DNA match is, it may be helpful to consider a hypothetical case in which the relevant population is unusually clear. The circumstances are such that exactly one thousand unrelated men — and no one else — could have left the incriminating DNA sample. This group is the relevant population. Let us assume that 1/2 of the men are Caucasian, 1/8 are Hispanic, 1/8 are African-American, and the remaining 1/4 have other ancestries. One of these men, designated *D*, is tested. Lo and behold, the DNA matches. We may denote this event as M_D. No one else is tested, but *D* is charged with the crime.

What must the jury know to assess the strength of this evidence, as expressed by the likelihood ratio in equation (1)? First, it needs the numerator, the chance of the match if *D* is indeed the source.[71] For an error-free test, this is $P(E|D) = P(M_D|D) = 1$.[72] The denominator is trickier. It is the probability that D would match if *someone else* were the source of the DNA.[73] Assuming that *D* was not selected for testing based on his genotype (or anything correlated to it), he can be regarded as a random draw from the relevant population.[74] Alas, we do not know the actual frequency of alleles or profiles in this group of one thousand. After all, if we knew all the profiles, we would not need any statistics. If the incriminating profile were found only once in the relevant population, the one individual with that profile would have to be the source. If more than one of the thousand men had the profile, then the DNA evidence would incriminate each of them equally — just as DNA evidence alone incriminates two monozygotic twins to the same degree.[75]

Nevertheless, as far as DNA types go, it may be reasonable to regard the relevant population of one thousand as a random sample of the general male population. The chance that a randomly selected member of a random sample has a given trait is just the frequency of the trait in the population.[76] Therefore, on the hypothesis of innocence, the probability that D would have the misfortune of sharing the actual assailant's DNA profile is simply the frequency f of this profile in the general population. Hence, $P(E|I) = f$,

71. *See* Lempert, *supra* note 67, at 1023-24; *see also* COMM. ON DNA FORENSIC SCI.: AN UPDATE, *supra* note 7, at 128.

72. *See* Lempert, *supra* note 67, at 1024.

73. *See id.*

74. COMM. ON DNA FORENSIC SCI.: AN UPDATE, *supra* note 7, at 127.

75. *See* COMM. ON DNA TECH. IN FORENSIC SCI., *supra* note 20, at 3-4, 144.

76. COMM. ON DNA FORENSIC SCI.: AN UPDATE, *supra* note 7, at 127.

and we conclude that the likelihood ratio is:

$$\frac{P(E \mid D)}{P(E \mid I)} = \frac{1}{f} \tag{2}$$

Consequently, the probative value of the match turns on f, the profile frequency in the general male population. This statistic would be relevant not for its own sake, but because it helps explain the meaning of E. But which general male population is involved? The entire country? A city? A state? DNA analysts are able to finesse such questions because DNA allele frequencies, classified by race, show largely the same distributions in collections of DNA samples from many sources including blood banks, paternity-testing laboratories, and convicted offenders.[77] With the allele frequencies in such samples, the "unmodified product rule" leads to a list of race-specific frequencies or likelihoods such as those in *Prince*.[78] Their logical relevance comes from their ability to indicate the magnitude of f. The legal theory is that the jury can make better use of the evidence of a DNA match if it knows how rare the profiles are in several major population groups (or even one such group) than if it has *no* frequency information and must resort to uninformed guessing.[79] The objection that the figures include races other than the perpetrator's misses the explanatory purpose for which the statistics are introduced.[80] The very fact that the frequencies are tiny in one race after another assists the jury in deciding what evidentiary value to attach to the finding of a match. Thus, the usual statistics possess explanatory relevance.

IV. PROBLEMS WITH RACE AS A "PRELIMINARY FACT"

The explication of explanatory relevance in Part III might seem like a tortuous way to reach the intuitively obvious conclusion that the frequency in the entire population of plausible suspects is relevant. If so, I can only say that the California courts could not agree on this fundamental point. The *Prince* court of appeal wrote that:

> [A]ny [frequencies] that do not [come from] the perpetrator's racial group are irrelevant, of themselves, to establish that the defendant is likely the

77. *See id.* at 58.

78. *See Venegas*, 954 P.2d at 542.

79. *See* COMM. ON DNA FORENSIC SCI.: AN UPDATE, *supra* note 7, at 166-67.

80. *See id.* at 187-88.

> perpetrator.... The selection of three individual ethnic databases, even assuming they represent the three largest population groups, is insufficient for this purpose because they have no value independent of the ethnicity of the perpetrator. All such evidence tells the jury is that the DNA profile is statistically rare in those population groups. It neither excludes nor includes the perpetrator as a member of any of those groups, nor does it specifically identify the defendant as being in the same population group as the perpetrator.[81]

This is not a claim that the frequency evidence is relevant but inadmissible because it will confuse the jury. It is not a claim that the population genetics model underlying the estimate is speculative or not generally accepted. It is not a claim that categories like "Caucasian," "Hispanic," and "African American" are socially constructed and devoid of essential meaning. It is not a claim that the frequencies fail to account for the probability of laboratory or handling error. The court did not bar the introduction of the racial and ethnic frequencies on any of these possible grounds. Rather, it asserted that the frequencies are *logically* irrelevant unless another fact first is established on the basis of other evidence: "The probative value (hence, the relevancy) of a profile's frequency in an ethnic population depends on proof that the perpetrator belongs to that ethnic group."[82]

Part III revealed why profile frequencies are relevant without regard to any proof of a perpetrator's race. This section shows that odd consequences follow from efforts to apply the *Pizarro* rule of conditional relevance. The rule could allow the foundational fact to be established in two ways. There might be evidence unrelated to the defendant that proves the perpetrator's race; or, the defendant himself might provide the link in that other evidence of his guilt would imply that it was someone (namely, the defendant) of his race who committed the crime. Each type of proof of the preliminary fact is awkward at best.

A. Foundation Evidence Unrelated to the Defendant

In a substantial number of cases, some evidence of the perpetrator's ethnicity will be available. If an eyewitness has seen a somewhat dark-skinned individual, the court's reasoning should permit the African-American and Hispanic frequencies to be admitted (but not, one might

81. 36 Cal. Rptr. 3d at 313-14.

82. *Id.* at 304; *see also id.* at 308 ("[T]he perpetrator's ethnicity should have been established independently.").

argue, the Caucasian frequency). However, racial identifications are fallible, and there is a broad range of skin coloration within all the racial groups. Therefore, even with an eyewitness to the crime, "[i]t is not always easy to say when the race of the perpetrator is 'known.'"[83] Moreover, even when the evidence is powerful, a defendant should be permitted to argue that a person of another race is responsible.[84] When the prosecution and the defense have competing racial hypotheses, the frequency of the DNA profile *in both groups* is informative. The Supreme Court in *Wilson* made this point forcefully:

> Excluding probability evidence about any but the most likely group could deprive the jury of potentially crucial evidence. If, for example, the jury believed it 51 percent likely the perpetrator was Caucasian, providing it with the probability only for Caucasians would leave it uninformed regarding the 49 percent possibility the perpetrator was of some other population group.[85]

Even the *Prince* court itself ultimately seems to reject its own demand for independent proof of the perpetrator's ethnicity.[86] It expands "the perpetrator's population" to encompass "the general population" when it writes that "it is not enough to show that the genetic profile is rare in a certain number of ethnic populations."[87] Instead, it must be shown either that the genetic profile is rare in the perpetrator's racial group *or in the general population*, as these are the perpetrator's populations."[88]

If the court is serious about allowing proof of the frequency "in the general population,"[89] then it would seem that the fact that "the genetic profile is rare" in a range of ethnic populations shows that it is rare in the general population. The court, nonetheless, suggests that something is lacking—"a showing that [the frequencies] represent not just major population groups but the general population as a whole."[90] Unfortunately, the opinion does not reveal what might constitute the requisite showing,

83. *Wilson*, 21 Cal. Rptr. 3d at 110, *aff'd*, 136 P.3d 864 (Cal. 2006). The Court of Appeal added that "[f]or instance, in *People v. Soto*, . . . a rape victim described her attacker as a white man with light hair. Soto, a neighbor of the victim who was ultimately convicted based at least partly on DNA evidence, was Hispanic with a dark complexion and black hair." *Id.*

84. *Wilson*, 136 P.3d at 870 ("If a defendant wanted to argue that the perpetrator might have been a member of a particular population group for which the odds were more favorable to the defense, surely it would be relevant and permissible to admit evidence of those odds.").

85. *Id.*

86. 36 Cal. Rptr. 3d at 313.

87. *Prince*, 36 Cal. Rptr. 3d at 306 n.17, 322.

88. *Id.* at 313.

89. *Id.* at 324.

90. *Id.* at 323.

although it alludes to the possibility of expert testimony.[91] Presumably, an expert might opine that inasmuch as the general population is composed of major population groups, the figures for these latter groups determine the figure for the population as a whole. But this is an algebraic truth rather than a substantive insight. If each major population group represents a proportion p_j of the total population, with a profile frequency f_j in each group, then the frequency f in the population is the sum these frequencies weighted by how often they occur: $f = \sum p_j f_j$. Because each p_j is between 0 and 1, this weighted average must lie between the largest and the smallest f_j. Inasmuch as courts can take judicial notice of mathematical theorems,[92] expert testimony is not even necessary when the "major population groups" exhaust the population.

In *Prince*, however, the criminalist did not give frequencies for every general racial population.[93] Only Caucasian, Hispanic, and African-American frequencies were presented.[94] Even so, if these estimates—each in the trillionths—are credible, then the frequencies in the omitted groups would have to be enormously higher for the incriminating genotype to be anything other than "rare . . . in the general population."[95] Consequently, the three numbers surely had some tendency to prove that the DNA match was unlikely to be the result of coincidence. Indeed, census data indicate that they arguably covered 94% of the population of Kern County, where the crimes occurred.[96] Even if the profile frequency among the people not included in the three main groups were a thousand times larger than the largest frequency in those groups, the frequency for the general country population would be only about 3 in 100 million.[97]

---

91. *Id.* at 312.

92. The doctrine of judicial notice permits a court to find an indisputable fact without hearing any testimony or other evidence and to instruct a jury in a criminal case that it may do likewise. For example, courts have taken judicial notice of historical facts, such as the dates on which wars began and ended, and of geographical facts, such as the boundaries of states, counties, and townships. 2 MCCORMICK ON EVIDENCE 378 (John Strong ed., 5th ed. 1999).

93. 36 Cal. Rptr. 3d at 322.

94. *Id.* at 309.

95. *Id.* at 317.

96. The opinion does not specify the locations of the residences of the five victims, but the DNA samples from the two victims were analyzed at the Kern County laboratory, and the trial was held in Kern County. *Prince*, 36 Cal. Rptr. 3d at 309. Table 1 indicates that expanding the boundaries beyond Kern County would not alter materially the racial and ethnic population proportions. *See infra* Appendix, tbl.1 at p. 322.

97. Using the percentages for Kern County in Table 1, Table 2 shows the profile frequency for a range of values of the frequency in the 6% of the population not covered by the criminalist's statistics. *See infra* Appendix, tbl.2, at p. 322.

B. Other Evidence of Defendant's Guilt as Foundation Evidence

Can the prosecution pull itself up by the defendant's bootstraps to create the foundation to admit DNA statistics for defendant's race? The argument might go like this: "Your Honor, we have ample evidence of the perpetrator's race. The race is the same as the defendant's. We know this because of the following non-DNA evidence that points to defendant: . . . ." Although this might seem like a cheap trick, the court in *Prince* itself used precisely this reasoning.[98] It maintained that the preliminary fact of the perpetrator's race was proved by evidence that linked the defendant and no one else in particular to the rapes:

> In the present case, there was no direct evidence of the perpetrator's ethnicity. However, direct evidence, such as a description from a percipient witness, is not the only means of establishing the preliminary fact. Instead, the requisite fact can also be established through other independent evidence (evidence not dependent upon the profile match, match frequency, or the defendant's ethnicity per se) that the defendant is the perpetrator. The logic is as follows: If independent evidence establishes that the defendant more likely than not is the perpetrator, and the defendant is Caucasian, then independent evidence establishes . . . that the perpetrator more likely than not is Caucasian. The preliminary fact of the perpetrator's ethnicity is thus sufficiently established so that match frequency statistics, computed from a Caucasian database, are relevant to prove the defendant's identity as the perpetrator.[99]

The "independent evidence" that the court deemed sufficient to establish the preliminary fact of the perpetrator's race was:

> [T]he actual evidence . . . that defendant is the perpetrator . . . . [E]vidence presented by means of victim testimony showed that the perpetrator wore a ski mask and used a flashlight with a colored lens. Appellant possessed such items; therefore, the evidence establishes that appellant could be the perpetrator.[100]

And so, the court concluded that:

---

98. 36 Cal. Rptr. 3d at 325.

99. *Id.* (referring to the standard in CAL. EVID. CODE § 403 (West 2008) that determines when there is sufficient evidence to sustain a finding of preliminary fact).

100. *Id.* at 326.

> [E]vidence [set out] at length in the unpublished portion of our opinion . . . . Suffice it to say that items seized from appellant's residence and car, as well as appellant's reaction when asked to give a buccal swab, are pertinent to our analysis. Although none is sufficient standing alone, when considered cumulatively, they meet the . . . standard.

Id.

> Since the record sufficiently establishes that appellant is the perpetrator and therefore shares the perpetrator's race, and appellant is Caucasian, the profile frequency statistics derived from the Caucasian database, as testified to by [the criminalist], were relevant. The evidence with respect to the other ethnic databases was not; however, its admission did not prejudice appellant.[101]

This reasoning could be applied in virtually every case in which some evidence beyond the DNA match incriminates the defendant.[102] Such "independent evidence" need not be extensive.[103] It need only "be sufficient evidence to enable a reasonable jury to conclude that it is more probable that the fact exists than that it does not."[104] The fact that the court that originated the preliminary-evidence rule embraced a defendant-based method to satisfy it speaks volumes. Because the rule lacks a meaningful justification, it is difficult to apply sensibly. In contrast, the rule that well grounded statistics from a variety of racial groups are admissible to convey the probative value of a DNA match in a general-population case makes good sense. But it is not the only way to indicate to the jurors the probative value of the fact of the DNA match. The remaining section enumerates some "race-blind" methods.

V. RACE-BLIND STATISTICS?

The *Prince* court seemed drawn to dispensing with statistics for races entirely. It wrote that:

> [W]hen the perpetrator's ethnicity is unknown, the most appropriate solutions would appear to be (1) to present the one most conservative frequency, without mention of ethnicity, or (2) assuming this method is scientifically valid and results in a frequency that is considered conservative, to present a single frequency calculation based on a general, nonethnic population database. When frequency calculations that do not reference ethnicity are employed, the profile frequency evidence no longer tells jurors that if the defendant and the perpetrator share ethnicity, the likelihood the defendant is the perpetrator is some number. Instead, the evidence tells jurors that regardless of the perpetrator's ethnicity, the likelihood the defendant is the perpetrator is some number. Jurors then have to decide whether, in their minds, the genetic profile is sufficiently

101 *Id.* (citing People v. Watson, 46 Cal. 2d 818, 836 (1956)).

102. *See id.* at 325-26.

103. *Id.*

104. *Prince*, 36 Cal. Rptr. 3d at 325.

rare so as to be persuasive as to identification of the defendant as the perpetrator.[105]

Although we have seen that these approaches are not required by the law of evidence, there is a certain appeal in simply describing the significance of the DNA match in terms of a statistic that does not refer to any particular races. A few comments on the court's proposal, and of other possibilities, therefore are in order.

The court's first approach is a less extreme version of the "modified ceiling principle"[106] for computing profile frequencies (with less of a guarantee of producing a ceiling). Of course, a single number suppresses the information on variability that comes from hearing how much the profile frequencies differ for three or four races. On the other hand, it is doubtful that the jury will find the variations of much importance, at least when the set of frequencies are in the millionths or beyond. Consequently, this proposal is not unreasonable.

The second approach, involving a racially blended database, is also possible. In effect, one would pool the reference databases for the various races, then use the allele frequencies in the pooled database to obtain a profile frequency estimate. However, the population supposedly represented by the pooled reference database is not randomly mating.[107] An adjustment for population structure would be needed. The procedure is more complicated than the unmodified product rule, but it is technically feasible.[108]

Finally, the expert could compute the probability that the defendant would match given that the true source is a close genetic relative. Even when the suspect population is composed entirely of unrelated individuals, this number can provide an upper bound on the frequency in that population. Thus, one group of scientists recommended using such a computation for a sibling when "there is uncertainty about the population

105. *Id.* at 313-14 (citation omitted). Of course, no population frequency can express "the likelihood the defendant is the perpetrator." *See, e.g.*, People v. Collins, 438 P.2d 33, 39-41 (Cal. 1968).

106. *See supra* Part II.A; *see also* COMM. ON DNA FORENSIC SCI.: AN UPDATE, *supra* note 7 (concluding that the ceiling principle is not needed); Eric S. Lander & Bruce Budowle, *Commentary: DNA Fingerprinting Dispute Laid to Rest*, 371 NATURE 735 (1994) (defending the ceiling principle).

107. *See supra* Part II.A.

108. *See* NATIONAL COMMISSION ON THE FUTURE OF DNA EVIDENCE, THE FUTURE OF FORENSIC DNA TESTING: PREDICTIONS OF THE RESEARCH AND DEVELOPMENT WORKING GROUP 5 (2000) ("[T]he necessity for group classification could be avoided by using an overall U.S. database and an appropriately increased value of [the population-structure parameter] θ. . . . A θ value of 0.03 would usually be appropriate."), *available at* http://www.ncjrs.gov/pdffiles1/nij/183697.pdf.

substructure, as with isolated tribes or communities, or possible unsuspected relatives."[109] The formulas for this "Sib Method" are well known:

> The conditional match probability for a pair of sibs is determined mainly by simple Mendelian rules and is relatively unaffected by allele frequencies (which may differ among population subgroups) and unsuspected substructure, inbreeding, or presence of relatives. Since no other relatives are as close as sibs, the match probability for sibs provides a rough upper limit for the actual match probability.[110]

One could go a step further and use this conservative approach in every case in which the DNA samples are of adequate size and quality for testing at a large number of loci. Because this likelihood is virtually certain to exceed the match probability in any population of unrelated individuals, it could be presented as an estimate of the frequency in either a general-population or a subpopulation case.

In short, it appears that there are scientifically defensible ways to avoid referring to racial population frequencies or probabilities and to convey a sense of the likelihood ratio in a general-population case. The California and the Federal Rules of Evidence should allow them to be used. But neither evidence code dictates that they must be used in lieu of the established procedure of estimating profile frequencies or probabilities in all of the applicable major races. In recognizing as much, *People v. Wilson* represents a welcome return to normal evidence doctrine.

---

109. *Id.* at 4.

110. *Id.*; *see also* Thomas R. Belin et al., *Summarizing DNA Evidence When Relatives Are Possible Suspects*, 92 J. AM. STAT. ASS'N 706, 707 (1997).

Appendix

Table 1. Population Percentages (2000)[111]

	Kern County	California	U.S.A.
White	49.5	46.7	69.1
Hispanic	38.4	32.4	12.5
Black	6.0	6.7	12.3
Total	93.9	85.8	93.9

Table 2. Kern County 9-locus Genotype Frequency for Possible Frequencies in the Groups Not Included in the Criminalist's Statistics

	Percent of County p_j	Genotype Frequency f_j	Product $p_j f_j$	Total $\sum p_j f_j$
1. White	49.5	5.26×10^{-10}	2.61×10^{-10}	
2. Hispanic	38.4	3.85×10^{-10}	1.48×10^{-10}	
3. Black	6.0	1.10×10^{-10}	6.59×10^{-12}	
4. Other	6.1	$f_l = 5.26 \times 10^{-10}$	3.21×10^{-11}	4.47×10^{-10}
		$10 f_l$	3.21×10^{-10}	7.36×10^{-10}
		$100 f_l$	3.21×10^{-9}	3.63×10^{-9}
		$1000 f_l$	3.21×10^{-8}	3.25×10^{-8}

111. U.S. Census Bureau, State and County Quick Facts, http://quickfacts.census.gov/qfd/states/06000.html & http://quickfacts.census.gov/qfd/states/06/06029.html.