Data Sharing in Forensic Science: Consequences for the Legal System

Andrea L. Roth[1], Edward J. Ungvarsky[2]
[1]Law Fellow, Stanford Law School, 559 Nathan Abbott Way, Stanford, CA, 94305
[2]Northern Virginia Capital Defender, 2300 Clarendon Blvd., Arlington, VA, 22201

Abstract

The recent National Academy of Sciences (NAS) Report documenting the deficiencies in "forensic science" noted that many forensic science fields are woefully lacking in validation through independent research. While the Report touts DNA testing as a uniquely reliable means of forensic identification, it is precisely for this reason that continued validation of DNA match statistics through research, and the continued use of DNA in exonerating the wrongfully convicted, is so critical. Notwithstanding recent studies on three state offender DNA databases that call into question key assumptions underlying currently used DNA frequency tables, state and federal law enforcement officials have refused to give statisticians and population geneticists further access to offender databases. Moreover, these same officials routinely refuse to allow litigants access to offender databases to facilitate post-conviction claims of actual innocence, even when such litigants have proven that key DNA evidence excludes them as a suspect. This Article urges the statistical community to engage the forensic science community about the need for access to data, both in DNA cases and in other fields of forensic science. Heightened involvement of the broader scientific community, combined with much-needed funding for independent academic research by the proposed National Institute of Forensic Science, will hopefully encourage an ethic of data sharing in the forensic science community.

Key Words: DNA, offender database, match statistic, data access

1. Introduction

In February of this year, the nonpartisan National Academy of Sciences issued a groundbreaking and scathing report documenting numerous deficiencies in forensic science (NAS Report 2009). The report noted that several fields of forensic "science," such as toolmark analysis, handwriting comparison, and even the pedigreed latent fingerprint examination, have been exposed to little "stringent scientific scrutiny." *Id.* Because many of these methods were developed in crime laboratories to aid in law enforcement rather than by academic scientists and mathematicians, "researching their limitations and foundations was never a top priority." *Id.* As a result, many "non-DNA forensic tests do not meet the fundamental requirements of science." *Id.*

The Report is quick to note that the lack of scientific rigor in most forensic science fields is not the result of malfeasance; rather, the over-taxed forensic science community simply "has had little opportunity to pursue or become proficient in the research that is needed to support what it does." *Id.* The Report notes that "few sources of funding exist for independent forensic research" and that most studies that do exist are commissioned

by the Department of Justice itself and "conducted by crime laboratories with little or no participation by the traditional scientific community." *Id.*

The lack of validation in most forensic science fields through independent scientific and statistical research is perhaps most obvious in the non-DNA areas listed above. But DNA testing itself, precisely because it is the gold standard by which other forensic testing methods are judged, also is in dire need of an ethic of data sharing with the broader scientific community in two respects. First, because DNA match statistics – when used as a tool of inclusion – are touted as highly reliable near-certain proof of guilt, the continuing need to ensure that the match statistics are accurate is especially critical. As discussed below, recent studies of three state offender DNA databases have called the reliability of some DNA match statistics into serious question, and several well-regarded scholars are echoing the need for more research on other databases. Second, because DNA testing – when used as a tool of exclusion – can provide conclusive exoneration of the wrongly convicted, the need for reasonable access to offender databases to facilitate post-conviction innocence claims where a litigant has demonstrated that highly relevant DNA evidence excludes him as a suspect is also critical.

Notwithstanding the clear need for data sharing in both the DNA inclusion and exclusion contexts, state and federal law enforcement – including the Federal Bureau of Investigation – have thus far refused to allow independent scientists and statisticians access to DNA offender databases. While a small number of courts have forced certain laboratories to run searches related to claims of innocence, no court has required a state or federal laboratory to allow access to a database for purposes of testing the statistical assumptions underlying reported match statistics.

This Article suggests that the broader community of scientists and statisticians engage the forensic science community on the issue of DNA database access and, eventually, data sharing in other forensic disciplines. Through a combination of guaranteed funding, pressure from the scientific community, and court enforcement when necessary, DNA laboratories should begin to allow much-needed access to their databases. Hopefully, the forensic community will eventually welcome such collaboration, as the ultimate goal will be to retain DNA testing as a powerful and well-regarded tool of both inclusion and exclusion in criminal cases.

2. The Need for Access to DNA Offender Databases To Ensure Continued Viability of DNA Testing as a Tool of Inclusion and Exclusion

2.1 The need for data access to test the independence and relatedness assumptions underlying reported DNA random match probabilities (RMPs)

DNA is a powerful tool of inclusion in criminal cases, providing what the United States Supreme Court has called "near certain[]"proof of identity in some cases (*District Attorney's Office v. Osborne* (2009)). The unparalleled inculpatory power of DNA evidence rests in large part on the fact that forensic DNA analysis has advanced to the point where estimates of random match probabilities (RMPs) with denominators in the quintillions or higher are routinely reported (Lynch et al. 2008). Some have argued that such infinitesimally small RMPs are likely the product of inaccurate assumptions about independence of loci or the amount of substructure in the relevant population to which the RMP ostensibly applies (Weir 2001). A Stanford University mathematician, when asked to comment on the introduction of RMPs with such large denominators, opined that "such unsophisticated use of the product rule rapidly takes theory well beyond the bounds of reality" and that such RMPs are "ludicrous" (Devlin 2006). Nonetheless, nearly all

challenges to the statistical assumptions underlying RMPs generated in Short Tandem Repeat (STR) DNA typing have been settled by courts in favor of the government.

The debate over the accuracy of RMPs has gained new life, however, as a result of recent comparisons of profiles in existing DNA databases. In 2001, an analyst in Arizona's state crime laboratory compared each profile in the Arizona database to every other profile and determined through such "pairwise" comparisons that, among 65,000 profiles, 122 pairs matched at 9 of the FBI's 13 loci, and 20 pairs matched at 10 loci (Felch & Dolan 2008). Similar tests on the Illinois and Maryland databases, with 220,000 and 30,000 profiles, respectively, showed close to 1,000 other pairs of profiles matching at nine or more loci and three pairs matching at all 13 FBI loci. *Id.*

Evolutionary biologist Laurence Mueller has suggested, based on his analysis of the Arizona data, that the data are not easily reconciled with currently reported RMPs and that further research must be conducted to determine whether the high number of partial matches indicates a need to adjust fundamental assumptions underlying the FBI's allelic frequency tables and use of the product rule to generate its reported RMPs (Mueller 2008). Meanwhile, a group of Berkeley researchers has also recently called for more research into the assumptions of independence among the FBI's chosen 13 STR loci, after determining that, assuming independence across the 13 loci, the average chance of a coincidental (and thus erroneous) attribution in a pure cold hit case – where the entirety or near-entirety of the prosecution's case is a database profile match – is 1 in 3.4 million (Song et al. 2009). Other scholars have followed suit. *See, e.g.*, Thompson (2008) ("An important source of uncertainty is the relatively small size of available statistical databases, which makes it impossible to perform sensitive tests of the statistical independence of markers across multiple loci.").

The obvious candidate for further study is the Combined DNA Index System (CODIS) database, a collection of state and national databases maintained by the FBI that, as of March 2009, contained over 6.8 million offender profiles (CODIS-NDIS Statistics 2009). Calling a pairwise study of CODIS "a matter of some urgency," Professor Devlin expects that CODIS "will contain not just one but several pairs that match on all 13 loci, contrary . . . to the prediction made by proponents of the currently much-touted RMP that you can expect a single match only when you have on the order of 15 quadrillion profiles" (Devlin 2006). Other scholars, in contrast, have predicted that several such partial matches are to be expected in any pairwise comparison study of CODIS, and that such discoveries would not call into question currently reported RMPs, analogizing to the famous "birthday problem" (Kaye 2009; Budowle 2009).

Yet, as of this writing, the FBI refuses to conduct such pairwise comparisons in the CODIS database, or to allow independent researchers to do the same (*Id.*; Kaye (2009); Ungvarsky (2007); Murphy (2009)). Even more, the FBI "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" (Kaye 2009). While a handful of litigants have challenged the government's refusal to allow scientific access to the databases to run pairwise comparisons (*United States v. Blackmon* (2009)), no court has required it. Indeed, the government routinely seeks to preclude mention of the Arizona data in front of the jury at trial (*People v. Santana* (2009)).

The FBI has also refused to allow access to the raw data underlying the profiles in its mitochondrial DNA (mtDNA) population database, even though independent researchers have found several errors in the reported profiles (Kittles et al. 2006). In contrast, laboratories such as the Armed Forces DNA Identification Laboratory and the Institute of Legal Medicine have recently begun to collect thousands of mtDNA sequences from populations that are underrepresented in the FBI's database, and have implemented impressive automated quality control mechanisms (Just et al. 2004, Parson & Dür 2007).

Both laboratories have made clear their intentions to share all raw data with either Genbank, a sequencing database run by the National Institute of Health, or the European DNA Profiling Group's publicly available EMPOP mtDNA population database. *Id.*

2.2 Responding to law enforcement's proffered reasons for refusing data access to independent researchers

The first argument typically put forth by the FBI and state governments to justify their refusal to allow access to offender database profiles is that the databases contain private information about the offenders. California state officials, for example, have recently argued that making database profiles available to independent experts to run pairwise comparisons would violate the privacy of the offenders (S.F. Chronicle 2008). There is some irony, of course, in law enforcement's invocation of offender privacy in refusing to allow independent research that might reveal that match statistics admitted against offenders in criminal cases have been overstated by law enforcement. In any event, it is certainly true that, with access to a person's entire genetic strand, the government can determine her "[p]hysique and ethnic origins," arguably sexual orientation, as well as whether she suffers from alcoholism, schizophrenia, or a genetic predisposition to criminality (McCartney 2006). But a person's DNA profile alone would presumably offer at most information about gender and ethnicity, and an anonymous profile could not be connected back to any particular offender. While the inclusion of a person's profile in the database indicates that he or she has likely been arrested or convicted, such data are typically already matters of public record, and a person's inclusion in the database would be difficult to discern without already having a DNA sample from the person for purposes of comparing against the database (Kaye 2009).

The Forensic Science Service (FSS), the British governmental institution charged with maintaining the U.K.'s national offender DNA database, has allowed at least five private firms access to the database (Hope 2008). A spokesman for the National Policing Improvement Agency defended the government's disclosure of profiles to private firms by arguing that the profiles are "completely anonymous" and "not identifiable in any way. After approval, they were made available for authorized research purposes demonstrating clear, potential operational benefit to the police in terms of detecting and solving crime" Hope (2008). Other scholars have similarly suggested anonymization of records as the solution to any privacy concerns. *See, e.g.*, Kaye (2009). Even the DNA Identification Act of 1994, which authorized CODIS to begin with, contemplated that the database would be available for identification research and protocol development purposes, or for quality control purposes if "personally identifiable information is removed" (Kaye 2009; 42 U.S.C. § 14132(b)).

The second concern typically put forth is that such comparisons would be unduly burdensome to government officials, particularly on the state level, because such an internal search would take a week or longer and would not allow for CODIS searches during that time (Konzak 2005). But the internal search could presumably be completed on a separate computer or computers, and – as discussed below – the new independent agency proposed in the NAS Report could pay for any costs associated with such searches. Notably, no law enforcement official has publicly suggested that the searches conducted on the Arizona, Illinois, or Maryland databases were prohibitively expensive.

A final potential concern with respect to running pairwise comparisons is that there are too many duplicative profiles, and profiles belonging to related individuals, to make an analysis of the CODIS profiles worthwhile for testing current RMP estimates (Budowle 2009). But duplicate profiles could be ignored in any analysis, and laboratories may be able to easily remove duplicates with use of other identifying information kept

private, such as social security numbers and unique accession numbers (Gilder 2005; Gilder 2008). Moreover, while the presence of duplicative profiles and related offenders makes CODIS less than ideal as a means of estimating allelic frequencies, Professor Kaye points out that the FBI's STR population databases themselves consist merely of convenience samples from a limited number of ethnic groups, and that analysis of CODIS would "still play a role in checking on basic product-rule (and more sophisticated) calculations of random-match probabilities" (Kaye 2009). Moreover, the existence of so many related individuals in CODIS (Kaye 2009) itself is worthy of additional statistical research.

2.3 Access to CODIS to facilitate post-conviction claims of actual innocence

Investigation of crime scenes by either the police or the defense will sometimes yield a DNA sample that matches neither the complaining witness nor the charged defendant. In still other cases, a defendant who has already been convicted of a crime may seek testing of previously untested biological material from a crime scene, sometimes yielding a DNA sample that matches neither the complaining witness nor the defendant. In such cases, defendants have asked law enforcement to run the non-matching sample through a state or federal database to determine whether another person matches, on the theory that he or she might be the true perpetrator (*Illinois v. Griffin*, *Coleman v. Bradshaw*, *Rivera v. Mueller*). No law enforcement agency of which the authors are aware has voluntarily agreed to such a database search at the request of the defense, even where the prosecution has agreed to the search (*Rivera v. Mueller*). While few courts have definitively ruled on the issue, two federal courts in Ohio and Illinois recently ordered the FBI to perform such a search (*Coleman v. Bradshaw*; *Rivera v. Mueller*).

State and federal laboratories have typically offered two reasons for refusing to conduct such innocence-claim-related searches. First, some laboratories have also argued that they cannot run a third party profile through a database because the sample from which the profile was developed was tested at an unaccredited laboratory, and in order to run the profile through the database the agency must first "adopt" the profile and actually incorporate it in the database (*Coleman v. Bradshaw*; *Rivera v. Mueller*). Placing a profile from an unaccredited laboratory into the database would, to be sure, violate many agencies' quality control standards. But as the FBI acknowledged in *Rivera*, laboratories are capable of running a "manual keyboard search" – which is neither "costly [n]or time-consuming" – in which the profile to be compared is not actually uploaded into the database. Indeed, the FBI routinely compares "inferior-quality profile[s]" of fewer than 13 loci against the profiles in the database, even though such partial profiles would not meet the quality control standards for inclusion in CODIS (*Rivera*). Second, as the court in *Coleman* ultimately ruled, the law enforcement agency itself could use its partner laboratory to retest a sample and develop a new profile, this time tested by an accredited laboratory, appropriate for inclusion in the database.

Second, laboratories have argued that any match between the DNA and a third party offender in the database would not necessarily be "relevant" (*Coleman v. Bradshaw*) or "crucial" (*Rivera v. Mueller*) to the litigant's claim of innocence. It seems reasonable to require defendants to show why the presence of third-party DNA would tend to cast doubt on the identity of the defendant as the perpetrator, change the type or gravity of offense with which the defendant is charged, or corroborate a central claim critical to the defendant's credibility. While the mere presence of someone else's DNA in a crowded store might not be relevant in showing that the defendant is not guilty of robbing the store, the presence of another person's DNA in semen recovered in a rape kit (*Rivera*) would presumably be material. Given that the government has an affirmative duty to

disclose material serological evidence favorable to the defendant (*Brady v. Maryland*), a showing of materiality by the defendant should be sufficient to trigger the government's duty to run the profile through the database. As the *Rivera* court recognized, "the mere fact of a match to someone other than [the litigant] by itself could be the basis for a powerful argument on the part of his attorneys at trial."

.

3. The Solution: Engagement and Funding

The NAS Report lamented that the forensic science community "has only thin ties to an academic research base that could undergird the forensic science disciplines and fill knowledge gaps" (NAS Report 2009). The problem, according to the committee, is mostly one of "under resourcing" – a lack of funding and staff that would allow already over-worked forensic laboratories to engage in the type of research needed to make the disciplines more rigorous. While the FBI and National Institute of Justice have limited research budgets, the "level of support has been well short of what is necessary . . . to establish strong links with a broad base of research universities and the national research community." Moreover, most such funding requires "law enforcement collaboration," which hinders "the pursuit of more fundamental scientific questions" necessary to subject forensic science to proper scientific scrutiny.

To facilitate a new era of independent research to support forensic science disciplines, and to otherwise improve the state of forensic science, NAS specifically proposed that Congress create a new independent federal agency, the National Institute of Forensic Science (NIFS), that would, among other tasks, "fund[] . . . independent research projects" and "promot[e] scholarly, competitive, peer-reviewed research" in several areas, including "[s]tudies establishing the scientific bases demonstrating the validity of forensic methods" (NAS Report 2009).

If the greater community of scientists and statisticians took an active interest in collaborating with the forensic science community on database research issues, and if such research endeavors were well-funded by sources outside existing and already strained forensic laboratory budgets, the authors are confident that laboratories would be much less resistant to an ethic of data sharing and that court enforcement would become less necessary. Indeed, the Report quotes the director of the Los Angeles County Sheriff's Department as acknowledging that "[w]e run the risk of our science being questioned in the courts because there is so little research." In the end, further research will only strengthen the position of those seeking admission of DNA evidence, whether as a tool of inclusion or exclusion. As Professor Kaye points out, if further pairwise comparisons in CODIS would only confirm the accuracy of the FBI's reported match statistics, "What is the FBI afraid of?" (Kaye 2009). Ultimately, better research, better methodologies, and more reliable match statistics help the prosecution and defendants alike in the quest for accuracy in criminal trials.

4. Conclusion

Convincing state and federal laboratories to allow greater access to government-controlled DNA offender databases, through heightened engagement of the forensic science community by the statistical profession and through better funding of independent research, will lead to one of two results: either "greater confidence in the method now used to estimate RMPs," or to "some revised, but more defensible form of these estimates" (Kaye 2009). In the same respect, allowing litigants access to these databases to facilitate post-conviction innocence claims when they have demonstrated

that they are excluded as contributors of material DNA evidence will similarly lead to greater accuracy and closure with respect to resolving the litigants' factual claims. As the NAS Report recognizes, each wrongful conviction "based on improperly interpreted [forensic] evidence is serious, both for the innocent person and also for society, because of the threat that may be posed by a guilty person going free" (NAS Report 2009). Either way, the "importance of DNA" to criminal trials is "too high to continue the policy of ignoring or keeping relevant data secret and unexplored" (Kaye 2009). We urge the statistical profession to join this discourse and encourage further data sharing in all aspects of forensic science.

Acknowledgements

The authors wish to thank Jennifer Friedman, Christine Funk, Jason Tulley for helpful case-related information, as well as Erin Murphy for her thoughtful comments.

References

Brady v. Maryland, 373 U.S. 83 (1963).

Budowle, B., Baechtel, S., & Chakraborty, R. 2009 Partial Matches in Heterogeneous Offender Databases Do Not Call into Question the Validity of Random Match Probability Calculations. *Int'l J. Legal Med.*, **123**, 59-63.

CODIS-NDIS Statistics, Federal Bureau of Investigation, http://www.fbi.gov/hq/lab/codis/clickmap.htm.

Cole, S.A. & Lynch, M. 2006 The Social and Legal Construction of Suspects. *Ann. Rev. of L. & Soc. Sci.*, **2**, 39-60.

Cole, S. A. 2001 *Suspect Identities: A History of Fingerprinting and Criminal Identification*, Harvard Univ. Press, Cambridge, Mass.

Coleman v. Bradshaw, No. 3:03cv299 (S.D. Ohio).

Devlin, K. 2006 *Damned Lies*, Devlin's Angle, Mathematical Association of America, http://www.maa.org/devlin/devlin_10_06.html.

District Attorney's Office for the Third Judicial District v. Osborne, No 08-6, 557 U.S. ---- (2009), slip op. at 8.

Editorial, *Should We Trust DNA?*, San Francisco Chronicle, July 28, 2008, at B4.

Felch, J. & Dolan, M., *How Reliable Is DNA in Identifying Suspects?*, Los Angeles Times, July 20, 2008, http://www.latimes.com/news/local/la-me-dna20-2008july20,0,5133446.story.

Gilder, J. 2008 What is necessary (and unnecessary) for analyses of offender databases. http://www.bioforensics.com/conference08/DB_Analysis/index.html.

Gilder, J. 2005 Declaration of Jason R. Gilder, M.S., *People v. Davis*, No. MCN 2122087/SCN 190226 (Cal. Super. Ct. Dec. 13, 2005).

Hope, C. 2008 Millions of Profiles from DNA Database Passed to Private Firms, *Daily Telegraph*,http://www./telegraph.co.uk/news/newstopics/politics/lawandorder/2459976.

Illinois v. Griffin, No. 00 CR 16901 (Cook County Cir. Ct. Ill.).

Just, R.S., Irwin, J. A., O'Callaghan, J. E., Saunier, J. L., Coble, M. D., Vallone, P. M., Butler, J. M., Barritt, S. M., Parsons, T. J. 2004 Toward Increased Utility of mtDNA in Forensic Identifications. *Forensic Sci. Int'l*, **146S**: S147-49.

Kaye, D. 2009 Trawling DNA Databases for Partial Matches: What Is the FBI Afraid of? *Cornell J. L. & Pub. Pol'y*, **19**, 1 (forthcoming).

Kittles, R., Kaestle, F., Roth, A., & Ungvarsky, J. 2006 Database Limitations on the Evidentiary Value of Forensic Mitochondrial DNA Evidence, *Amer. Crim. L. Rev.*, **43**, 53-88.

Konzak, K. C. 2005 Motion to Quash SDT – Supplemental Declaration of Kenneth C. Konzak, *People v. Davis*, No. MCN 2122087/SCN 190226 (Cal. Super. Ct. Dec. 5, 2005).

Lynch, M., Cole, S. A., McNally, R., & Jordan, K. 2008 *Truth Machine: The Contentious History of DNA Fingerprinting*. Univ. of Chicago Press, Chicago.

McCartney, C. 2006 *Forensic Identification and Criminal Justice: Forensic Science, Justice, and Risk.* Willan, Devon, U.K.

Mueller, L. D. 2008 Can Simple Population Genetic Models Reconcile Partial Match Frequencies Observed in Large Forensic Databases? *J. Genetics*, **87**, 101-08.

Murphy, E. 2009 *Give Scholars Access to the National DNA Database*, S. F. Chronicle.

National Research Council, National Academic of Sciences, 2009 *Strengthening Forensic Science in the United States: A Path Forward*. National Academy Press, Washington, D.C.

Paoletti, D. R., Doom, T. E., Raymer, M. L., Krane, D. 2006 Assessing the Implications for Close Relatives in the Event of Similar But Nonmatching DNA Profiles. *Jurimetrics*, **46**, 161-75.

Parson, W. & Dür, A. 2007 EMPOP – A Forensic mtDNA Database. *Forensic Sci. Int'l*, **Genetics 1**, 88-92.

People v. Reeves, 109 Cal. Rptr. 2d 728 (Cal. Ct. App. 2001).

People v. Alejandro Santana, Crim. No. 00F06961 (Sacramento County Sup. Ct. May 27, 2008)

Rivera v. Mueller, 596 F. Supp. 2d 1163 (N.D. Ill. 2009).

Song, Y.S., Patil, A., Murphy, E., & Slatkin, M. 2009 The Average Probability That a Cold Hit in a DNA Database Results in Erroneous Conviction, *J. Forensic Sci.* **54**, 22-27.

Thompson, W. C. 2008 The Potential for Error in Forensic DNA Testing (and How That Complicates the Use of DNA Databases for Criminal Identification), *Council for Responsible Genetics*, http://www.councilforresponsiblegenetics.org/pageDocuments/H4T5EOYUZI.pdf.

Ungvarsky, E. J. 2007 *What Does One in a Trillion Mean?*, *Genewatch*, **20**, 1.

United States v. Blackmon, No. 2008-CF1-21355 (D.C. Super. Ct. Mar. 11, 2009)

Weir, B. S. 2001 DNA Match and Profile Probabilities: Comment on Budowle et al. (2000) and Fung and Hu (2000), *Forensic Sci. Comm.*, **3**, 1.