J Forensic Sci, January 2009, Vol. 54, No. 1
doi: 10.1111/j.1556-4029.2008.00917.x
Available online at: www.blackwell-synergy.com

Yun S. Song,[1] Ph.D.; Anand Patil,[2] Ph.D.; Erin E. Murphy,[3] J.D.; and Montgomery Slatkin,[4] Ph.D.

Average Probability that a "Cold Hit" in a DNA Database Search Results in an Erroneous Attribution

ABSTRACT: We consider a hypothetical series of cases in which the DNA profile of a crime-scene sample is found to match a known profile in a DNA database (i.e., a "cold hit"), resulting in the identification of a suspect based only on genetic evidence. We show that the average probability that there is another person in the population whose profile matches the crime-scene sample but who is not in the database is approximately $2(N - d)p_A$, where N is the number of individuals in the population, d is the number of profiles in the database, and p_A is the average match probability (AMP) for the population. The AMP is estimated by computing the average of the probabilities that two individuals in the population have the same profile. We show further that if *a priori* each individual in the population is equally likely to have left the crime-scene sample, then the average probability that the database search attributes the crime-scene sample to a wrong person is $(N - d)p_A$.

KEYWORDS: forensic science, population genetics, DNA fingerprinting, database search

In just over 20 years, DNA typing has emerged as a powerful forensic tool in criminal cases (1). Initially, DNA typing was used to bolster the case against a suspect identified through traditional means of investigation and evidence gathering, because the match of the DNA profile (i.e., the genotype at several genetic loci) of an incriminating crime-scene sample to a known suspect's profile provides strong corroborative evidence of the suspect's guilt. But the probative value of DNA evidence is no longer limited to this confirmatory role. In the United States, crime-scene profiles are routinely compared against profiles kept in state and federal databases, known as the Combined DNA Index System (CODIS) (2). Technically, CODIS refers to the software used to search the databases (2). At present, 13 tetranucleotide short tandem repeat (STR or microsatellite) CODIS loci on 12 chromosomes are customarily typed in the United States (2).

CODIS databases contain DNA profiles of two general kinds: known profiles and crime-scene profiles. The sources of the known profiles include voluntary submissions and mandatory contributions from certain convicted offenders and in some cases arrestees (3). Crime-scene profiles are collected by crime-scene technicians and are evidentiary samples connected to unsolved offenses according to certain standards related both to the quality of the evidence and the clarity of its connection to the crime (3). Both categories have grown rapidly in the past 5 years. As of January 2008, the national database held over 5 million offender profiles and over 190,000 crime-scene profiles (2).

As the CODIS databases have expanded, law enforcement has increasingly turned to DNA typing as an investigatory tool. In a "database search," the DNA profile from a crime-scene sample is compared to the profiles in databases to determine if a match exists. Such matches, whether offender-to-scene or scene-to-scene, are called "cold hits" (3). The number and frequency of cold hits has grown rapidly. In November 2004, the FBI reported a total of 19,500 cold hits, including both scene-to-scene and offender-to-scene matches (4); by March 2007 that number had risen to over 47,000 (5). States have experienced similar growth. For example, it took the state of Virginia from 1993 to 2001 to reach its first 1000 cold hits, but the second thousand occurred within the following 18 months (6).

Cold hits raise several legal and scientific questions. Because it is possible to recover biological material from well-preserved evidence, many cold-hit cases have arisen from offenses that occurred years or even decades ago. In fact, the federal government has funded programs to encourage states to reopen unsolved cases to determine whether a conviction can now be secured (7). Unfortunately, clear data on the number, frequency, and outcome of cold-hit cases are difficult to obtain. Law enforcement and prosecutorial functions are highly diffused in the American criminal justice system, and there is no centralized authority to track the outcome of cold hits. Some information exists, however. In Virginia, a survey of the outcome of the first 1000 cold hits revealed that 100 resulted in convictions through plea or trial, seven yielded not guilty verdicts, and 53 were never prosecuted; 752 were pending at the time of the survey (6).

Cold-hit cases have prompted courts to confront the question of whether a genetic match constitutes sufficient evidence to uphold a conviction. By comparison, the Supreme Court has previously ruled that a confession, standing alone, cannot serve as the sole basis for conviction (8). Although it has not yet been fully decided, most courts have resolved the cold-hit question in the affirmative. As one court observed, "the perils of eyewitness identification testimony far exceed those presented by DNA expert testimony" (9). And another, while recognizing that DNA evidence is not "infallible," explained that "[v]irtually no evidence is absolutely conclusive" (10). Of course, cold-hit cases raise justice-related concerns, especially since mounting a defense to a crime that occurred in the past becomes increasingly difficult as time progresses.

[1] Departments of EECS and Statistics, University of California, Berkeley, CA 94720-1776.

[2] Department of Zoology, University of Oxford, The Tinbergen Building, South Parks Road, Oxford, OX1 3PS, U.K.

[3] School of Law, University of California, Berkeley, CA 94720-7200.

[4] Department of Integrative Biology, University of California, Berkeley, CA 94720-3140.

Received 6 Oct. 2007; and in revised form 24 Jan. 2008; accepted 2 Aug. 2008.

Cold-hit cases can vary markedly in terms of the evidentiary value of the DNA match itself, particularly in light of the existence or quality of other corroborating evidence. Such evidence may be weak—for example, information that the defendant lived near the victim (11)—or strong, for example, that the defendant matches a detailed physical description given by the victim at the time of the offense. In *People v. Johnson* (10), a 15-year-old rape victim provided descriptions of the assailant's car and physical features, including distinctive tattoos, but no suspect was identified. Five years later, a cold hit identified the defendant, and subsequent investigation revealed that he had lived in the area at the time of the offense, had matching tattoos, and owned a vehicle that fit the description given by the victim.

Courts across the country have also upheld convictions based only upon cold-hit evidence. In cases of sexual assault, courts have reasoned that the intimate nature of the sample forecloses arguments that it might have been left accidentally or inadvertently (12–14). For example, in *State v. Hunter*, an appellate court upheld a rape conviction based on a semen sample collected after a sexual assault in 1995, which was matched 8 years later to the defendant, although "literally no other evidence" linked him to the crime (15). In upholding a rape conviction based on DNA evidence alone, another court explained, "we cannot conceive of an innocent reason for the defendant's DNA to be found on swabs taken from the victim's anal area." At the same time, however, the court specifically disclaimed "an iron-clad principal [sic] that DNA evidence, without corroboration, is always sufficient," since "[p]ractically infinite factual variations can arise" (16).

Some courts have upheld prosecutions based on cold hits not only in the absence of additional evidence but also in the face of contrary evidence. For instance, in Michigan, the government tested a biological sample collected from the hand of the victim in a 36-year-old murder case. The test revealed two profiles: one of the man ultimately prosecuted for the offense and another of an individual who was 4 years old at the time of the murder. Despite the unexplained presence of the second profile and the absence of any additional evidence, the jury convicted the defendant (17). Likewise, in *United States v. Jenkins*, the court sanctioned a murder and burglary prosecution based largely on genetic evidence, even though a day after the incident another man was found in possession of the decedent's credit cards and other items taken from the home (18). Although the prosecution was allowed to proceed, the jury ultimately could not reach a unanimous verdict (19).

DNA evidence carries such persuasive power in court because the random match probability (RMP), defined to be the probability that a person picked at random has the same DNA profile as the evidentiary sample, is very low if several unlinked loci are typed. For a 13-locus CODIS profile, typical RMPs are on the order of 10^{-14} to 10^{-15} (20). Such a low RMP implies that a particular DNA profile has a high probability of being unique (21,22), although the lack of certainty makes claims of uniqueness improper to make in the presentation of DNA evidence in court.

To compute the RMP, the recommendation of the second National Research Council Report (NRC II) (23) is usually followed. Allele frequencies at each locus are estimated for the ethnic group of the suspect, whether identified by other evidence or by a database search. From the estimated allele frequencies at each locus, the probability that a randomly chosen individual from the same ethnic group has the same genotype at that locus is computed using the "theta correction":

$$\Pr(\mathrm{A}_i\mathrm{A}_i|\mathrm{A}_i\mathrm{A}_i) = \frac{[2\theta + (1-\theta)p_i][3\theta + (1-\theta)p_i]}{(1+\theta)(1+2\theta)}, \qquad (1a)$$

$$\Pr(\mathrm{A}_i\mathrm{A}_j|\mathrm{A}_i\mathrm{A}_j) = \frac{2[\theta + (1-\theta)p_i][\theta + (1-\theta)p_j]}{(1+\theta)(1+2\theta)}, i \neq j \qquad (1b)$$

(equation 4.10, Ref. 24; equation 8.1, Ref. 23). In Eq. (1), A_i represents the ith allele at the locus, p_i is the estimated frequency of that allele in the same ethnic group, and θ is a parameter that takes account of deviations from Hardy–Weinberg frequencies caused by population subdivision and other factors. NRC II recommended that $\theta = 0.01$ be used for most ethnic groups in the United States and that $\theta = 0.03$ be used for Native Americans. The probability that a randomly chosen individual from the same ethnic group has the same genotype at all loci, which is the RMP, is obtained by multiplying the per-locus probabilities.

Although there is a consensus about using the recommendation of NRC II for computing the RMP, there remains a controversy about how to present the RMP as evidence when a suspect has been identified by a database search. NRC II (24) recommended that the RMP be multiplied by the number of profiles in the database searched, resulting in a higher but still small probability of a match if the suspect were not the source of the crime-scene sample. This recommendation has been supported by some later commentators (25) but criticized by others who have argued that a single match in a database search provides stronger evidence that the suspect is the source than suggested by NRC II because everyone else in the database can be excluded (26–28).

Regardless of what probability is attached to a cold hit, such evidence has sufficed to convict defendants in many cases. Our goal here is to determine the consequence of this practice by asking what happens if cold-hit evidence becomes a regular basis for conviction. We consider a hypothetical population containing N individuals. For the U.S. population, N is approximately 300 million, but it may be appropriate to consider subsets such as only males or only individuals within a specified age range. We assume there is a series of cases in which a crime is committed by one individual in the population, which makes our model equivalent to the model used in the "island problem" (23). There is a database containing d DNA profiles of individuals randomly chosen from the population. The crime-scene profile matches only one profile in the database.

We can ask two different but closely related questions. The first is: what is the average probability that at least one individual not in the database but in the population also has the same profile as the crime-scene sample? The second is: what is the average probability of an erroneous attribution, meaning that the crime was actually committed by someone in the population whose profile is not in the database?

The answers to both questions are derived formally in the Appendix. The average probability that someone not in the database but in the population has the same profile is approximately $2(N - d)p_A$, where p_A is the average match probability (AMP), defined to be the probability that two randomly chosen individuals in the population have the same DNA profile. The AMP differs from the RMP. The RMP is the probability that the DNA profile of another individual matches a particular DNA profile. Therefore, the probability that a randomly chosen individual from the population has the same DNA profile as a crime-scene sample in a particular case is the RMP. It is the RMP that is the relevant statistic to present as evidence in a case involving DNA evidence. The AMP is the average of the RMPs of all profiles in the population. It is not a legally relevant statistic in any particular case, and therefore should not be presented as evidence. Rather, the AMP is a probability relevant to determining the reliability of a series of cases each involving DNA evidence, as we are doing here. It is the difference between asking whether a particular defendant was erroneously attributed to be the source of the crime-scene sample and assessing the average rate at which erroneous attribution is likely to occur in a series of cases.

To find the probability of an erroneous attribution, we need an additional assumption, namely that, *a priori*, each individual in the population is equally likely to have committed the crime. That assumption is obviously not true but is made in such calculations to make the results conservative (21,23,24,29).

The probability of an erroneous attribution also depends on the AMP as $(N - d)p_A$, which is half the probability that an individual not in the database has the same profile. The factor of one half reflects the assumption that, in the absence of other information, the two individuals with the same profile are equally likely to have been the source of the crime-scene sample. This assumption is particularly appropriate for a cold-hit case, in which there is little or no evidence beyond the genetic evidence to support the belief that a particular defendant is guilty.

To compute the AMP for an ethnically mixed population, it is necessary to take account of differences in allele frequency among subgroups and the numbers of individuals in each subgroup. In the Appendix, we estimate the AMP for 13-locus CODIS profiles of the U.S. population to be approximately 9.94×10^{-16}. We used Eq. (1) for groups for which CODIS allele frequencies are available. If we assume $N = 300$ million and $d = 5$ million, we obtain the probability of an erroneous attribution of approximately 2.93×10^{-7} or approximately 1 in 3.4 million. Assuming a smaller N will result in a smaller probability.

We conclude then that, if the recommendation of NRC II is used as the basis for computing the AMP, the chance of an erroneous attribution is very small even under the conservative assumptions we have made. The recommendation of NRC II has two parts, the match probabilities computed from Eq. (1) for each locus and multiplication of those probabilities across loci. It is relatively easy to test for the appropriateness of Eq. (1). Many methods have been developed and applied (23), leading to a consensus that the theta correction using the recommended values of theta provides a conservative basis for computing the per-locus match probabilities. Testing for appropriateness of multiplying across loci is more difficult and has been done only with sample sizes much smaller than the current sizes of the offender databases (24,30). The largest such study is by Weir (31), who examined an ethnically mixed data set of over 12,000 nine-locus profiles assembled by the Australian forensic agencies. He computed the numbers of individuals that matched at one or both alleles at one to nine loci, and compared the observations with predictions. The probability that both alleles at a locus were the same was computed from Eq. (1) with $\theta = 0.01$, and the probability that one of two alleles were the same was computed using an expression derived in the appendix of reference (31). The probabilities of matches of one or two alleles were obtained from a multinomial distribution, which assumes independence across loci. Weir (31) found good agreement between observed and predicted numbers when $\theta = 0.01$ was used, even though significant deviations from Hardy–Weinberg proportions were detectable at each locus.

As has been pointed out (23,32,33), failure to reject the hypothesis of independence across loci is not equivalent to verifying that the RMP computed from the product rule is accurate. Only by carrying out more studies of the type done by Weir (31) can we be sure that multiplying across loci produces accurate results in data sets of size comparable to those of the offender databases. Furthermore, the theories accounting for the uncertainty that arises in interpreting DNA evidence have just begun to be developed (32,34).

Although our result appears to support existing practice, it also calls attention to the critical role of the assumption of independence across loci in justifying the use of cold hits to obtain convictions. Even though the AMP is small, the risk that the wrong individual is identified as a suspect depends on the product of the AMP and the number of individuals not in the offender databases, which is a very large number. Slight deviations from independence across loci could result in a higher AMP. For example, if the AMP were as large as 10^{-9}, our calculations show there is a considerable risk that someone not in the database has the same profile. Given the importance of the assumption of independence across loci in this context, continued testing of that hypothesis is recommended.

Acknowledgement

YSS has been supported in part by NIH Grant K99/R00-GM080099 and MS has been supported in part by NIH Grant R01-GM40282.

References

1. Jeffreys AJ. Genetic fingerprinting. Nat Med 2005;11:1035–9.
2. http://www.fbi.gov/hq/lab/codis/clickmap.htm.
3. Butler J. Forensic DNA typing. Burlington, MA: Elsevier Academic Press, 2005.
4. NDIS Statistics, FBI, Measuring Success, 2005. Available at http://www.fbi.gov/hq/lab/codis/success.htm, accessed January 5, 2005.
5. NDIS Statistics, FBI, Measuring Success, 2007. Available at http://www.fbi.gov/hq/lab/codis/success.htm, accessed May 21, 2007.
6. Murphy E. The new forensics: criminal justice, false certainty, and the second generation of scientific evidence, 56 Cal. L. Rev. 721, 2007:742.
7. Justice for All Act of 2004, codified at 42 U.S.C. § 3797k *et seq*.
8. *Smith v. United States*, 348 U.S. 147, 152 (1954).
9. *Roberson v. State*, 16 S.W.3d 156, 170 (Tex. Crim. App. 2000).
10. *People v. Johnson* 139 Cal.App.4th 1135 (Cal. Ct. App. 2006).
11. *Riggs v. State*, 809 N.E.2d 322 (Ind. 2004).
12. *State v. Davis,* 698 N.W.2d 823 (Wisc. 2005).
13. *State v. Toomes*, 191 S.W.3d 122 (Ct. Crim. App. Tenn. 2005)
14. *State v. Labarbera*, 115 P.3d 1038 (Wash. Ct. App. 2005).
15. *State v. Hunter*, 861 N.E.2d 898 (Ohio App. Ct. 2006).
16. *State v. Toomes*, 191 S.W.3d 122, 131 & n.4 (Ct. Crim. App. Tenn. 2005).
17. Convicted Murderer Seeks Retrial, Kalamazoo Gazette, May 10, 2006; Sect. Local News.
18. *United States v. Jenkins*, 887 A.2d 1013 (D.C. 2005).
19. Cauvin HE. Jury deadlocks in case that relied on DNA. Washington Post, April 15, 2006; Sect. B:10.
20. Chakraborty R, Stivers DN, Su B, Zhong YX, Budowle B. The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. Electrophoresis 1999;20:1682–96.
21. Budowle B, Chakraborty R, Carmody G, Monson KL. Source attribution of a forensic DNA profile. Forensic Sci Com 2000;3.
22. Balding DJ. When can a DNA profile be regarded as unique? Sci Justice 1999;39:257–60.
23. Evett IW, Weir BS. Interpreting DNA evidence. Sunderland, MA: Sinauer, 1998.
24. National Research Council Committee on DNA Technology in Forensic Science. The evaluation of forensic DNA evidence. Washington, DC: National Academy Press, 1996.
25. Stockmarr A. Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search. Biometrics 1999;55:671–7.
26. Donnelly P, Friedman RD. DNA database searches and the legal consumption of scientific evidence. Michigan Law Rev 1999;97:931–84.
27. Balding DJ. The DNA database search controversy. Biometrics 2002;58:241–4.
28. Balding DJ, Donnelly P. Evaluating DNA profile evidence when the suspect is identified through a database search. J Forensic Sci 1996;41:603–7.

29. Balding DJ, Donnelly P. Inference in forensic identification. J R Stat Soc Series A 1995;158:21–53.
30. Holt CL, Stauffer C, Wallin JM, Lazaruk K, Nguyen T, Budowle B, et al. Practical applications of genotypic surveys for forensic STR testing. Forensic Sci Int 2000;112:91–109.
31. Weir BS. Matching and partial-matching DNA profiles. J Forensic Sci 2004;49:1009–14.
32. Curran JM, Buckleton JS, Triggs CM, Weir BS. Assessing uncertainty in DNA evidence caused by sampling effects. Sci Justice 2002;42:29–37.
33. Triggs CM, Buckleton JS. Logical implications of applying the principles of population genetics to the interpretation of DNA profiling evidence. Forensic Sci Int 2002;128:108–14.
34. Triggs CM, Curran JM. The sensitivity of the Bayesian HPD method to the choice of prior. Sci Justice 2006;46:169–78.
35. Budowle B, Shea B, Niezgoda S, Chakraborty R. CODIS STR loci data from 41 sample populations. J Forensic Sci 2001;46:453–89.
36. U.S. Census Bureau, Census 2000 Summary File 1; generated by Anand Patil; using American Factfinder; http://www.factfinder.census.gov/, accessed March 11, 2007.
37. U.S. Census Bureau, Census 2000 Summary File 3; generated by Anand Patil; using American Factfinder; http://www.factfinder.census.gov/, accessed March 11, 2007.
38. U.S. Census Bureau. We the People: American Indians and Alaska natives in the United States. Washington, DC: Census 2000 Special Reports, 2002.
39. U.S. Census Bureau. The Asian population: 2000. Washington, DC: Census Brief 9, 2002.
40. *People v. Rush*, 672 N.Y.S.2d 362, 363 (App. Div. 1998).

Additional information and reprint requests:
Montgomery Slatkin, Ph.D.
Department of Integrative Biology
University of California
Berkeley, CA 94720-3140
E-mail: slatkin@berkeley.edu

Appendix

Derivation of Average Probabilities

Suppose that a population of size N is partitioned into two subsets D and R, where D corresponds to the set of individuals in a database and R the rest of the population. The sizes of D and R are denoted by d and r, respectively.

We assume that a crime is committed by exactly one individual in the population. In what follows, we use C to denote the actual criminal. Note that either $C{\in}D$ or $C{\in}R$, but not both. Suppose that the DNA profile $\Phi(C)$ of C has been obtained from the crime scene, and let $E_X{}^k$ denote the event that *exactly* k individuals in the set X have DNA profiles that match $\Phi(C)$. The notation E_X (with no superscript) denotes the event that the DNA profile of *at least* one individual in X matches $\Phi(C)$.

Existence of Matching DNA Profiles Outside a Database

Let Q denote a probability distribution which takes the dependency of DNA profiles into account, and q a probability distribution which assumes that DNA profiles are independent. Analytically computing $Q[E_R|E_D{}^1]$ is a challenging problem, so we wish to approximate it by $q[E_R|E_D{}^1]$. First, we show that $q[E_R|E_D{}^1]$ is a lower bound on $Q[E_R|E_D{}^1]$. Note that

$$\begin{aligned} Q[E_R|E_D^1] &= Q[E_R|E_D^1 \wedge (C \in D)]P[C \in D] \\ &\quad + Q[E_R|E_D^1 \wedge (C \in R)]P[C \in R]. \end{aligned}$$

The probabilities $P[C \in D]$ and $P[C \in R]$ can be viewed as prior probabilities, and we assume

$$P[C \in D] = \frac{d}{r+d} \text{ and } P[C \in R] = \frac{r}{r+d}. \qquad (2)$$

Since $Q[E_R|E_D{}^1{\wedge}(C{\in}R)] = q[E_R|E_D{}^1{\wedge}(C{\in}R)] = 1$, to show $q[E_R|E_D{}^1] \le Q[E_R|E_D{}^1]$, we need to show $q[E_R|E_D{}^1{\wedge}(C \in D)] \le Q[E_R|E_D{}^1{\wedge}(C{\in}D)]$. Rewrite $Q[E_R|E_D{}^1{\wedge}(C{\in}D)]$ as follows:

$$\begin{aligned} Q[E_R|E_D^1 \wedge (C \in D)] &= 1 - Q[E_R^0|E_D^1 \wedge (C \in D)] \\ &= 1 - \frac{Q[E_R^0 \wedge E_D^1 \wedge (C \in D)]}{Q[E_D^1 \wedge (C \in D)]} \\ &= 1 - \frac{Q[E^0_{(R\cup D)\setminus\{C\}}|C \in D]}{Q[E^0_{D\setminus\{C\}}|C \in D]}. \end{aligned}$$

Now, we are concerned with the case in which dependency of DNA profiles increases the probability of observing matching DNA profiles, so we conclude

$$\begin{aligned} &Q[E_R^0|C \in D] \leqslant q[E_R^0|C \in D], \\ &Q[E^0_{(R\cup D)\setminus\{C\}}|C \in D] \leqslant Q[E_R^0|C \in D]Q[E^0_{D\setminus\{C\}}|C \in D], \end{aligned}$$

while independence implies

$$q[E^0_{(R\cup D)\setminus\{C\}}|C \in D] = q[E_R^0|C \in D]q[E^0_{D\setminus\{C\}}|C \in D].$$

Using the above results, we obtain

$$\begin{aligned} \frac{Q[E^0_{(R\cup D)\setminus\{C\}}|C \in D]}{q[E^0_{(R\cup D)\setminus\{C\}}|C \in D]} &\leqslant \frac{Q[E_R^0|C \in D]Q[E^0_{D\setminus\{C\}}|C \in D]}{q[E_R^0|C \in D]q[E^0_{D\setminus\{C\}}|C \in D]} \\ &\leqslant \frac{Q[E^0_{D\setminus\{C\}}|C \in D]}{q[E^0_{D\setminus\{C\}}|C \in D]}, \end{aligned}$$

which implies $q[E_R|E_D{}^1{\wedge}(C{\in}D)] \le Q[E_R|E_D{}^1{\wedge}(C{\in}D)]$. Therefore, we conclude that $q[E_R|E_D{}^1]$ is a lower bound on $Q[E_R|E_D{}^1]$.

In what follows, we obtain an analytic expression for $q[E_R|E_D{}^1]$. First, rewrite $q[E_R{}^k|E_D{}^1]$ as follows:

$$\begin{aligned} q[E_R^k|E_D^1] &= \frac{q[E_R^k{\wedge}E_D^1{\wedge}(C{\in}D)]+q[E_R^k{\wedge}E_D^1{\wedge}(C{\in}R)]}{q[E_D^1{\wedge}(C{\in}D)]+q[E_D^1{\wedge}(C{\in}R)]} \\ &= \frac{q[E_R^k{\wedge}E_D^1|C{\in}D]\times P[C{\in}D]+q[E_R^k{\wedge}E_D^1|C{\in}R]\times P[C{\in}R]}{q[E_D^1|C{\in}D]\times P[C{\in}D]+q[E_D^1|C{\in}R]\times P[C{\in}R]}. \end{aligned} \qquad (3)$$

Let p denote the RMP for $\Phi(C)$. Then, for $k > 0$, note that

$$\begin{aligned} q[E_D^1|C \in D] &= (1-p)^{d-1}, \\ q[E_R^k \wedge E_D^1|C \in D] &= (1-p)^{d-1}q[E_R^k|C \in D] \\ &= (1-p)^{d-1}\binom{r}{k}p^k(1-p)^{r-k}, \\ q[E_R^k \wedge E_D^1|C \in R] &= \binom{r-1}{k-1}p^{k-1}(1-p)^{r-k}q[E_D^1|C \in R], \\ q[E_D^1|C \in R] &= dp(1-p)^{d-1}. \end{aligned}$$

Using Eq. (2) and the above formulas, Eq. (3) can be written as

$$q[E_R^k|E_D^1] = \frac{\binom{r}{k}p^k(1-p)^{r-k}(1+k)}{1+rp}, \qquad (4)$$

from which it follows that

$$q[E_R|E_D^1] = \sum_{k=1}^{r} q[E_R^k|E_D^1] = \frac{1 + rp - (1-p)^r}{1 + rp}.$$

For $p, rp \ll 1$,

$$q[E_R|E_D^1] \approx 2rp = 2(N-d)p.$$

This expression is for a particular value of RMP p associated with a particular profile $\Phi(C)$. Averaging Eq. (5) over the probability distribution of $\Phi(C)$ gives $2(N - d)p_A$, where p_A denotes the AMP. The distribution of $\Phi(C)$ is likely to be quite complex; in Section Computation of the AMP, we provide an estimate of p_A based on assumption (2).

Erroneous Attribution

Let G denote the event that the sole individual in D whose DNA profile matches $\Phi(C)$ is the actual criminal C. If n individuals in the population have DNA profiles that match $\Phi(C)$, then assume that each one of those n individuals is equally likely to be the actual criminal C. Under this assumption, the probability of G given $E_D{}^1$ is

$$P[G|E_D^1] = \sum_{k=0}^{r} \frac{1}{1+k} P[E_R^k|E_D^1],$$

and the probability of an erroneous attribution is

$$P[\text{Erroneous Attribution}|E_D^1] = 1 - P[G|E_D^1] = \sum_{k=1}^{r} \frac{k}{k+1} P[E_R^k|E_D^1].$$

Using Eq. (4) as an approximation of $P[E_R{}^k|E_D{}^1]$, we obtain

$$P[\text{Erroneous Attribution}|E_D^1] = \frac{rp}{1+rp}. \qquad (5)$$

Hence, for $p, rp \ll 1$,

$$P[\text{Erroneous Attribution}|E_D^1] \approx rp = (N-d)p.$$

As in the previous section, averaging this expression over the probability distribution of $\Phi(C)$ gives $(N - d)p_A$.

Computation of the AMP

In Section Derivation of Average Probabilities, we obtained several analytic results under the assumption that DNA profiles are independent. In this section, we describe the computation of the AMP p_A, taking the dependency between alleles into account.

Population Data

Estimates of population sizes for the groups considered by Budowle et al. (35) were obtained from the U.S. Census Bureau website. To avoid double-counting, only individuals reporting a single race or ethnicity were included in our analysis. In addition, only the atomic populations, as opposed to aggregated populations such as "General Asian," were considered. Populations outside the United States were excluded from our analysis. Budowle et al. "Japanese1" and "Japanese2" populations were excluded because the meaning of these categories is unclear. The Minnesota African-American, New York and Minnesota Caucasian, Minnesota Hispanic, and Minnesota Native American populations were excluded because the data for locus D16S539 were not available.

The estimated population sizes used in our analysis are summarized in Table A1. The estimated African-American, Caucasian, and Hispanic population sizes were obtained from Summary File 1 (36) of the 2000 U.S. Census, and the estimated Michigan Native American population size was obtained from Summary File 3 (37) of the same census. The estimated Native American population sizes for Navajo and Apache were obtained from the U.S. Census 2000 Special Report (38), while the estimated Asian population sizes were obtained from the U.S. Census 2000 Brief (39).

The "universe" of possible offenders was assumed to be the union of all populations listed in Table A1. We assumed that individuals in the CODIS database were drawn at random from that universe, such that each population is represented in the database proportional to its size.

Method

Suppose that there are m populations and that K_ℓ distinct alleles are observed at locus ℓ in the total population. Greek indices α, β $\in \{1,\ldots, m\}$ denote population labels, while Roman indices $i, j \in \{1,\ldots, K_\ell\}$ denote allele types at locus ℓ.

Let $p_{\alpha,i}{}^{(\ell)}$ denote the frequency of allele i at locus ℓ in population α. Then, the probability of sampling genotype A_i A_i at locus ℓ from population α and the analogous probability for sampling genotype A_i A_j are, respectively, given by

$$\begin{aligned} P(A_iA_i) &= p_{\alpha,i}^{(\ell)}[\theta_\alpha + (1-\theta_\alpha)p_{\alpha,i}^{(\ell)}] \quad \text{and} \\ P(A_iA_j) &= 2(1-\theta_\alpha)p_{\alpha,i}^{(\ell)}p_{\alpha,j}^{(\ell)} \end{aligned} \qquad (6)$$

where θ_α denotes the probability that two alleles randomly sampled from population α are identical by descent (Ref. [23]; bottom of page 119). Let $G_{\alpha,\beta}{}^{(\ell)}$ (respectively, $H_{\alpha,\beta}{}^{(\ell)}$) denote the probability that a random individual from population α and a random individual from population β both have homozygous (respectively, heterozygous) genotypes at locus ℓ that are the

TABLE A1—*Estimated population sizes within the United States used in our analysis.*

African-American		Caucasian		Hispanic		Asian		Native American	
Population	Size	Population	Size	Population	Size	Population	Size	Population	Size
Alabama	1,150,076	Alabama	3,125,819	Arizona	1,295,617	Chinese	2,314,537	Apache	57,199
California	2,181,926	California	15,816,790	California	10,956,556	Korean	1,076,872	Michigan	53,421
Florida	2,264,268	Florida	10,458,509	Florida	2,682,715	Vietnamese	1,122,528	Navajo	276,775
Illinois	1,856,152	Michigan	7,806,691	Michigan	323,877				
New York	2,812,623	Virginia	4,965,637	New York	2,867,583				
Virginia	1,376,378								

same. We compute ${G_{\alpha,\beta}}^{(\ell)}$ and ${H_{\alpha,\beta}}^{(\ell)}$ as follows. If $\alpha \neq \beta$, we assume that the probability of alleles from different populations being identical by descent is negligibly small and use Eq. (6) separately for each individual. If $\alpha = \beta$, on the other hand, we use Eqs (1) and (6) to compute the joint probabilities $P(A_i\ A_i, A_i\ A_i)$ and $P(A_i\ A_j, A_i\ A_j)$. More exactly, we use the following formulas for ${G_{\alpha,\beta}}^{(\ell)}$ and ${H_{\alpha,\beta}}^{(\ell)}$:

$$G^{(\ell)}_{\alpha,\beta} = \begin{cases} \sum_{i=1}^{K_\ell} p^{(\ell)}_{\alpha,i}[\theta_\alpha + (1-\theta_\alpha)p^{(\ell)}_{\alpha,i}]p^{(\ell)}_{\beta,i}[\theta_\beta + (1-\theta_\beta)p^{(\ell)}_{\beta,i}], & \text{if } \alpha \neq \beta, \\ \sum_{i=1}^{K_\ell} \frac{p^{(\ell)}_{\alpha,i}\left[\theta_\alpha+(1-\theta_\alpha)p^{(\ell)}_{\alpha,i}\right]\left[2\theta_\alpha+(1-\theta_\alpha)p^{(\ell)}_{\alpha,i}\right]\left[3\theta_\alpha+(1-\theta_\alpha)p^{(\ell)}_{\alpha,i}\right]}{(1+\theta_\alpha)(1+2\theta_\alpha)}, & \text{if } \alpha = \beta, \end{cases} \tag{7}$$

$$H^{(\ell)}_{\alpha,\beta} = \begin{cases} 4(1-\theta_\alpha)(1-\theta_\beta) \sum_{i=1}^{K_\ell - 1} \sum_{j=i+1}^{K_\ell} p^{(\ell)}_{\alpha,i} p^{(\ell)}_{\alpha,j} p^{(\ell)}_{\beta,i} p^{(\ell)}_{\beta,j}, & \text{if } \alpha \neq \beta, \\ \sum_{i=1}^{K_\ell - 1} \sum_{j=i+1}^{K_\ell} \frac{4(1-\theta_\alpha)p^{(\ell)}_{\alpha,i}p^{(\ell)}_{\alpha,j}\left[\theta_\alpha+(1-\theta_\alpha)p^{(\ell)}_{\alpha,i}\right]\left[\theta_\alpha+(1-\theta_\alpha)p^{(\ell)}_{\alpha,j}\right]}{(1+\theta_\alpha)(1+2\theta_\alpha)}, & \text{if } \alpha = \beta. \end{cases} \tag{8}$$

In our analysis, allele frequencies ${p_{\alpha,i}}^{(\ell)}$ were taken from Budowle et al. (35). As suggested there, θ_α was assumed to be 0.03 for Native American populations and 0.01 for all other populations.

Using the product rule, the probability $M_{\alpha,\beta}$ that a random individual from population α and a random individual from population β have matching genotypes at all 13 CODIS loci is given by

$$M_{\alpha,\beta} = \prod_{\ell=1}^{13} \left[G^{(\ell)}_{\alpha,\beta} + H^{(\ell)}_{\alpha,\beta} \right],$$

and the 13-locus AMP p_A is given by

$$p_A = \sum_{\alpha=1}^{m} \sum_{\beta=1}^{m} q_\alpha q_\beta M_{\alpha,\beta} \tag{9}$$

where q_α denotes the proportion of population α relative to the total population. Our Python code for carrying out the above computation is available upon request.