TRAWLING DNA DATABASES FOR PARTIAL MATCHES: WHAT IS THE FBI AFRAID OF?

*David H. Kaye**

DNA evidence is often presented as the "gold standard" for forensic science. But this was not always the case. For years, eminent scientists complained that the estimates of the tiny frequencies of DNA types were unfounded. It took scores of research papers, dozens of judicial opinions, and two committees of the National Academy of Sciences to resolve the dispute by the mid-1990s. Since 2000, however, reports have surfaced of shocking numbers of "partial matches" among samples within large DNA databases, and some scientists have complained that the infinitesimal figures used in court to estimate the probability of a random match are no better than alchemy. To study the partial-match phenomenon further, defendants have sought to discover all the DNA records (with personal identifiers removed) kept in offender databases. The FBI has responded by branding the proposed research as useless and the release of the data as an illegal invasion of privacy. The media have reacted by calling for congressional hearings and, possibly, criminal charges against FBI officials.

This Article reviews the existing research findings and considers the scientific, legal, and ethical objections to disclosure of the DNA data. It concludes that the arguments against further research are unpersuasive. At the same time, it finds that the claims of dramatic departures from the expected numbers of partial matches are exaggerated and predicts that new research will not reveal unknown flaws in the procedure for estimating the chance of a match to an unrelated individual. In view of the importance of DNA evidence to the criminal justice system, this Article recommends using the databases for more statistical research than has been undertaken so far. It also calls for dissemination of the anonymized records for this purpose.

* Distinguished Professor and Weiss Family Scholar, Pennsylvania State University, Dickinson School of Law. This paper was presented at Pennsylvania State University and at the Arizona State University Consortium for Science, Policy & Outcomes. Thanks are due to Bruce Weir, John Buckleton, Charles Brenner, and Marjan Sjerps, who helped in various ways.

INTRODUCTION

Across the globe, many countries have established DNA databases—collections of computer-searchable records of the DNA profiles of suspected or convicted offenders.[1] England started the first national criminal DNA database in 1995.[2] In the United States, the state

[1] For brevity, this Article sometimes refers to all these databases as "offender databases." On the constitutionality of including nonoffender DNA in "offender databases," see David H. Kaye, *Who Needs Special Needs? On the Constitutionality of Collecting DNA and Other Biometric Data from Arrestees*, 34 J. L., MED. & ETHICS 188, 192–94 (2006); D.H. Kaye, *The Constitutionality of DNA Sampling on Arrest*, 10 CORNELL J.L. & PUB. POL'Y 455, 463–81 (2001).

[2] In proportion to the population, the British database (NDNAD) is "the largest of any country: 5.2% of the UK population is on the database compared with 0.5% in the USA . . . . By the end of 2005 over 3.4 million DNA profiles were held on the database—the profiles of the majority of the known active offender population." U.K. Home Office, The National DNA Database, http://www.homeoffice.gov.uk/science-research/using-science/dna-database/ (last visited Sept. 3. 2009). In late 2008, the Home Secretary reported that the number exceeded 4.6 million. Sarah Lyall, *European Court Rules Against Britain's Policy of Keeping DNA Database of Suspects*, N.Y. TIMES, Dec. 4, 2008, at A16. How the authorities will remove records from the database in response to the decision of the European Court of Human Rights in *S. v. The United Kingdom*, 48 Eur. Ct. H.R. 50 (2008), remains to be seen. *See* HOME OFFICE, KEEPING THE RIGHT PEOPLE ON THE DNA DATABASE: SCIENCE AND PUBLIC PROTEC-

and federal databases as combined in the National DNA Index System (NDIS) hold over seven million short tandem repeat (STR) profiles from convicted offenders as well as a growing number of people who were merely arrested or detained.[3] When investigators recover a DNA sample from the scene of a crime, they can search these databases to discover if any of the recorded profiles match. Such "cold hits" from these database trawls have led police to serial rapists and murderers who have long eluded detection.[4] Indeed, even dead men have been "accused" through this technology.[5] In addition, database trawls have considerable potential to solve common property crimes.[6] In one case, an observant police inspector in Finland noticed a dead mosquito in a stolen vehicle.[7] The mosquito's body contained human blood from its last meal. Testing the blood against Finland's database yielded a DNA profile match, giving the police a likely suspect.[8]

These databases have another possible use—as a research tool. When a defendant's DNA matches that from a crime-scene, it is standard practice to introduce the probability of a random match in the general

TION 14–18 (2009), http://www.homeoffice.gov.uk/documents/cons-2009-dna-database (presenting proposals).

[3] *See* Federal Bureau of Investigation, CODIS–NDIS Statistics, http://www.fbi.gov/hq/lab/codis/clickmap.htm (last visited Sept. 3, 2009).

[4] In 2005, a 58-year-old man became a suspect in over 24 rapes in three states dating back to 1973 as a result of a coordinated database search. Associated Press, *DNA Leads to Arrest 32 Years After Attack*, USA TODAY, Apr. 26, 2005, at A3. He was promptly convicted of a 32-year-old rape in New York. Julia Preston, *After 3 Decades, Guilty Verdict in Rape Case, With Help From DNA*, N.Y. TIMES, Nov. 10, 2007, at B1. In 2009, Los Angeles police reported linking a state insurance claims adjuster to two waves of slayings in the 1970s and '80s. Andrew Blankstein & Joe Mozingo, *Suspect May Be Linked to 30 Killings*, L.A. TIMES, Apr. 30, 2009, at A1. Systematic data on the effectiveness of the databases, however, are harder to come by. The FBI reports tens of thousands of "investigations aided," but the actual impact of the information on investigations is not known. *See* FBI, Today's FBI: Law Enforcement Support & Training, http://www.fbi.gov/facts_and_figures/law_enforcement_support.htm (last visited Sept. 6, 2009) (reporting that "[a]s of April 2008, CODIS has achieved 68,860 investigations aided, over 50,000 total offender hits, and more than 12,000 forensic hits").

[5] *See, e.g.*, Associated Press, *Illinois: DNA Match in 1984 Murders*, N.Y. TIMES, Feb. 12, 2009, at A27 (reporting that, on the basis of a search of NDIS, "a man who died more than five years ago killed and sexually assaulted two young Decatur girls on Halloween night in 1984, the police said").

[6] *See, e.g.*, Shaila K. Dewan, *As Police Extend Use of DNA, A Smudge Could Trap a Thief*, N.Y. TIMES, May 26, 2004, at A1; Alison Gendar, *DNA Test Buoys Rob Busts*, N.Y. DAILY NEWS, Mar. 29, 2005, at 25. *See generally* JOHN K. ROMAN ET AL., THE URBAN INSTITUTE, THE DNA FIELD EXPERIMENT: COST-EFFECTIVENESS ANALYSIS OF THE USE OF DNA IN THE INVESTIGATION OF HIGH-VOLUME CRIMES (2008), http://www.urban.org/UploadedPDF/411697_dna_field_experiment.pdf (cataloguing the effectiveness of DNA in property crimes).

[7] *See* BBC News, *Mosquito Blood 'Identifies Thief'*, Dec. 22, 2008, http://news.bbc.co.uk/2/hi/europe/7795725.stm (last visited Sept. 16, 2009).

[8] *See id.*

population.[9] The numbers bandied about in court boggle the mind. Reported match probabilities involve quadrillionths ($1/10^{15}$), quintillionths ($1/10^{18}$), sextillionths ($1/10^{21}$), and even septillionths ($1/10^{24}$).[10] These numbers are smaller than the radius of an electron,[11] and it is easy to be skeptical of such extreme claims. Keith Devlin, a mathematician at Stanford University, calls them "total nonsense" and a "damned lie."[12] In Devlin's view, it is "disgraceful" that courts allow experts to provide such small random-match probabilities: "They may as well admit alchemy and astrology."[13]

There is something to the notion that one should not take the number of zeroes in the random-match probabilities too seriously. It might be wiser for an expert to stop the multiplication at one in a million or one in a billion for the chance of a match with a randomly selected, unrelated individual, or to avoid estimating the random-match probability and simply give the much larger probability of a match between two full sib-

[9] *See* David H. Kaye, *The Role of Race in DNA Evidence: What Experts Say, What California Courts Allow*, 37 SW.U. L. REV. 302, 304 (2008) (discussing the relevance of this statistic, and alternatives to it).

[10] *See, e.g.*, People v. Nelson, 48 Cal. Rptr. 3d 399, 404 n.2 (Cal. Ct. App. 2006) ("[T]his profile would occur at random among unrelated individuals in about 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. There are 21 zeros in a sextillion and 24 zeros in a septillion."), *aff'd*, 185 P.3d 49 (Cal. 2008); United States v. Davis, 602 F. Supp. 2d 658, 680 n.26 (D.Md. 2009) ("[T]he allelic frequencies for the evidentiary samples were calculated to be . . . in the quadrillions to quintillions").

[11] *See* National Institute of Standards & Technology, Fundamental Physical Constants, http://physics.nist.gov/cgi-bin/cuu/Value?re—search_for=classical+electron+radius (last visited Sept. 16, 2009) (listing the classical electron radius as 2.8×10^{-15} m).

[12] *See* Keith Devlin, *Damned Lies*, MAA ONLINE, Oct. 2006, http://www.maa.org/devlin/devlin_10_06.html. Dr. Devlin directs Stanford University's Center for the Study of Language and Information. He recounts how the students in his "evening adult education course at Stanford," consisting of "30 or so experienced and numerically sophisticated scientists, technologists, engineers, and others" burst out laughing when he "displayed on the screen a probability figure that prosecutors typically give in court: 1/15,000,000,000,000,000" for "the probability that a DNA profile match between (say) a defendant and a DNA sample taken from a crime scene was a result of an accidental match, rather than because the crime-scene sample came from the defendant." *Id.*

The number given above, however, is not the probability that a match is the result of the defendant's bad luck in having, quite by accident, the same DNA profile as the true source of the crime-scene sample. *See, e.g.*, DAVID H. KAYE, DAVID E. BERNSTEIN & JENNIFER L. MNOOKIN, THE NEW WIGMORE: A TREATISE ON EVIDENCE: EXPERT EVIDENCE § 12.3.1 (Aspen Pub. Co. 2004) [hereinafter KAYE ET AL, THE NEW WIGMORE] (describing this "transposition fallacy"). It is the probability that if the defendant and the true source are unrelated, then their profiles will, by accident, be the same. Nonetheless, for extremely small probabilities, there may be little practical difference between the two characterizations of the statistic. *See* David H. Kaye, *"False, But Highly Persuasive": How Wrong Were the Probability Estimates in* McDaniel v. Brown*?*, 108 MICH. L. REV. FIRST IMPRESSIONS 1 (2009), http://www.michiganlawreview.org/firstimpressions/vol108/kaye.pdf [hereinafter Kaye, *"False, But Highly Persuasive"*].

[13] Devlin, *supra* note 12.

lings.[14] But Devlin goes further. He points to an informal and unpublished study of an offender database that he thinks is the only one of its kind and that "dramatically" contradicts (or seems to) "the astronomical, theoretical figures given by the naive application of the product rule."[15]

This study, as well as the reaction of the FBI to it, has become notorious—and notoriously misunderstood. Defense lawyers have used the study, with occasional success, to argue that they should have access to convicted-offender databases to put the theoretical estimates to an empirical test.[16] Calling the study "meaningless"[17] and suggesting that it would be illegal and unethical to disclose the assembled DNA data, the FBI has discouraged this initiative—with a very heavy hand. The Bureau reportedly has threatened states with cutting off their participation in the national database system that pools the state and federal data if they release their databases to outside scientists or to defendants.[18] The misperception that the study is a smoking gun for the usual random-match probabilities,[19] combined with the FBI's defensiveness, prompted one prominent law professor to demand "an immediate congressional investigation" that "could raise questions of appeal in hundreds of cases and [could] lead to some FBI officials being fired."[20] Likewise, the *San Francisco Chronicle* branded the FBI's opposition to the use of large, offender databases for population-genetics research "ridiculous and reprehensible" if not "criminal."[21]

This Article examines this controversy. Part I explains why the usual random-match probabilities (RMPs) are relevant in criminal prose-

[14] *See* NATIONAL RESEARCH COUNCIL COMMITTEE ON DNA FORENSIC SCIENCE: AN UPDATE, THE EVALUATION OF FORENSIC DNA EVIDENCE 113 (1996) [hereinafter NRC] (providing formulas indicating that two full siblings possess a greater chance of having inherited the same DNA sequences than do two unrelated individuals).

[15] Devlin, *supra* note 12.

[16] *See* Erin Murphy, *The New Forensics: Criminal Justice, False Certainty, and the Second Generation of Scientific Evidence*, 95 CAL. L. REV. 721, 782 (2007); Edward Ungvarsky, *What Does One in a Trillion Mean?*, GENEWATCH, Feb. 2007, at 10.

[17] Jason Felch & Maura Dolan, *FBI Resists Scrutiny of 'Matches'*, L.A. TIMES, July 20, 2008, at A20.

[18] *See* Murphy, *supra* note 16, at 782–83; Editorial, *Should We Trust DNA?*, SAN FRANCISCO CHRONICLE, July 28, 2008, at B4, [hereinafter *Should We Trust DNA?*]; Edward Humes, *Guilt by the Numbers: How Fuzzy Is the Math that Makes DNA Evidence Look So Compelling to Jurors?*, CALIFORNIA LAWYER, Apr. 2009, at 20–24; Chris Smith, *DNA's Identity Crisis*, SAN FRANCISCO MAG., Sept. 2008, http://www.sanfranmag.com/story/dna%E2%80%99s-identity-crisis; Felch & Dolan, *supra* note 17.

[19] *See* Editorial, *DNA Evidence, What Are the Real Chances of Mistakes?*, LAS VEGAS REV. J., July 29, 2008, *available at* http://www.lvrj.com/opinion/26025944.html.

[20] Jonathan Turley, Res Ipsa Loquitur, *FBI Accused of Trying to Bury Findings that Raised Questions Over DNA Claims*, http://jonathanturley.org/2008/07/23/fbi-accused-of-trying-to-bury-findings-that-raised-questions-over-dna-claims/ (July 23, 2008).

[21] *Should We Trust DNA?*, *supra* note 18.

cutions and how they are computed. Part II discusses the findings that now are said to undermine these computations. It shows that much of the drama in the "dramatic" departure from the theoretical expectations is the product of a cognitive fallacy involving certain probabilities. Part III reviews the existing research and describes what it indicates about RMPs. It concludes that current research does not seriously undermine the vanishingly small theoretical estimates of RMPs. Part IV considers whether existing laws and bioethical principles stand in the way of open research with offender databases to assess the validity of the theoretical RMPs. It maintains that the release of the data, stripped of personal identifiers, for population-genetics research is permissible. In short, this Article suggests that other than the inevitable and already manifest problem of explaining a complex issue in genetics and statistics to lay judges, the FBI has nothing to fear and should reverse its policy of not researching the issue and maintaining the secrecy of the data. Furthermore, even if the predictions about the likely results of appropriate studies with the databases are erroneous, the public and the legal community need to know that all reasonable efforts have been made to verify the accuracy of the numbers that are given to police, prosecutors, judges, and juries.[22] Disclosure of the databases in anonymized form is the best policy.

I. THE "RANDOM-MATCH PROBABILITY"

The random-match probability is the probability that a randomly selected, unrelated individual in the general population (or some part of it) would have a particular DNA profile—the one found in a crime-scene sample. For example, in *People v. Nelson*,[23] a DNA profile from semen on the sweater of a teenager who "had been raped and drowned in mud"[24] had a random-match probability of "one in 930 sextillion (93 followed by 22 zeros)."[25] After discovering that it matched the recorded profile of Dennis Louis Nelson, a convicted sex-offender, the state of California charged Nelson [ ] with the murder. A possible defense argument was that someone unrelated to Nelson is actually the source, and we can designate this hypothesis as H^0. To refute H^0 and support the state's hypothesis that Nelson was the source of the semen (H^1), the prosecution introduced the RMP.[26] Thus, the RMP is relevant not for its own

[22] I have taken this position in letters requested by defense counsel in the District of Columbia and San Francisco.

[23] 185 P.3d 49 (Cal. 2008).

[24] *Id.* at 53.

[25] *Id.* at 52.

[26] *See id.* at 53–54. Other probabilities come into play with respect to other rival hypotheses such as "The defendant is not the source—his uncle is," or "The DNA samples match because the laboratory contaminated the crime-scene sample with the defendant's DNA." For an analysis of other hypotheses that could explain the reported match, see KAYE ET AL, THE

sake, but only as an aid to the jury in assessing the probative value of the circumstantial evidence—the match itself.[27]

In the early years of DNA testing, the computation of the random-match probability sparked an extended controversy.[28] To estimate RMPs, the FBI and other laboratories assembled DNA profiles from a few hundred people.[29] Each profile was so rare that no two individuals in these data sets shared the same profile, but the "alleles"—the variations in the DNA at each of the genetic locations (the "loci") that made up the overall profile—occurred often enough to permit the allele fre-

NEW WIGMORE, *supra* note 12, at § 12.3.1. Moreover, even if the source-level hypothesis is resolved correctly, other reasoning is required to reach a verdict of guilt or innocence. In the Swedish mosquito case, for example, the suspect presented a perfectly innocent explanation—he was in the car as a hitchhiker. *See* BBC News, *supra* note 7. In *Nelson*, the defendant asserted that he "had consensual sexual intercourse" with the victim but proposed "that someone else abducted, raped, and murdered her." 48 Cal. Rptr. 3d at 404.

[27] Articulating how the RMP supports H_1 relative to H_0 is not as simple as it might seem. The RMP bears on the choice between these hypotheses only because it also is a conditional probability—the chance of a (correctly ascertained) match given that the true source is unrelated to the suspect. If we abbreviate this conditional probability as *Pr*(*match* given *unrelated*), or even more tersely, as $P(M—H_0)$, then its connection to the hypothesis H_0 becomes explicit. From the classical hypothesis-testing standpoint, an event that has a very small chance of arising under the "null hypothesis" H_0 justifies the rejection of that hypothesis. It would be surprising to witness the event if the hypothesis were true. *See*, *e.g.*, Persi Diaconis, *Theories of Data Analysis: From Magical Thinking Through Classical Statistics*, *in* EXPLORING DATA TABLES, TRENDS, AND SHAPES 1 (David C. Hoaglin et al. eds., John Wiley & Sons, 1985).

A better theory of why a small RMP supports H_1 relative to H_0 is based on the concept of "likelihood." *See generally* DAVID H. KAYE, THE DOUBLE HELIX AND THE LAW OF EVIDENCE (Harvard University Press, 2010) [hereinafter KAYE, THE DOUBLE HELIX] ; KAYE ET AL, THE NEW WIGMORE, *supra* note 12; Richard Lempert, *Modeling Relevance*, 75 MICH. L. REV. 1021 (1977). From the likelihood perspective, the probative value of the match depends on the relative magnitude of the probabilities for the evidence conditioned on the parties' hypotheses. A match that is far more probable when the defendant is the source than when he is genetically unrelated to the source is strongly probative of identity. One that is more or less equally probable under both hypotheses provides no firm basis for choosing between the hypotheses. It is not particularly probative. This ratio of the conditional probabilities is known as the "likelihood ratio" in probability and statistics. *See*, *e.g.*, RICHARD M. ROYALL, STATISTICAL EVIDENCE: A LIKELIHOOD PARADIGM 3 (Chapman & Hall/CRC, 1997).

The likelihood ratio here is $LR = P(M— H_1) / P(M— H_0)$. The numerator is the chance of a (correctly ascertained) match given that the suspect is the source. Putting aside the possibility of laboratory error or fraud, this probability is 1. The denominator is the RMP, so the smaller the RMP, the larger the likelihood ratio $LR = P(M— H_1) / P(M— H_0) = 1 / \text{RMP}$. A profile that has a random-match probability of one in 930 sextillion, as in *Nelson*, is 930 sextillion times more probable when the suspect is the source than when he is unrelated to the source. The evidence lends enormous support to H_1 compared to H_0. This is why the RMP matters. It is relevant because it gives the judge or jury information about the strength of the scientific evidence.

[28] *See*, *e.g.*, JAY ARONSON, GENETIC WITNESS: SCIENCE, LAW, AND CONTROVERSY IN THE MAKING OF DNA PROFILING (Rutgers University Press 2007); KAYE, THE DOUBLE HELIX, *supra* note 27.

[29] KAYE, THE DOUBLE HELIX, *supra* note 27, at 87.

quencies to be estimated with reasonable precision.[30] The individual allele frequencies then were combined to give an estimate of the profile frequency according to a population-genetics model.[31] In the simplest such model, everyone is assumed to choose mates without regard to their DNA profiles (or any factor that would be correlated with these profiles).[32] In this situation, the profile frequency is, very roughly, the product of the allele frequencies.[33]

Critics of this product-rule calculation pointed out that if the allele frequencies varied across subgroups within the population, and if these subpopulations mated largely among themselves, then the choice of mates could be correlated with DNA profiles.[34] The then-available data, however, demonstrated that strong "population structure" of this kind was unlikely,[35] and subsequent studies confirmed that population structure in the United States does not radically change the estimates of the frequencies of DNA profiles in the major racial or ethnic groups.[36] In addition, an "affinal model" of the population can be used to modify the product-rule so that it accounts for population structure.[37] By the mid-to-late 1990s, courts were persuaded that either basic product rule or the

[30] *Id.* at 87–88.

[31] *Id.* at 89.

[32] *Id.* at 91–92.

[33] A factor of 2 is included for each locus in which the profile has two distinct alleles. These factors of 2 arise because there are two equally probable ways for two random draws from a pool of alleles A and B to produce the pair of alleles A and B. We could draw an A then B, or a B then A. Hence, the probability of an A and a B in an individual is twice the product of the allele proportions. One might think that when the two alleles at any locus appeared to be the same (AA, BB, etc.), the allele frequency would be multiplied by itself, and no factor of 2 would be applied. For the genetic system then in use (known as Variable Number Tandem Repeats, or VNTRs), however, the allele frequency was multiplied only once, and the factor of 2 was used. As compared to a naive product rule, this practice benefited defendants. *See*, *e.g.*, NRC, *supra* note 14.

[34] The argument is analyzed and assessed in KAYE, THE DOUBLE HELIX, *supra* note 27, at 92–139.

[35] *See*, *e.g.*, Bernard Devlin & Kathryn Roeder, *DNA Profiling: Statistics and Population Genetics*, *in* 1 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY 710 (David L. Faigman et al. eds., West Group, 1997); KAYE, THE DOUBLE HELIX, *supra* note 27, at 125–26.

[36] NRC, *supra* note 14. When estimating random match probabilities within a single subpopulation such as Italian-Americans, multiplying allele frequencies derived from the larger population could introduce greater error. D.H. Kaye, *DNA Evidence: Probability, Population Genetics, and the Courts*, 7 HARV J.L. & TECH. 101, 134 (1993). The most difficult case is an isolated group, such as the Inuit people of Alaska's Northwest Arctic and North Slope. The simple product rule applies within the group, but unless samples from this group are available, the allele frequencies will not be known directly. In this situation, a committee of the National Academy of Sciences recommended using estimates from groups with related ancestry for which allele-frequency data are available. *See* NRC, *supra* note 14.

[37] NRC, *supra* note 14.

affinal method produces generally accepted and valid estimates of random-match probabilities.[38]

As a result of the large databases, however, this seemingly settled issue is being revisited. Investigative reporters are asking, "How reliable is DNA in identifying suspects?"[39] Radio talk show hosts are discussing "disturbing doubts about DNA."[40] One law professor is demanding "an immediate congressional investigation."[41] More circumspectly, another simply states that "recent evidence calls into question the accuracy of . . . match probabilities."[42] A defendant charged with robbery, carjacking, and related firearms violations in Maryland produced expert opinion that FBI "statements or inferences of uniqueness [might] be fundamentally incorrect."[43] But are the new doubts justified? The next part traces the finding that prompted the current outcry and uses a well known result in probability theory to show that the concern, while appropriate, is exaggerated.

II. THE IMPLICATIONS OF PARTIAL MATCHES IN DATABASE TRAWLS

A. *The Arizona Experience*

The commotion began after staff from the Arizona Department of Public Safety's DNA laboratory posted an announcement of "A Nine STR Locus Match Between Two Apparently Unrelated Individuals" at an annual scientific meeting on DNA identification methods in Phoenix in 2001.[44] It is not clear from published accounts how this nine-locus match—between a white and a black man with felony convictions—was discovered.[45] The database records consisted of a list of the alleles at thirteen distinct loci. When the same thirteen loci can be typed in a crime-scene sample, a mere nine-locus match will not generate a suspect. In fact, the discrepancies in the full profile at the other four loci will

[38] *See* KAYE, THE DOUBLE HELIX, *supra* note 27, at 139.

[39] Felch & Dolan, *supra* note 17.

[40] Airtalk, *How Reliable is DNA Testing?*, Southern California Public Radio, July 23, 2008, available by contacting Southern California Public Radio at http://www.scpr.org/contact.

[41] Turley, *supra* note 20.

[42] Murphy, *supra* note 16, at 781.

[43] *United States v. Davis*, 602 F. Supp. 2d 658, 681 (D.Md. 2009) (quoting Laurence Mueller, but rejecting this argument).

[44] *See* Kathryn Troyer et al., Lab Analysts, Arizona Department of Public Safety, *A Nine STR Locus Match Between Two Apparent Unrelated Individuals Using AmpFlSTR Profiler Plus™ and Cofiler™*, PROCEEDINGS OF THE PROMEGA 12TH INTERNATIONAL SYMPOSIUM ON HUMAN IDENTIFICATION (2001).

[45] According to Murphy, *supra* note 16, at 782, "an alert analyst happened to observe" it. Smith, *supra* note 18, states that Troyer was "doing a routine check of the state's criminal offender database." The director of the state crime laboratory did not respond to my request for clarification. Probably, the finding resulted from an expanded version of a standard check to see, by comparing all the profiles in the database, whether there were any duplicates.

exclude a suspect as a possible source of crime-scene DNA. But even partial matches at nine loci are almost never encountered in case work. Indeed, according to one report, the RMP for the matching nine-locus genotype in Arizona was "1 in 754 million in Caucasians, 1 in 561 billion in African Americans, and 1 in 113 trillion in Southwest Hispanics."[46] The discovery of such a match in the state database surely seemed anomalous.

In 2005, Bicka Barlow, a public defender in San Francisco, contacted Kathryn Troyer, the DNA analyst who came across the partial match in the Arizona database.[47] Barlow, herself a "molecular biologist turned lawyer,"[48] was defending a man in a California case in which the state had only typed nine loci.[49] Troyer told Barlow that she had found more nine-locus matches.[50] Barlow asked Troyer to send her more information.[51] Todd Griffith, the head of the Arizona Department of Public Safety lab, interceded.[52] He notified Barlow that no further information would be released. Barlow applied to an Arizona court for a subpoena.[53] At a hearing in Phoenix, Troyer stated that she had found "approximately 90" nine-locus, partial matches. Barlow was astounded. "'I almost fell over when I heard that,' Barlow says now, with a laugh. 'I was thinking she had 10 matches, or 20. That would have been huge, right?'"[54] The Arizona lab produced no testimony to counter Barlow's representation to the Arizona judge that the 90 matches meant that "the FBI's population-rarity statistics are 'only an estimate, and the estimate is wrong.'"[55] The court ordered the lab to produce a summary of the numbers of partial hits at nine or more loci among all pairs of people in the Arizona convicted-offender database. The report stated that 122 pairs matched at nine loci (and did not match at the other four loci); another 20 pairs matched at ten loci (and did not match at the remaining three loci).[56]

[46] Murphy, *supra* note 16, at 782. The figure of 1 in 113 billion is mentioned in other accounts. *See, e.g.*, Charles Brenner, ARIZONA DNA DATABASE MATCHES, Jan. 8, 2007, http://dna-view.com/ArizonaMatch.htm.

[47] *See* Smith, *supra* note 18.

[48] *Id.*

[49] *See* Murphy, *supra* note 16, at 782.

[50] *See* Smith, *supra* note 18.

[51] *See id.*

[52] *See id.*

[53] *See id.*

[54] *Id.*

[55] *Id.*

[56] The crux of the report, as described by Ungvarsky, *supra* note 16 at 12, is that the offender database "profiles of . . . 65,493 persons . . . had 122 pairs of people who matched at 9 out of the 13 loci, 20 that matched at 10 loci, 1 that matched at 11 loci, and 1 that matched at 12 loci. The last two matches were confirmed to be between pairs of siblings." *See also* Bruce Budowle et al., *Partial Matches in Heterogeneous Offender Databases Do Not Call into Question the Validity of Random Match Probability Calculations*, 123 INT'L J. LEGAL MED. 59, 61 (2009).

Lawyers, and at least one mathematician, have found this number to be "remarkable"[57] and "startling."[58] If the RMP for a nine-locus match is anything like "one in 754 million for whites, and one in 561 million for blacks,"[59] how can it be that a database as small as "a mere 65,493 entries"[60] produces even one such match? "Scary isn't it?" asked Devlin.[61]

B. *The Combinatorial Explosion: All-pairs Trawls and the Birthday Problem*

This scenario is only scary if we conflate the size of the database with the number of comparisons being made to find a match.[62] Three distinct situations for DNA matches can arise, and no end of confusion results if they are not disentangled. They are pictured in Figure 1.

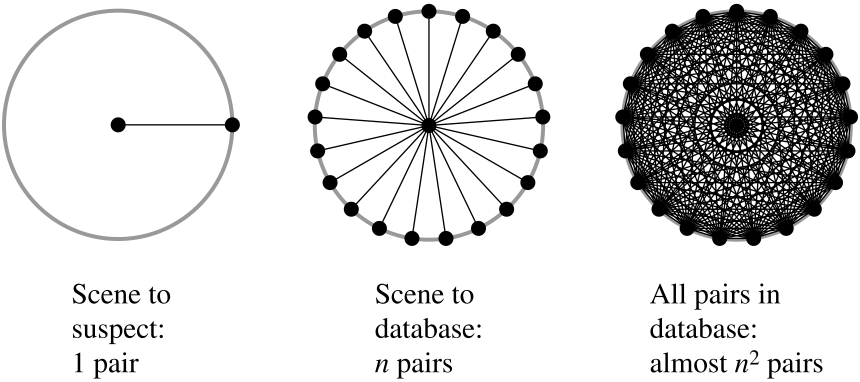
Scene to suspect: 1 pair
Scene to database: n pairs
All pairs in database: almost n^2 pairs

Figure 1. Comparisons in a confirmation case, an ordinary database trawl, and an all-pairs database trawl. Each line represents a comparison. There is 1 comparison in the first situation, n in the second, and about half of n^2 in the third.

First, a confirmation case involves a one-to-one search. The crime-scene profile is compared to the suspect's profile. This is shown by the single line from the crime-scene sample to the suspect in the left-hand panel of Figure 1. The probability of a match when an unrelated individual is the source of the crime-scene DNA is some number p. This is the random-match probability. Second, a database trawl is a one-to-n search for a match to the crime-scene profile among the n profiles in the

[57] Ungvarsky, *supra* note 16 at 12.

[58] Devlin, *supra* note 12.

[59] Smith, *supra* note 18. Professor Murphy provides different numbers. *See* Murphy, *supra* note 16, at 782 ("Under the statistical models then in place, a person picked at random would match that nine loci profile at a rate of 1 in 754 million in Caucasians, 1 in 561 billion in African Americans, and 1 in 113 trillion in Southwest Hispanics.").

[60] Devlin, *supra* note 12.

[61] *Id.*

[62] *See* Brenner, *supra* note 46.

database. Because there are *n* independent comparisons, as shown by the lines radiating from the crime-scene sample at the center of the circle of database samples in the middle panel, these trawls can be expected to produce approximately *np* matches when no one in the database is the source and no one is related to each other or to the source. Finally, in a trawl through all possible pairs in a database, every profile in the database is compared with every other profile. Instead of the *n* comparisons to a single crime-scene profile in the simple database trawl, the all-pairs trawl entails nearly n^2 comparisons. This combinatorial explosion, shown by the crisscrossing lines in Figure 1, creates a vastly greater number of opportunities for a match among profiles. Hence, the database need not be so huge before one can expect many matches that have very small random-match probabilities. One-to-one comparisons (the testing of a known suspect) and one-to-*n* searches (for a cold hit) are markedly different from large *n*-to-*n* comparisons (the all-pairs trawls), although commentary confusing the latter two types of searches is all too common.[63]

All-pairs trawling—an artificial form of searching that is not used in criminal investigations—is analogous to the famous "birthday problem."[64] The problem is to determine the minimum number of people in a room such that the odds favor there being at least two of them who were born on the same day of the same month.[65] In its simplest form, the birthday problem assumes that equal numbers of people are born every day of the year. Since the random-match probability for a specified birthday is about 1/365, most people think that more than 180-some people must be in the room.[66] Indeed, one might think that for a match to be likely, the number should be larger still. After all, the chance of a match between two randomly selected individuals having a given birthday (say, January 1) is a miniscule 1/365 × 1/365 = 1/133,225.

But a precise calculation shows that it takes only 23 people before it is more likely than not that at least two people in the room share a birthday.[67] The actual number is this small because the matching birthday can be any one of the 365 days in the year and because the number of

[63] *See, e.g.*, Humes, *supra* note 18, at 20 (misreporting that "the birthday problem" caused the "the odds of a coincidental match" in an ordinary cold-hit case to go from 1 in 1.1 million to "a whopping 1 in 3").

[64] *See, e.g.*, BRUCE BUDOWLE ET AL., CLARIFICATION OF STATISTICAL ISSUES RELATED TO THE OPERATION OF CODIS, PRESENTATION TO THE 17TH INTERNATIONAL SYMPOSIUM ON HUMAN IDENTIFICATION 6 (2006), http://www.promega.com/GENETICIDPROC/ussymp17proc/oralpresentations/budowle.pdf [hereinafter CLARIFICATION] (using the "birthday problem" to illustrate the high probability of DNA profile matches).

[65] *See id.*

[66] *See id.*

[67] *See, e.g.*, Persi Diaconis & Frederick Mosteller, *Methods for Studying Coincidences*, 84 J. AM. STAT. ASS'N 853, 857 (1989).

comparisons among birthdays scales as n^2 with an increasing number n of people in the room.[68]

Thus, all-pairs trawling in databases makes it much less surprising to come across partial (or even full) matches among unrelated people than in ordinary casework.[69] Just think about how many distinct pairs can be formed, and then compared, with the "mere 65,493 entries"[70] in the Arizona database. For example, number each individual 1 through 65,493. Number 1 can be paired with number 2, 3, and so on. That is 65,492 pairs right there. Number 2 can be paired with number 3, 4, and so on. That is another 65,491 pairs. The exact formula for the number of distinct pairs of n items is $n(n–1)/2$.[71] For $n = 65{,}493$ items, the number of distinct pairs therefore is $65{,}493 \times 65{,}492/2 = 2{,}144{,}633{,}778$. But that is not all. For each pair, there is only one way to match all thirteen loci, but there are many more ways to get a nine-locus partial match. The profiles in the pair might match at the first nine loci and not match at the next four; they might not match at the first four but then match at the next nine; and so it goes for the $(13!)/(9!)(4!) = 715$ distinct combinations of nine items out of thirteen. With no particular set of nine loci that need to match, we perform $715 \times 2{,}144{,}633{,}778$ comparisons, which gives us more than 1.53×10^{12} opportunities to find some nine-locus matches. If the chance of any nine-locus match for any pair were "one in 754 million,"[72] then the expected number of nine-locus matches would be not just one. It would not even be 90, as Troyer mentioned, or 122, as the court-ordered report stated.[73] It would be $(1.53 \times 10^{12}) / (7.54 \times 10^8) = 2{,}034$ nine-locus matches. It seems as if random-match probabilities are even smaller than the theoretical estimates.

[68] *See id.*

[69] The more cynical view, attributed to Erin Murphy, a "graduate" of the District of Columbia Public Defender's Office, and now Acting Professor of Law at the University of California at Berkeley, is that the forensic science community is nefariously engaging in "goal-post shifting on a grand scale." Smith, *supra* note 18. "'The story before Arizona was that a nine-locus match was, like, a one-in-a-trillion thing," [Murphy] says. 'Now the story is, we expected that all along.'" *Id.* The real story is that DNA analysts, who are not necessarily skilled in probability and statistics, failed to distinguish between ordinary casework and all-pairs searching within a database. The failure of forensic analysts to apply appropriate statistical procedures is all too common. *See, e.g.*, NATIONAL RESEARCH COUNCIL OF THE NATIONAL ACADEMIES, COMMITTEE ON SCIENTIFIC ASSESSMENT OF BULLET LEAD ELEMENTAL COMPOSITION COMPARISON, FORENSIC ANALYSIS: WEIGHING BULLET LEAD EVIDENCE 31–35 (National Academies Press 2004) (criticizing the FBI's "chaining method" for matching bullet fragments); Kaye, *"False, But Highly Persuasive"*, *supra* note 12 (discussing errors in the computation and presentation of DNA random-match probabilities).

[70] Devlin, *supra* note 12.

[71] The first member of a pair is any one of the n items. The second member is any of the $n–1$ remaining items. That gives $n(n–1)$ pairs. But these include pairs such as (1, 50) and (50, 1), so we divide by two to get the total number of distinct pairs.

[72] Smith, *supra* note 18.

[73] *See id.*

But that conclusion cannot rest on these particular numbers. The theoretical random-match probability is not 1 in 754 million. That estimate pertained to Caucasians with the genotype seen in the first nine-locus match back in 2001. Each profile has its own random-match probability in each population group. A detailed study would need more than the summary statistics to ascertain whether the observed numbers of partial matches exceed those predicted by the basic product rule for unrelated individuals. To account for the vast number of possible pairings, we could refer to the individual profiles to compute the different RMP for each pair of profiles and derive an expected number of partial matches. Or, we could use a more efficient procedure that merely requires the number of loci at which all the pairs match and the number of loci at which they partially match.[74] In principle, more detailed studies with individual-level data on profiles and relatedness (or suitably detailed counts) might suggest that the theoretical RMPs are just fine, are too high (benefitting defendants), or are too low.[75]

III. THE FBI'S PURISM AND STATISTICAL RESEARCH WITH OFFENDER DATABASES

The FBI opposes statistical studies of offender databases that might undercut the theoretical RMPs that were the subject of a decade of litigation and scores of academic papers. Its arguments range from the purist to the practical. At a Promega conference five years after the Arizona report was posted, the leader of the FBI's scientific work on DNA evi-

[74] *See* B.S. Weir, *Matching and Partially-matching DNA Profiles*, 49 J. FORENSIC SCI. 1009 (2004) [hereinafter Weir, *Matching and Partially-matching DNA Profiles*]. The phrase "partial match" can be confusing. A match at every allele at 9 out of 13 loci, for example, is a partial match in the sense there is a full match at nine loci and something less than a full match at the other four loci. But how much less? Without additional specification, a nine-locus partial match is consistent with various possible numbers of "partial matches" at the four incompletely matching loci. Let "9/4" mean that nine loci are fully matching and that the pair of profiles contain exactly one matching allele at each of the other four loci. The total partial match thus consists of nine fully matching loci and four partly matching ones. Next down on the scale of nine-locus partial matches is a "9/3" match, that is, a full match at nine loci, partial matches (for exactly one of the two alleles) at another three loci, and a complete mismatch at the remaining locus. A "9/0" represents the weakest nine-locus partial match, being a full match at nine loci and a full mismatch at the other four loci. This notation is borrowed from James M. Curran et al., *Empirical Testing of Estimated DNA Frequencies*, 1 FORENSIC SCI. INT'L: GENETICS 267, 268 (2007). The most informative comparison of expected to observed numbers of partial matches is possible only when all degrees x/y of partial matching have been counted. *See* Weir, *supra*.

[75] The family relationships of the individuals whose profiles are in a database are potentially important. Assuming that everyone is unrelated can result in underestimates of the expected numbers of matches. More partial matches would be likely because close relatives will match more often than will the unrelated individuals to whom the theoretical RMPs apply. Therefore, further analysis would be required to interpret an excess of partial matches as a defect in the theory for computing RMPs. *See infra* Part III.

dence, Bruce Budowle, and various coauthors issued a "Clarification of Statistical Issues Related to the Operation of CODIS."[76] These investigators categorically claimed that "no valid analyses using such a repository can be carried out regarding the reliability of current statistical practices because there are duplicate profiles and profiles from relatives (of varied and unknown kinship category) contained within the database, and the population data are heterogeneous."[77] In contrast, the much smaller reference databases used to estimate allele frequencies in the major population groups are (or should be) carefully constructed to ensure that no individual provides more than one sample, that the census-type racial category of each individual is known, and that no close relatives are included.

Offender databases are not designed with these constraints in mind. For example, it does not matter if the database contains two records of the same individual's profile because of clerical error or the individual's use of an alias at one time. If there is a hit to this profile, then the two names will emerge, and the criminal investigation will reveal that one name is a mistake or an alias. And, to enhance the chance that anyone who might leave DNA at a crime-scene is located, it is good to have relatives in the database. From the investigators' perspective, the more, the merrier. From the population geneticists' standpoint, however, the database is a mess. The people represented in it are not from any randomly mating population or any real admixed population. Instead, the sample reflects a strange population structure, and the frequencies of alleles and profiles reflect unknown types and degrees of relatedness. Thus, the Budowle group argues:

> [A]ny results obtained from studies assessing the number of observed and expected genotypes at a number of loci from the offender database would be virtually irrelevant and would be misleading for either supporting or refuting current forensic DNA statistical practices. . . . [B]efore any legitimate inferences could be drawn from such databases, it would be imperative to remove as many as possible of the duplicates or profiles contributed by close family members.[78]

With the demand for clean and tidy research databases in place, the group followed up with a very practical argument: "Identifying and

[76] "CODIS" stands for the "Combined DNA Index System" of loci and software that the FBI developed to facilitate multi-state database trawls. *See* CLARIFICATION, *supra* note 64.

[77] *Id.* at 3.

[78] *Id.* However, a modest number of duplicates is not a substantial problem in comparing the expected and observed numbers of partially matching profiles. Curran et al., *supra* note 74, at 268.

resolving (and then removing) these profiles would be a monumental task and would not be readily possible to accomplish."[79]

The choice, however, is not between ideal research databases and worthless, motley assemblies of samples. The allele-frequency databases are themselves convenience samples with less-than-perfect group classifications, but they still have a place in producing reasonable estimates of allele frequencies and in verifying the assumptions of statistical independence.[80] Likewise, the offender databases could play a role in checking on basic product-rule (and more sophisticated) calculations of random-match probabilities.[81]

It is therefore worth describing the studies that already have been conducted. As a whole, they do not cast much doubt on the entrenched methods for estimating match probabilities. However, the existing research is limited in scope, and further research could reinforce or undermine this conclusion. In addition, databases now are large enough that they could be used to supplement if not replace the theoretical random-match probabilities with direct empirical estimates that do not rely on any population-genetics models. The remainder of this part, therefore, reviews the existing research and describes what it indicates about RMPs.[82] It begins with a few studies of data from Arizona and New Zealand that are hampered by the lack of individual-level data on profiles, then proceeds to more probative studies that begin with such data. As a whole, the research supports the small theoretical RMPs for multi-locus profiles in unrelated individuals.

Part IV then considers whether existing laws and bioethical principles stand in the way of further and open research. It concludes that they do not. Further empirical tests with offender databases may be less criti-

[79] CLARIFICATION, *supra* note 64, at 3.

[80] Using partial-matches to validate the independence assumptions predates the rise of offender databases. Neil Risch & Bernard Devlin, *On the Probability of Matching DNA Fingerprints*, 255 SCIENCE 717 (1992), for example, used the basic product rule on the FBI's and Lifecodes' reference databases of five VNTR loci for Blacks, Caucasians, and Hispanics. They estimated how many pairs of people within each group would match at two loci, three loci, four loci, and five loci. The research database sizes were only in the thousands, but this enabled them to make millions of comparisons. The observed numbers of partially matching pairs meshed with the expected numbers of matching pairs.

[81] *See* Brenner, *supra* note 46; *see also* Bruce Weir, *The Rarity of DNA Profiles*, 1 ANNALS APPLIED STAT. 358 (2007) [hereinafter Weir, *The Rarity of DNA Profiles*].

[82] Although one geneticist recently wrote that "testing the predictions of rarity provided by simple models like the product rule [by studying] the frequency of partial matches" is "new," Laurence D. Mueller, *Can Simple Population Genetic Models Reconcile Partial Match Frequencies Observed in Large Forensic Databases?*, 87 J. GENETICS 101, 102 (2008), much earlier studies on partial matches in smaller, research databases supported the use of the simple product rule, albeit at far fewer than nine loci. *See, e.g.*, Risch & Devlin, *supra* note 80, at 719–20.

cal than the public exposés suggest, but this line of research is legally and ethically sound.

A. *Studies Without Individual-level Data*

1. The Arizona Numbers

Several studies of the Arizona database have been conducted. Using the basic product rule with an estimated uniform single-locus match frequency of about 1 in 14, Charles Brenner reported that the excess number of nine-locus partial matches could be explained by a very small degree of dependence.[83] A more detailed, but also unpublished, analysis by Steven Myers reached the same conclusion.[84] Myers employed the affinal model for a structured population and considered some values for the proportion of full siblings in the Arizona database.[85] Bruce Weir also found no excess in the observed number of partial matches in Arizona.[86] A more complex study by Laurence Mueller used simulations with varying numbers of subpopulations, degrees of population structure, and proportions of siblings and parent-child pairs in the Arizona database.[87] By adjusting these parameters, he was able to fit the numbers of observed nine-locus partial matches or ten-locus partial matches with his model—but not both at the same time.[88] To put it another way, all his simulations were consistent with 150 nine-locus matches and 15 ten-locus ones.[89] The actual numbers of nine-locus and ten-locus matches in the Arizona database, it will be recalled, were 122 and 20, respectively.[90] Thus, there were 20% fewer nine-locus matches and 25% more ten-locus matches than a theoretical model with product-rule RMPs would predict. These discrepancies may be statistically significant,[91] but they present a very different picture than the original portrayal of inexplicably "huge" numbers of partial matches in Arizona. The studies of the few numbers

[83] *See* Brenner, *supra* note 46.

[84] *See* Steven P. Myers, Cal. Dep't Justice, Address to the California Association of Criminals and the Forensic Science Society: Felon-to-Felon Partial Profile Matches in the Arizona Database (May 9, 2006).

[85] *See id.*

[86] *See* Weir, *The Rarity of DNA Profiles*, *supra* note 81, at 361–65 (using a large value of 0.03 for the co-ancestry coefficient *è*).

[87] *See* Mueller, *supra* note 82.

[88] For example, when he tuned the model to give the observed number of ten-locus matches, he found that the comparisons in the Arizona produced too few nine-locus matches. *See id.* at 104–05. In other words, for these parameter values, rather than there being an excess of nine-locus matches as Barlow and others believed, there were too few. *See id.*

[89] *See id.* at 106.

[90] *See* Smith, *supra* note 18.

[91] *See* Mueller, *supra* note 82 at 105. Statistical significance means that the differences are improbable if all the modeling assumptions are correct. It is not a good indicator of whether the differences are of any practical importance. *See* KAYE ET AL, THE NEW WIGMORE, *supra* note 12, § 11.8.3(b), at 417.

reported in Arizona do not demonstrate that the theoretical computations yield absurdly small estimates of the true probabilities of a match among unrelated individuals.

Perhaps the simplest way to see this is to eschew the theoretical RMPs and estimate the average probability of a partial match from the observed numbers. As explained in Part II, there were more than 1.53×10^{12} opportunities to find some nine-locus matches in the database. The observed number of 122 such matches therefore yields an estimated RMP of $122/(153 \times 10^{10})$, or about one in ten billion.[92] The 20 ten-locus matches produce an empirical estimate of about 3 in 100 billion.[93] These empirical estimates are a kind of average RMP for the particular collection of related and unrelated individuals in the Arizona database. Their appeal is that they do not use the product rule or any other modeling assumptions.[94] Like the product-rule RMPs, these estimates are quite small.

2. The New Zealand Database Exercise

An early report of an all-pairs trawl in a large database comes from New Zealand.[95] At the time, 10,907 samples had been typed at six loci.[96] This means that there were about 59 million distinct pairs in the national database.[97] Since the theoretical random-match probability was about 1 in 50 million, if all the individuals represented in the database were unrelated, one would expect that an exhaustive comparison of the profiles for these 59 million pairs would produce only about one match. In fact, the 59 million comparisons revealed ten matches.[98] The excess number of matches is evidence that either not all the individuals in the database were unrelated, that the true match probability was smaller than the theoretical calculation, or both. In fact, eight of the pairs were twins or brothers.[99] The ninth was a duplicate (because one person gave a

[92] In a related calculation, two FBI scientists and the population geneticist Ranajit Chakraborty concluded that "[t]he RMP . . . values for 122 matching 9 locus pairs are between 1.523×10^{-12} and 6.769×10^{-14}." Budowle et al., *supra* note 56, at 62.

[93] There are 13! / (10!)(3!) = 286 ways to obtain a ten-locus match and 2,144,633,778 pairs of profiles. The empirically estimated RMP is therefore 20 (the observed number) divided by the $286 \times 2{,}144{,}633{,}778 = 6.13 \times 10^{11}$ possible ways to obtain the 20 matches.

[94] The estimates are not the probability that any pair of individuals in the Arizona database would have nine or ten loci in common. This probability does not account for the fact that the trawl is for any nine-locus or ten-locus match. This adjustment is necessary in estimating an overall RMP because in case work, a search would be conditioned on the specific nine or ten loci in a degraded sample.

[95] *See* Simon Walsh & John Buckleton, *DNA Intelligence Databases*, *in* FORENSIC DNA EVIDENCE INTERPRETATION 439, 463 (John Buckleton et al. eds., CRC Press 2005).

[96] *Id.*

[97] *Id.*

[98] *Id.*

[99] *Id.*

sample as himself and then again pretending to be someone else).[100] The tenth was apparently a match between two unrelated people.[101] This exercise thus confirmed the theoretical computation of the random-match probability. On average, the theoretical match probability for unrelated people was about 1/50,000,000, and the rate of matches in the unrelated pairs within the database was 1/59,000,000.[102]

B. *Studies with Individual-level Data*

None of the analyses of the matches in Arizona and New Zealand incorporates the match probabilities for each pair of profiles or takes advantage of data to the extent of single-allele matches at different numbers of loci.[103] Two published studies of databases in Australia and New Zealand fill this gap and therefore constitute the most convincing analyses to date.[104]

1. The First Australian Database Study

In 2004, Bruce Weir reported a study of the Australian national offender database, which then contained 14,768 nine-locus profiles from disparate ethnic groups.[105] This permitted 109,039,528 pairs to be compared without regard to ethnicity.[106] There were hardly any partial matches at more than five loci.[107] The basic product rule provided "good overall fit" to the observed numbers up to that point, "with some sets of loci having more matches than expected and some having less."[108] With the affinal model,[109] the number of partial matches was always less than

[100] *See* Walsh & Buckleton, *supra* note 95, at 463.

[101] *See id.* at 463–64.

[102] *Id.* at 454.

[103] *See* Weir, *Matching and Partially-matching DNA Profiles*, *supra* note 74 and accompanying text; *see also supra* note 78 and accompanying text.

[104] At a recent meeting, researchers from the Netherlands Forensic Institute described an all-pairs study of the Dutch national database. The database of suspects and offenders had 72,317 complete ten-locus profiles. Comparing the more than 2.5 billion resulting pairs, the group found a large number of ten-locus matches (773), which they dismissed as probable duplicates. As for partial matches, they found no nine-locus matches, and even without accounting for relatives, they reported substantial agreement between the observed and expected numbers of both partial and complete matches at eight and fewer loci. *See* Marjan Sjerps et al., Neth. Forensic Inst., Address at Fifth European Academy of Forensic Science Conference: Observed and Expected Numbers of (Partially) Randomly Matching Profiles in the Dutch DNA Database and in International DNA Searches (Sept. 11, 2009).

[105] *See* Weir, *The Rarity of DNA Profiles*, *supra* note 81, at 1010.

[106] *See id.* at 1011

[107] *See id.*

[108] *Id.* at 1010.

[109] Using *è* = 0.01.

the predicted value.[110] This study supports the use of the theoretical RMPs, at least with respect to matches at up to nine loci.

2. A Later Australian and New Zealand Database Study

James Curran and his colleagues obtained similar results when they examined the numbers of partial matches in national databases in Australia and New Zealand. There were no matches at nine loci in the recorded profiles. Dividing the profiles into "broad racial groups,"[111] Curran, Walsh, and Buckleton fit a model for both substructure and relatedness within these groups to conclude that "the fit of observed [numbers of partial matches] and [expected numbers] is remarkable and gives substantial support to the reliability of current methods."[112]

In sum, the research to date gives little reason to doubt the adequacy of the existing model for computing multilocus STR frequencies. Nevertheless, still more powerful tests could be conducted with the U.S. national database. NDIS contains approximately seven million offender profiles,[113] allowing some 25 quadrillion pairwise comparisons.[114] Assuming that supercomputers are up to the task, a small number of matches in this many comparisons would directly confirm (or provide) estimates of random-match probabilities for 13-locus matches. But many commentators have questioned the propriety of using DNA data on non-

[110] Weir concluded that the affinal model with $\hat{e}$ = 0.03 should be conservative for as many as nine loci. *See* Weir, *The Rarity of DNA Profiles*, *supra* note 81, at 1011.

[111] The profiles came from 17,501 Australian Caucasians, 8,630 Australian Aborigines, 28,696 Western Australian Caucasians, 10,763 Western Australian Aboriginals, 9,840 New Zealand Caucasians, 11,042 New Zealand Eastern Polynesians, and 2,797 New Zealand Western Polynesians. Curran et al., *supra* note 74, at 268.

[112] *Id.* at 267. Possible relatedness among the individuals represented in the offender databases seemed to be more important than the corrections in the affinal model. In fact, once relatedness was estimated, the standard corrections for population substructure within the broad groups appeared to be conservative, yielding theoretical probabilities that predicted more partial matches than were observed at the high end. James M. Curran et al., *Empirical Support for the Reliability of DNA Evidence Interpretation in Australia and New Zealand*, 40 AUSTRALIAN J. FORENSIC SCI. 99, 102-06 (2008); *cf.* Gordan Lauc et al., *Empirical Support for the Reliability of DNA Interpretation in Croatia*, 3 FORENSIC SCI. INT'L: GENETICS 50, 51, 53 (2008) (concluding from the 259 profiles of single relatives of missing people in Croatia (yielding 33,411 pairs) that "[i]t may be timely to start discussion about further practical ways, beyond the substantial efforts already made by some laboratories, in which the importance of relatedness may be appropriately and proportionately accommodated in casework and reported to court").

[113] Federal Bureau of Investigation, NDIS-CODIS Statistics, http://www.fbi.gov/hq/lab/codis/clickmap.htm (last visited Mar. 3, 2009).

[114] The 1996 NAS committee recommended against such straightforward estimates based on counting within databases. *See* NRC, *supra* note 14. However, that was years ago, when the databases were too small to give useful estimates of very rare events.

consenting individuals,[115] and the FBI has cited the privacy of individual records as a reason to keep the offender databases out of the hands of defendants.[116] The final part contends that these concerns are not significant barriers to conducting statistical research with NDIS.

IV. ETHICAL AND LEGAL CONSTRAINTS ON USING AND DISCLOSING DATABASE RECORDS

Having seen that statistical research with the offender databases can increase confidence (or shake it) in the theoretical RMPs, the question becomes whether legal or ethical strictures preclude this effort. This part shows that current law authorizes this research when conducted by the government itself. It also argues that disclosure of an anonymous list of STR profiles to scientists outside of law enforcement is legally and ethically permissible.

A. *Legality of Internal and External Research on Offender Databases*

Police agencies are empowered to prevent crimes and apprehend criminals,[117] and statutes are not needed to prescribe every procedure that they may use to obtain, preserve, and utilize information that could be helpful in discharging their duties.[118] Nevertheless, DNA databases,

[115] *See* David H. Kaye, *Behavioral Genetics Research and Criminal DNA Databanks*, *in* THE IMPACT OF BEHAVIORAL GENETICS AND NEUROLOGY IN CRIMINAL LAW 355, 357 (Nita Faranhy ed., Oxford University Press 2009) [hereinafter Kaye, *Behavioral Genetics Research*].

[116] *See* Felch & Dolan, *supra* note 17, at A1; *see also* Martha Neil, *Are DNA Tests as Accurate as We Thought?*, ABAJ LAW NEWS NOW, July 23, 2008, http://abajournal.com/news/are_dna_tests_as_accurate_as_we_thought1.

[117] *See, e.g.*, ARIZ. REV. STAT. § 11-441(A) (LexisNexis 2008) ("The sheriff shall: (1) Preserve the peace. (2) Arrest and take before the nearest magistrate for examination all persons who attempt to commit or who have committed a public offense.").

[118] In *State ex rel. Mavity v. Tyndall*, 66 N.E.2d 755 (Ind. 1946), for example, a man arrested in misdemeanor charges involving gaming was photographed, fingerprinted, and forced to give a signature. Copies of the fingerprints were sent to the Indiana State Police and the FBI. *See id.* at 756. Soon afterward, the charges were dismissed. *See id.* at 757. He demanded "the return of the data on file with the Indianapolis Police Department and that its officers obtain the return of the copies sent to the Federal Bureau of Investigation and the Indiana State Police." *Id.* The Indiana Supreme Court refused to order the release and return of the information. *See id.* at 762–63. It explained that "[i]n the absence of statutory direction or regulation the power to maintain and operate a city police system carries with it the right and duty to exercise reasonable discretion in such maintenance and operation. Courts should be cautious about interference with such an executive discretion." *Id.* at 757. Cases like this suggest that police may collect information on suspects and access that information (consistently with the Fourth Amendment, of course) it even if no statute specifically provides for these activities. *But cf.* State *ex rel.* Bruns v. Clausmier, 57 N.E. 541, 542 (Ind. 1900) (holding that no civil cause of action lies against a sheriff for photographing a prisoner later acquitted of the charge of forgery and placing that photograph in the "Rogues' Gallery" because "[a] sheriff, in making an arrest for a felony on a warrant, has the right to exercise a discretion, not only as to the means taken to apprehend the person named in the warrant, but also as to the means

at the state and federal levels, are creatures of statutes adopted in the 1980s or 1990s. These statutes generally address both the use of the offender databases for population-genetics research and the disclosure of the records in these databases. The DNA Identification Act of 1994,[119] which applies to the federal government and all states,[120] empowers the FBI to "establish an index of . . . DNA identification records."[121] The statute indirectly indicates the uses that can be made of the records by limiting the database to records received from "federal, state, and local criminal justice agencies [that] allow disclosure of stored DNA samples and DNA analyses only—

> (A) to criminal justice agencies for law enforcement identification purposes;
> (B) in judicial proceedings, if otherwise admissible pursuant to applicable statutes or rules;
> (C) for criminal defense purposes, to a defendant, who shall have access to samples and analyses performed in connection with the case in which such defendant is charged; or
> (D) if personally identifiable information is removed, for a population statistics database, for identification research and protocol development purposes, or for quality control purposes."[122]

Research on the frequencies of partial matches within the database falls into subsection D's category of "identification research and protocol development."[123] As such, the FBI or state agencies could perform these studies in-house, and they could engage the services of outside researchers in the investigation of "population statistics."[124]

Whether the database administrators can disclose the DNA records to the general scientific community is less clear. The first three subsections seem to be aimed at disclosure of the DNA "samples and analyses"

necessary to keep him safe and secure after such apprehension until lawfully discharged; and he has the right to take such steps and adopt such measures as, in his discretion, may appear to be necessary to the identification and recapture of persons in his custody if they escape").

[119] 42 U.S.C. §§ 14131-14136e (2006).

[120] A state cannot participate in CODIS—as every state does—unless it operates "pursuant to rules that allow disclosure of stored DNA samples and DNA analyses only" for the purposes enumerated in 42 U.S.C. § 14132(b). *See id.* § 14132(c). Under § 14135e(c), "[a] person who knowingly—(1) discloses a sample or result described in subsection (a) of this section in any manner to any person not authorized to receive it; or (2) obtains, without authorization, a sample or result described in subsection (a) of this section, shall be fined not more than $100,000."

[121] *Id.* § 14132(a).

[122] *Id.* § 14132(b)(3); *see also id.* § 14133 (providing the same "privacy protection standards" apply to the federal records produced by the FBI itself).

[123] *See id.* § 14132(b)(3).

[124] *See id.*

involving a small number of personally identified individuals who might have been thought to be the source of DNA recovered at a crime scene.[125] Subsection C, for instance, limits "access to samples and analyses performed in connection with the case in which such defendant is charged."[126] Plainly, this subsection does not entitle the defendant to the millions of DNA samples and profiles that come from *other* investigations or convictions.

Subsection D is a more plausible source of authority for releasing entire databases for statistical studies. It allows for an anonymized version of a state database or of the NDIS to be used as "a population statistics database" or as a tool for "identification research."[127] The privacy interest in anonymous DNA profiles is minimal,[128] and research confined to operating and improving the system of DNA identification advances rather than harms legitimate personal interests.[129] Therefore, subsection D contains none of the restrictions of the previous three subsections that deal with personally identified data.[130] Thus, subsection D should allow database administrators to create and release the entire set of records, stripped of all identifiers, for the research desired by some defense counsel.[131]

B. *Ethical Propriety of Research on Offender Databases*

Researchers are not free to experiment on human beings, especially vulnerable ones such as prisoners. The horrors exposed in the Nazi Doctors' Trial before the Nuremberg Military Tribunals should stamp this fact indelibly on the memory of humanity.[132] The core human right to be free from unwanted bodily invasions, however, has been extended and diluted in efforts to protect informational privacy. The possibility of sig-

[125] *See* 42 U.S.C. § 14132(b)(3) (2006).

[126] *See id.* As used throughout the Act, the word "analysis" means the testing of DNA molecules for their identifying features—that is to say, the production of the DNA profiles that go into the database. *See, e.g.*, *id.* § 14135(d) (describing the requirements for grants to states for the "[a]nalysis of samples"); *id.* § 14135a(2) (in relation to the "[c]ollection and use of DNA identification information from certain Federal offenders, . . . 'DNA analysis' means analysis of the deoxyribonucleic acid (DNA) identification information in a bodily sample"). Hence, "analysis" does not mean the comparison of profiles from DNA that already has been analyzed.

[127] *See id.* § 14132(b)(3).

[128] *See infra* Part IV.B.1.

[129] *See* Kaye, *Behavioral Genetics Research*, *supra* note 115; D.H. Kaye, *Bioethics, Bar, and Bench: Selected Arguments in* Landry v. Attorney General, 40 JURIMETRICS J. 193, 216–17 (2000) [hereinafter Kaye, *Bioethics*].

[130] *See* 42 U.S.C. § 14132(b) (2006).

[131] *See id.*

[132] *See, e.g.*, George J. Annas, *The Legacy of the Nuremberg Doctors' Trial to American Bioethics and Human Rights*, 10 MINN. J. L. SCI. & TECH. 19 (2009); Edmund D. Pellegrino, *The Nazi Doctors and Nuremberg: Some Moral Lessons Revisited*, 127 ANNALS INTERNAL MED. 307 (1997).

nificant psycho-social harm from the release of information about an individual and the notion that no scientific research on human subjects should be undertaken without their consent have been invoked to challenge the use of information derived from DNA samples in genetic research. These concerns have grave implications for some types of research, but they carry little weight as applied to research into DNA-identification profile frequencies.

1. Privacy—of What?

The interest of individuals in keeping their STR profiles secret is comparable to their interest in not revealing their blood groups or fingerprints. To some, this may seem like a major concern; others may think it involves only trivial information. Whatever the gravity of this interest may be, using anonymized lists of STR profiles does not implicate it. With personal identifiers removed, these records pose no realistic threat to individuals. This is so notwithstanding the suggestion that because DNA is uniquely identifying, there can be no such thing as anonymous DNA records or samples.[133] Consider the most extreme case of public disclosure, in which the list of DNA profiles were to be published on the Internet.[134] The only invasion of privacy could come from (1) taking a sample of DNA from a known individual (for example, by going through a neighbor's trash for used dental floss), (2) having a private laboratory analyze the sample, and (3) searching the profile against the posted profiles. This procedure would reveal that the neighbor is (or is not) a convicted offender or another individual subject to inclusion in a law enforcement database. Yet, even in this strained scenario, the genetic information on the individual (the STR profile) was not obtained from the publicly available information. The posted data simply indicated that the individual's profile is probably in the offender database as well. Inasmuch as the vast majority of the profiles in the database come from convictions that are matters of public record anyway, any reputational or other harm to the individual from the nosy neighbor's discovery will be quite rare.

Several researchers have warned of a different mechanism that might link individuals to the STR profiles.[135] They write:

[133] *See* Bernice S. Elger & Arthur L. Caplan, *Consent and Anonymization in Research Involving Biobanks*, 7 EMBO REPORTS 661, 664 (2006) ("If samples contain any trace of DNA, they are not truly anonymous, because it is always possible to identify the donor through DNA fingerprinting."). For a rejoinder, see J. Jaap Nietfeld, *What Is Anonymous?*, 8 EMBO REPORTS 518 (2007).

[134] *Cf.* Catherine Clabby, *DNA Research Commons Scaled Back*, 97 AM. SCIENTIST 113 (2009) (reporting on the National Institutes of Health closure of open-access, pooled genomics databases).

[135] *See* Budowle et al., *supra* note 56, at 63.

> Extremely important are that concerns exist and will arise about the privacy and confidentiality of data retrieved from matches found during a pair-wise comparison of offender DNA profiles. The names of individuals with matching and partial matching profiles would have to be disclosed to scientists and police when there is no criminal investigation underway. The names would be obtained because of a "research experiment." To further annotate such data may not be possible.[136]

Apparently, the authors assume that the names will have to be revealed to ascertain whether the matches are due to duplications or twins, or whether the partial matches are due to relatives.[137] Such an effort would be required only if one insists on a pristine database for research into the RMPs. In addition, it seems extravagant to brand requests to police to investigate whether two records in NDIS are duplicates as "[e]xtremely important" breaches of "the privacy and confidentiality of data."[138] Production of an imperfect, but still useful, anonymized database of STR profiles is therefore feasible and ethical.

2. Consent—for What Reason?

Invocations of a right to informed consent permeate the literature on the research uses of the databases and databanks.[139] But it is not enough to assert, in the broadest possible terms, that "[t]he right to consent or refuse to take part in research is an important right for individuals and for society."[140] The reasons for insisting on informed consent in medical or other scientific research on human subjects do not apply when samples are legally compelled and the information extracted from them is used solely to ensure that the very system that justifies this compulsion is

[136] *Id.*

[137] *See id.*

[138] *Id.* The FBI or its contractors identified duplicates when several surprisingly similar fingerprints emerged in a research study with a subset of the national fingerprint database. *See* David H. Kaye, *Questioning a Courtroom Proof of the Uniqueness of Fingerprints*, 71 INT'L STAT. REV. 521 (2003).

[139] *E.g.*, Michelle Hibbert, *DNA Databanks: Law Enforcement's Greatest Surveillance Tool?*, 34 WAKE FOREST L. REV. 767, 821–22 (1999) ("If states decide they would like to release samples for research purposes, however, then these states should only release the samples in accordance with federal regulations governing the use of stored tissue samples, which in most cases would likely require the informed consent of the offenders before releasing the samples.").

[140] KRISTINA STALEY, GENEWATCH UK, THE POLICE NATIONAL DNA DATABASE: BALANCING CRIME DETECTION, HUMAN RIGHTS AND PRIVACY 8 (2005), http://www.genewatch.org/uploads/f03c6d66a9b354535738483c1c3d49e4/NationalDNADatabase.pdf; *see also id.* at 46 ("Consent should have to be obtained from the individuals on the database before genetic research is allowed to go ahead.").

working as it should.[141] Informed consent serves to waive individual rights that stand between the subject and the researcher. These include the rights to be free from intentional bodily harm, from offensive touching or intrusion, from unnecessary confinement and physical restraint, and from serious and reasonable emotional distress. Thus, physicians are not at liberty to perform experimental (or even clinically accepted) surgery on their patients even when the surgery is the patient's only hope.[142] In the case of statistical analysis of offender databases, however, there is no threat to bodily integrity and no psychological harm to the individual. In the absence of a legal or moral right to keep the information secret, there is no ethical breach in using the offender DNA data to study the frequency of matching DNA types in the population.

CONCLUSION

The government's current policy of obstructing research into the number of partial matches in large databases is ill-advised. Studies to date tend to support the accepted method for computing random-match probabilities, but these studies are limited by sample size and the lack of details on individual profiles and familial relationships of the individuals whose DNA profiles are in the databases. Studying the profiles in the large national database, NDIS, could overcome or ameliorate the first two limitations. The third limitation biases the outcomes toward a larger-than-expected proportion of matches within offender databases (overstating random-match probabilities and underpredicting the number of chance matches), especially since relatives could comprise a third or more of the database.[143] Nonetheless, if the partial-match rate in all-pairs trawls of offender databases continues to be very small, it will constitute additional evidence that the probability of a match between two unrelated individuals at a substantial number of STR loci is very small. The proposed research into this issue is compatible with existing law and bioethical precepts. Indeed, under conventional scientific norms that encourage the availability of research data, the government should make an anonymized version of NDIS available to all researchers.[144] A policy of

[141] *See* Kaye, *Bioethics*, *supra* note 129.

[142] *See*, *e.g.*, Schloendorff v. Soc'y of N.Y. Hosp., 105 N.E. 92, 93 (N.Y. 1914) ("Every human being of adult years and sound mind has a right to determine what shall be done with his own body; and a surgeon who performs an operation without his patient's consent commits an assault, for which he is liable in damages.").

[143] It is likely that close relatives comprise a substantial fraction of the offender databases. *See* ALLEN BECK ET AL., BUREAU OF JUSTICE STATISTICS, SURVEY OF STATE PRISON INMATES 9 (1991) (37% of inmates reported having a parent or sibling "who had served time"); DORIS L. JAMES, BUREAU OF JUSTICE STATISTICS, PROFILE OF JAIL INMATES 9 (2002) (national survey of jail inmates found that 46% of the inmates had a sibling or parent who had been incarcerated).

[144] Strictly speaking, researchers could conduct all-pairs studies of the robustness of the theoretical RMPs without access to the profiles themselves. *See* Weir, *Matching and Par-*

openness will permit a robust inquiry that will ultimately produce either greater confidence in the method now used to estimate RMPs, or some more defensible form of these estimates. The stakes in many criminal cases and the importance of DNA evidence to the criminal justice system are too high to continue the policy of ignoring or keeping relevant data secret and unexplored.

tially-matching DNA Profiles, *supra* note 74 and accompanying text. However, open access enables independent researchers to verify the accuracy of any summary statistics generated by the database managers and to develop additional testing procedures and lines of inquiry.