ASSESSING THE IMPLICATIONS
FOR CLOSE RELATIVES
IN THE EVENT OF SIMILAR
BUT NONMATCHING DNA PROFILES

David R. Paoletti, Travis E. Doom, Michael L. Raymer, and Dan E. Krane[*]

ABSTRACT: A complete match between the STR DNA profile of an evidence sample and that of an individual included in a database of profiles of convicted offenders has clear utility as an investigative tool. Very similar but nonetheless nonmatching DNA profiles also can provide useful information by suggesting that a close relative of the individual may be the source of the evidence sample. This article describes a general framework for determining the relative likelihood that an individual's close relative is the source of an imperfectly matching DNA profile.

CITATION: David R. Paoletti, Travis E. Doom, Michael L. Raymer, and Dan E. Krane, Assessing the Implications for Close Relatives in the Event of Similar but Nonmatching DNA Profiles, 46 Jurimetrics J. 161–175 (2006).

Short tandem repeat (STR) sequences have become the genetic markers of choice for human identification in forensic investigations.[1] Length polymorphisms, because of differences in the number of tandem repeats of four-nucleotide (tetranucleotide) core sequences, are detected after polymerase chain reaction

---

    * David R. Paoletti is a Doctoral Candidate, Department of Computer Science and Engineering, Wright State University. Travis E. Doom is Associate Professor, Department of Computer Science and Engineering, Wright State University. Michael L. Raymer is Assistant Professor, Department of Computer Science and Engineering, Wright State University. Dan E. Krane is Associate Professor, Department of Biological Sciences, Wright State University. Requests for additional information or reprints should be addressed to the last author (dan.krane@wright.edu).
    1. JOHN M. BUTLER, FORENSIC DNA TYPING: BIOLOGY, TECHNOLOGY, AND GENETICS OF STR MARKERS 4 (2d ed. 2005).

(PCR) amplification. A set of thirteen STR loci are typically genotyped with commercially available kits, and length polymorphisms are identified with machines such as the Applied Biosystems 310 or 3100 capillary electrophoresis systems.[2] These STR genotypes are easily archived in searchable databases such as the Combined DNA Index System (CoDIS).[3] Perfect matches between evidentiary material and a profile in such a database are known as "cold hits."[4]

Individuals with DNA profiles that differ from the profile observed in an evidence sample are simply excluded as being possible contributors to that sample[5] regardless of whether they are found not to match during a database search or a more conventional investigation. However, since STR alleles are inherited in a strictly Mendelian fashion, it is possible that the most likely explanation for a nearly perfect match is that the source of an evidence sample is a close relative of the individual whose DNA profile is available for comparison.

Consider the following case. In 2003, North Carolina's State Bureau of Investigation performed postconviction DNA testing on evidence samples collected from a victim who was raped and murdered in 1984. The resulting 13-locus STR genotype exonerated Darryl Hunt after he had already served 18 years of a life prison sentence. No perfect matches to any of the 40,000 individuals in the state's convicted-offender database were found, but the single best match was to Anthony Dennard Brown. Clearly, Anthony Dennard Brown was excluded as the rapist and murderer. But, were the 16 out of a possible 26 alleles that he shared with the perpetrator sufficiently unusual to implicate close relatives such as his brother, Willard Brown? North Carolina investigators thought so. They

---

2. JOHN M. BUTLER ET AL., NAT'L INST. OF STANDARDS & TECH., REVIEW FORENSIC DNA TYPING BY CAPILLARY ELECTROPHORESIS: USING THE ABL PRISM 310 AND 3100 GENETIC ANALYZERS FOR STR ANALYSIS 3–5 (2004), *available at* http://www.cstl.nist.gov/div831/strbase/pub_pres/Butler2004a.pdf.

3. Nearly three million complete, 13-locus STR-DNA profiles of convicted offenders have already been entered into the CoDIS database. Federal Bureau of Investigation Combined DNA Index System (CoDIS), http://www.fbi.gov/hq/lab/codis/index1.htm (last visited Feb. 14, 2006). Similar databases are maintained by foreign countries such as Great Britain and Australia. *See* The Forensic Science Service, http://www.forensic.gov.uk (forensic service used by police in England and Wales) (last visited Feb. 14, 2006); CrimTrac, National Criminal Investigation DNA Database, http://www.crimtrac.gov.au/dna.htm (forensic service used by Australian police department) (last visited Feb. 14, 2006).

4. The appropriate way to describe the significance of such a DNA profile match in such cases has been a topic of considerable debate. NATIONAL RESEARCH COUNCIL, COMMITTEE ON DNA TECHNOLOGY IN FORENSIC SCIENCE, DNA TECHNOLOGY IN FORENSIC SCIENCE 111–30 (1992) [hereinafter NRC I]; NATIONAL RESEARCH COUNCIL, COMMITTEE ON DNA FORENSIC SCIENCE: AN UPDATE, THE EVALUATION OF FORENSIC DNA EVIDENCE 166–211 (1996) [hereinafter NRC II]; David J. Balding, *Errors and Misunderstandings in the Second NRC Report,* 37 JURIMETRICS 469, 470–73 (1997); David J. Balding & Peter Donnelly, *Evaluating DNA Profile Evidence When the Suspect Is Identified Through a Database Search,* 41 J. FORENSIC SCI. 603, 604 (1996); IAN W. EVETT & BRUCE S. WEIR, INTERPRETING DNA EVIDENCE: STATISTICAL GENETICS FOR FORENSIC SCIENTISTS 219 (1998).

5. William C. Thompson et al., *Part 1: Evaluating Forensic DNA Evidence: Essential Elements in a Competent Defense Review*, 27 THE CHAMPION 16, 17 (2003) [hereinafter *Part 1*]; William C. Thompson et al., *Part 2: Evaluating Forensic DNA Evidence: Essential Elements in a Competent Defense Review*, 27 THE CHAMPION 24, 24 (2003) [hereinafter *Part 2*].

generated Willard Brown's DNA profile from cigarette butts he had discarded and found that his 13-locus STR profile corresponded perfectly with the evidence from the 1984 crime (to which he has subsequently pled guilty and is serving a life sentence plus ten years).[6]

Great Britain's Forensic Science Service refers to investigations such as those that led to the arrest of Willard Brown as "familial searches" and performs them routinely.[7] Policies regarding familial searches within the United States range from not allowing them at all (federal law bars the FBI from using information from all but perfect matches) to specifically encouraging them (as authorized by specific statutes in New York and Massachusetts).[8] When familial searches are permitted, the requisite threshold of similarity tends to be ambiguously defined and described in terms such as matches needing to "be very, very close" (Virginia), "appear useful" (California), or be at 21 or more out of 26 alleles (Florida).[9]

Many researchers have considered the probability that a perfectly matching suspect has the same DNA profile as a close relative, such as a nontwin brother,[10] though only Sjerps and Kloosterman specifically consider when a similar but not perfectly matching genotype is sufficient grounds to demand DNA samples from relatives.[11] Their foundational work, however, does not explicitly provide a general formula for the likelihood that a relative of an excluded suspect is a contributor to a specific evidence profile. Furthermore, the approach described by Sjerps and Kloosterman[12] does not consider the size of the alternative suspect pool. Here we consider the effect of the alternative suspect pool and provide explicit formulae for calculating likelihood ratios of the actual perpetrator being a sibling or a parent or child (versus a randomly chosen, unrelated member of the population) for an arbitrary pair of suspect and evidence profiles. We also describe the results of simulations that provide statistical bounds on both the number and rarity of the alleles shared between an evidence sample and an

---

6. Richard Willing, *Suspects Get Snared by a Relative's DNA*, USA TODAY, June 8, 2005, at 1A.

7. Press Release, The Forensic Science Service, DNA Technology to Progress More Cold Cases (Dec. 8, 2004) (on file with Lucy Stokes, Press Officer), *available at* http://www.forensic.gov.uk/ forensic_t/inside/news/list_press_release.php?case=31&y=2004; Press Release, The Forensic Science Service, Double the Number of DNA Profiles Processed with Automation (Nov. 3, 2004) (on file with Lucy Stokes, Press Officer), *available at* http://forensic.gov.uk/forensic_t/inside/news/list_press _release.php?case=28&y=2004.

8. Willing, *supra* note 6.

9. *Id*.

10. NRC I, *supra* note 4, at 86–87; David J. Balding & Richard A. Nichols, *DNA Profile Match Probability Calculation: How to Allow for Population Stratification, Relatedness, Database Selection and Single Bands*, 64 FORENSIC SCI. INT'L. 125, 128 (1994); John F.Y. Brookfield, *The Effect of Relatives on the Likelihood Ratio Associated with DNA Profile Evidence in Criminal Cases*, 34 J. FORENSIC SCI. SOC'Y 193, 195–96 (1994); Peter Donnelly, *Nonindependence of Matches at Different Loci in DNA Profiles: Quantifying the Effect of Close Relatives on the Match Probability*, 75 HEREDITY 26, 26 (1995); Ian W. Evett, *Evaluating DNA Profiles in a Case Where the Defence Is "It Was My Brother,"* 32 J. FORENSIC SCI. SOC'Y 5, 8 (1992).

11. Marjan Sjerps & Ate D. Kloosterman, *On the Consequences of DNA Profile Mismatches for Close Relatives of an Excluded Suspect*, 112 INT'L J. LEGAL MED. 176, 179 (1999).

12. *See id*.

excluded suspect necessary to determine if significant suspicion is cast on the excluded suspect's relatives.

Weir considers the likelihood ratios for 9-locus profiles in an Australian data set, for both full siblings and first cousins, and reports that these values are "essentially the same" as those for unrelated individuals.[13] However, Weir only considers the proportion of profile pairs in which the likelihood ratio exceeds 1 (for a given number of loci matching at 0, 1, or 2 alleles), not the relatively small probability of unrelated pairs of individuals actually having a high number of shared alleles overall.

I. LIKELIHOOD RATIO

Given an STR profile of a suspect that is similar to (but does not perfectly match) an evidence sample, it would be useful to know which is more likely to generate an exact match: (1) a sibling of the initial suspect or (2) a randomly chosen, unrelated individual. The ratio of these probabilities can be written as follows:

$$LR = \frac{P(E \mid sib)}{P(E \mid random)}, \tag{1}$$

where E refers to the event "initial suspect has type Y, evidence is type X," *sib* refers to the event "evidence was deposited by a sibling" and *random* refers to the event "evidence was deposited by an unrelated individual." This likelihood ratio can be computed for each independent locus and the results multiplied over all loci.

The probability of a randomly chosen, unrelated individual matching the evidence type X (the "random match probability")[14] can be written:

$$P(E \mid random) = P_a P_b HF, \tag{2}$$

where P_a and P_b are the frequencies of the alleles a and b observed in the evidence sample, and *HF* is the heterozygosity factor of these alleles (1 for homozygotes; 2 for heterozygotes). However, when considering the probability of a related individual matching the evidence, we must consider how closely the suspect matches the evidence. For a given locus, the initial suspect may match the evidence at zero, one, or both alleles. A different formula is necessary to calculate $P(E|sib)$ in each of these situations.

Others have explored the likelihood of individuals of varying degrees of relatedness having identical DNA profiles. The following derivation is consistent with the results of these other methods.[15]

---

13. Bruce Weir, *Matching and Partially-Matching DNA Profiles*, 49 J. FORENSIC SCI. 1009, 1009, 1013 (2004).

14. NRC II, *supra* note 4, at 127; Balding, *supra* note 4, at 474; Balding & Donnelly, *supra* note 4, at 605; EVETT & WEIR, *supra* note 4, at 116–23; *Part 1*, *supra* note 5, at 17–18.

15. EVETT & WEIR, *supra* note 4, at 108. Although theirs are general descriptions of 4-allele descent measures, ours is specific to the problem of familial searches.

If the evidence sample and the initial suspect share 0 alleles, then the initial suspect will have two alleles not seen in the evidence, call them c and d. The parents of the initial suspect must have alleles c and d, one to each parent. Without loss of generality, assign allele c to Parent 1 and d to Parent 2. To produce another child who would match the evidence sample, these parents must possess alleles a and b also, one to each parent. This leads to Table 1:

Table 1. Initial Suspect (c, d) Matches Evidence (a, b) at 0 Alleles

Parent 1	Parent 2	Probability of Child (a, b)
(c, a)	(d, b)	¼
(c, b)	(d, a)	¼

From this, the probability of a single additional offspring (for example, a sibling of the initial suspect) of the possible parents matching the evidence sample (a, b) is

$$P(E \mid sib) = \frac{P_a P_b}{4} + \frac{P_b P_a}{4} = \frac{P_a P_b}{2}.$$

(3)

This assumes that a and b represent different alleles. If a and b represent the same allele, then the two cases in Table 1 are identical and thus must be counted only once. *HF* can be used to incorporate this:

$$P(E \mid sib) = \frac{P_a P_b HF}{4}.$$

(4)

Next, consider the case where the initial suspect shares one of two alleles with the evidence sample (a, b). Without loss of generality, the shared allele can be denoted as a and the unshared allele as b. Here the initial suspect will have a and some other allele c. As in the case of 0 shared alleles, the parents of initial suspect (a, c) must themselves possess alleles a and c, one per parent. Table 2 shows how likely such parents are to produce an (a, b) sibling of the initial suspect.

Table 2. Initial Suspect (a, c) Matches Evidence (a, b) at 1 Allele

Parent 1	Parent 2	Probability of Child (a, b)
(a, a)	(c, b)	½
(a, b)	(c, a)	¼
(a, b)	(c, b)	¼
(a, not a or b)	(c, b)	¼

Thus, the probability of an offspring of these potential parents matching the evidence sample (a, b) is

$$P(E \mid sib) = \frac{P_a P_b}{2} + \frac{P_b P_a}{4} + \frac{P_b P_b}{4} + \frac{(1 - P_a - P_b) P_b}{4}.$$

(5)

$$P(E \mid sib) = \frac{2P_a P_b + P_a P_b + P_b^2 + P_b - P_a P_b - P_b^2}{4}. \tag{6}$$

$$P(E \mid sib) = \frac{2P_a P_b + P_b}{4}. \tag{7}$$

As discussed previously, we include *HF* as a correction to allow for the possibility that *a* and *b* are not different alleles:

$$P(E \mid sib) = \frac{P_a P_b HF + P_b}{4}. \tag{8}$$

Finally, consider the case where the initial suspect and the evidence sample match at both alleles *a* and *b*. Here one parent must have allele *a*, while the other has allele *b*, and the remaining alleles must be either *a*, *b*, or some other allele(s), not yet observed:

Table 3. Initial Suspect (*a*, *b*) Matches Evidence (*a*, *b*) at 2 Alleles

Parent 1	Parent 2	Probability of Child (*a*, *b*)
(*a*, *a*)	(*b*, *b*)	1
(*a*, *a*)	(*b*, *a*)	½
(*a*, *a*)	(*b*, not *a* or *b*)	½
(*a*, *b*)	(*b*, *b*)	½
(*a*, *b*)	(*b*, *a*)	½
(*a*, *b*)	(*b*, not *a* or *b*)	¼
(*a*, not *a* or *b*)	(*b*, *b*)	½
(*a*, not *a* or *b*)	(*b*, *a*)	¼
(*a*, not *a* or *b*)	(*b*, not *a* or *b*)	¼

Table 3 yields

$$P(E \mid sib) = P_a P_b + \frac{P_a P_a}{2} + \frac{P_a (1 - P_a - P_b)}{2} + \frac{P_b P_b}{2} + \frac{P_b P_a}{2} + \frac{P_b (1 - P_a - P_b)}{4}$$

$$+ \frac{(1 - P_a - P_b) P_b}{2} + \frac{(1 - P_a - P_b) P_a}{4} + \frac{(1 - P_a - P_b)(1 - P_a - P_b)}{4}. \tag{9}$$

Simplifying this, and converting $2P_a P_b$ to $P_a P_b\ HF$ yields

$$P(E \mid sib) = \frac{1 + P_a + P_b + P_a P_b HF}{4}. \tag{10}$$

This probability is identical to that given by Weir[16] for this same case (where siblings match at both alleles for a given locus).

Combining the three possible degrees of similarity into a single equation yields:

$$P(E \mid sib) = \begin{cases} \dfrac{P_a P_b \, HF}{4}, & \text{if} \quad \text{shared} = 0 \\[3mm] \dfrac{P_b + P_a P_b \, HF}{4}, & \text{if} \quad \text{shared} = 1 \\[3mm] \dfrac{1 + P_a + P_b + P_a P_b \, HF}{4}, & \text{if} \quad \text{shared} = 2 \end{cases} \qquad (11)$$

Note that when one allele is shared between the evidence and initial suspect, the allele b, and thus P_b, must refer to the unshared allele; in any other case, the designation of a and b can be made arbitrarily.

While Equation (11) deals exclusively with siblings, other kinds of relatedness can be evaluated in a similar fashion. For instance, consider the possibility that the initial suspect is known to have a parent (or child) that might have contributed an evidence sample. Following a methodology identical to that presented above, the equation for each locus is:

$$P(E \mid parent \ or \ child) = \begin{cases} 0, & \text{if} \quad \text{shared} = 0 \\[3mm] \dfrac{P_b}{2}, & \text{if} \quad \text{shared} = 1 \\[3mm] \dfrac{P_a + P_b}{2}, & \text{if} \quad \text{shared} = 2 \end{cases} \qquad (12)$$

If no alleles are shared, the probability that the contributor to the evidence is a parent or child of the initial suspect is 0 (since mutations are not considered). If two alleles are shared, the probability is the average of the allele frequencies, which agrees with another formula already derived by Weir.[17] When only one allele is shared (that allele being a, as when we derived Equation 8), the

---

16. BRUCE S. WEIR, GENETIC DATA ANALYSIS II: METHODS FOR DISCRETE POPULATION GENETIC DATA 221 (1996).

17. *Id*.

probability is $P_b/2$. Similar equations can be derived for half-siblings, first cousins, et cetera, and used in a manner similar to the above formulae.

II. EXPERIMENTAL RESULTS

It is necessary to generate threshold values that will allow correct prediction with a stated degree of confidence to provide a useful framework for using these formulae. We provide empirical guidelines for such thresholds through extensive simulation. Such simulation requires a substantial quantity of STR DNA profiles for unrelated individuals and siblings. A data set of individuals from various populations made available by the Federal Bureau of Investigation (FBI)[18] consists of the complete, 13-locus STR genotypes of 959 (presumably unrelated) individuals.[19]

In our experiments involving unrelated individuals, we guaranteed the removal of all associations between alleles (particularly those due to identity by descent) by creating five randomized data sets, each equivalent in both size (959 genotypes each) and allele frequencies to the FBI data set. In this randomization, the alleles observed at each locus in the original FBI data set (without respect to racial classification) were randomly redistributed (without replacement) to produce a new set of genotypes equal in number to the original data set. An example of one possible redistribution of this sort is shown in Table 4. Allele frequencies in this randomized dataset are the same as in the original data set, but individuals are unequivocally unrelated by descent (alleles are not the same because they have been faithfully passed from a common ancestor). Instead, any allele sharing can arise only randomly through identity by state (alleles are the same because there is a finite number of different alleles that can be detected).

Table 4. Example of Alleles Being Redistributed among Three Individuals

	vWA Locus	
Individual	Original	Redistributed
A	18,19	15,17
B	17,18	18,18
C	14,15	14,19

---

18. Bruce Budowle et al., *Population Data on the Thirteen CoDIS Core Short Tandem Repeat Loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians*, 44 J. FORENSIC SCI. 1277 (1999); Bruce Budowle & Tamyra R. Moretti, *Genotype Profiles for Six Population Groups at the 13 CoDIS Short Tandem Repeat Core Loci and Other PCR-Based Loci*, 1 FORENSIC SCI. COMM. (1999), *available at* http://www.fbi.gov/hq/lab/fsc/backissu/july1999/budowle. htm. Data for genotype profiles for PCR-based loci in six population groups can be found at Table 1: Genotype Profile Data for PCR-Based Loci in Six Population Groups, http://www.fbi.gov/hq/lab/fsc/ backissu/july1999/dnaloci.txt (last visited Feb. 16, 2006).

19. David R. Paoletti et al., *Empirical Analysis of the STR Profiles Resulting from Conceptual Mixtures*, 50 J. FORENSIC SCI. 1361, 1361–66 (2005). One pair of individuals (ID #s 2118 and 2163) in the FBI's Bahamian population is reported to share 25 of 26 alleles suggesting either an error in the database, a very unusual pair of unrelated individuals, or the presence of a pair of closely related individuals. *Id*. at 1363.

In experiments involving siblings, profiles were generated by randomly selecting two profiles from the randomized data set of unrelated individuals to use as "parents." At each locus, one allele was chosen randomly from each parent to simulate a locus for a possible child of these two "parents." This process was performed twice across all loci to produce a complete STR genotype for two children (siblings).

The number of all possible pair-wise combinations of the individuals in each data set is $n(n-1)/2$, where n is the number of profiles in the data set. The original FBI database used for this study includes 959 individuals. Five different randomized databases of 959 unrelated individuals were created using the approach outlined above to produce a total of 2,296,805 (5 databases×459,361 pair-wise comparisons per database) sets of parents (and thus sibling pairs) for simulation.

In these trials, either a pair of sibling profiles (as generated above) or a pair of unrelated profiles were selected, with an equal number of trials of each type. One profile was then randomly selected to be the initial suspect and the remaining profile to be the source of the evidence profile. Then, using only the DNA profiles of the "initial suspect" and the "evidence," the question, "Is the actual source of the evidence more likely to be a sibling of the initial suspect than a randomly chosen, unrelated individual?" was addressed.

As previously noted, we are concerned with the hypotheses that "A sibling is the source of the evidence sample" and that "An unrelated individual is the source of the evidence sample." A likelihood ratio greater than 1 lends support to the former hypothesis. Using LR>1 as the decision criterion leads to the results in Table 5.

Table 5. Accuracy of Concluding that a Sibling Is the Source When LR>1

		True state	
		Evidence from unrelated individual	Evidence from sibling
Decision	Evidence from unrelated	~ 98% [Correct decision]	~ 4% [Type II error;
	Evidence from sibling	~ 2% [Type I error; false positive]	~ 96% [Correct decision]

Sjerps and Kloosterman recognize that the determination of what constitutes an appropriate false positive rate is a matter for judicial authorities, not forensic scientists.[20] We provide false-positive rates as a function of the LR threshold. As each of these trials was being run, all 459,361 LR values were saved for analysis

---

20. Sjerps & Kloosterman, *supra* note 11, at 176, 179.

at the end of the run. Using these values, a LR threshold is empirically calculated to result in a desired accuracy, such as 99% or 99.9% for unrelated individuals or siblings. Each of these situations is shown, along with their effect on the experimental classification rate, in Figure 1.
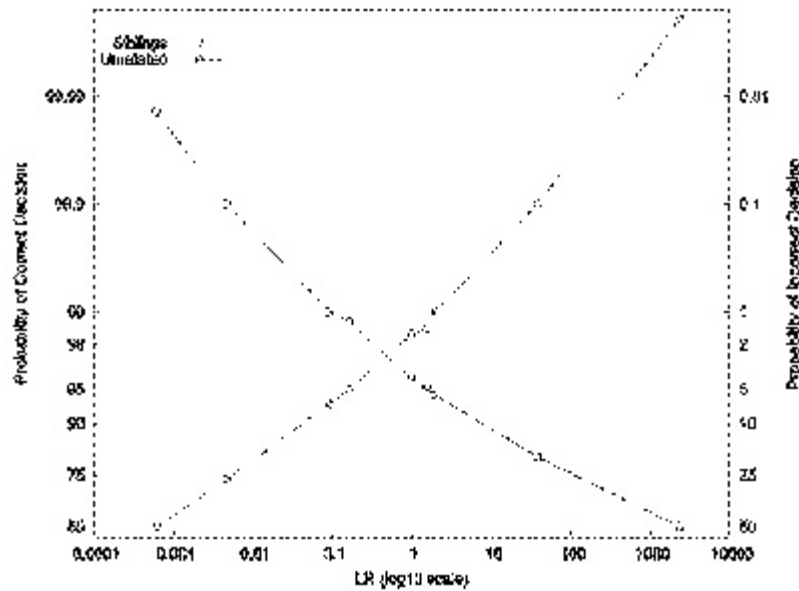

Figure 1. Effect of Using Various LR Thresholds on Accuracy (All values are for the case of a single initial suspect and do not otherwise take the size of a searched database into account. The probabilities on the y-axis are plotted on a logit scale: p% is plotted at the point $log_{10}(p/(100-p))$.).)

The classification rates shown in Figure 1 reflect two types of errors. A false positive (Type I) error occurs when it is determined that DNA profiling should be performed on an initial suspect's sibling even though an unrelated individual is actually the source of the evidence sample (dashed line). Similarly, a false negative (Type II) error occurs when it is determined that DNA profiling should not be performed on an initial suspect's sibling even though a sibling is, in fact, the source of the evidence sample (solid line). Reducing the rate of false positive errors increases the rate of false negative errors, and vice versa.

Another issue of interest is the effect of common versus rare alleles being shared between the evidence and initial suspect. We define "common" alleles to occur at a frequency of 25% in the population of unrelated individuals, "rare"

alleles to have a frequency of 1%,[21] and "average" alleles to occur with frequency 20.3% (the average allele frequency of all the alleles in the FBI data set).

Figure 2 illustrates that an evidence sample and a single initial suspect with DNA profiles that match at as few as five "rare" alleles will result in an LR greater than 1. Similar pairs of profiles, however, must match in at least 15 of 26 "common" alleles to meet the same benchmark or at 13 of 26 alleles for "average" profiles.
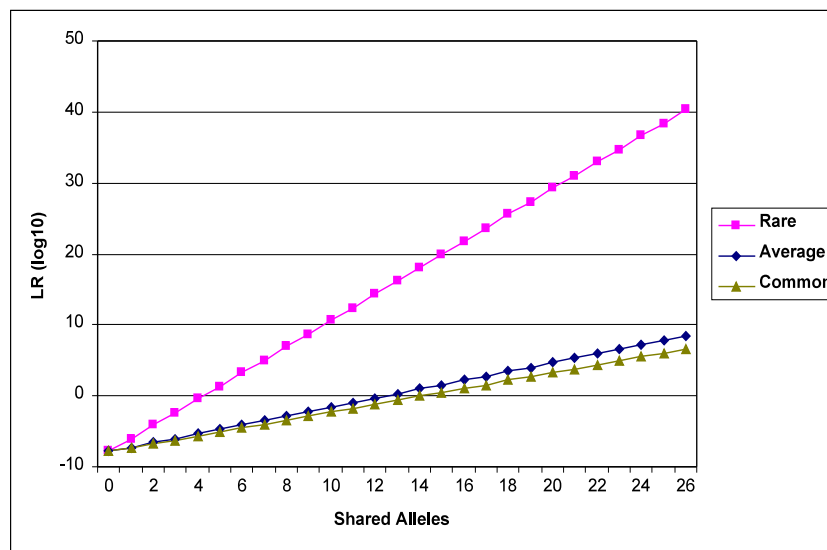

Figure 2. Effect on the LR of All Shared Alleles Being Either Rare or Common. The likelihood ratio (LR) (y-axis) is shown for pairs of synthetic genotypes with the number of shared alleles (x-axis) ranging from 0 to 26. The top line ("Rare") shows the LR when the frequency of all shared alleles is fixed at 1%. The bottom line ("Common") shows the LR when the frequency of all shared alleles is fixed at 25%. The "Average" case sets the allele frequency at 20.3% (the average across all alleles in the FBI data set).

III. BAYESIAN CONSIDERATIONS

The LR method described here compares the likelihoods of two alternative

hypotheses. However, it does not reveal the probability that a sibling of the single initial suspect is the source of the evidence sample. Suppose that in a particular case the population of possible suspects includes one sibling of a single initial suspect with a 10% probability of matching the evidence sample and ten unrelated

---

21. These frequencies are merely illustrating that some specific alleles in the original FBI data set have a higher frequency (for example, the 8 allele at the TPOX locus has a frequency of 0.431) or a lower frequency (many occur only once, corresponding to a frequency of 0.000521) than our assumptions would indicate.

individuals each with a 1% probability of matching. In this case, the sibling is the most likely individual suspect, but the odds are even that the match will be found in one of the ten unrelated individuals. If the alternate suspect pool included 100 unrelated individuals, the match would be more likely to be found in the alternate suspect pool. To determine the probability that the source of the evidence sample is a sibling of the single initial suspect, we can apply Bayes' rule:

$$P(sib \mid E) = \frac{P(E \mid sib) \cdot P(sib)}{P(E \mid sib) \cdot P(sib) + P(E \mid random) \cdot P(random)}, \qquad (13)$$

where the prior probabilities are

$$P(sib) = \frac{s}{popsize}, \qquad (14)$$

$$P(random) = \frac{popsize - s}{popsize}, \qquad (15)$$

where *s* is the number of siblings of the initial suspect. The value of *s* = 1 is used for these illustrations, but formulae are provided so that they can be customized for families with a larger number of siblings, or countries with high fertility rates.[22]

The likelihoods, *P*(*E*|*sib*) and *P*(*E*|*random*), are those used in calculating LR (Equations 2 and 11, respectively). However, to determine the posterior probability that the source of the sample is a sibling of the initial suspect, the prior probabilities *P*(*sib*) and *P*(*random*) must be considered. *P*(*sib*) represents the probability that the source of the evidence is a sibling of the initial suspect without consideration of the evidence itself. Likewise, *P*(*random*) is the probability that the source of the evidence is a random individual, unrelated to the initial suspect, without considering the genotype of the evidence sample. These values are dependent on the size of the pool of potential perpetrators for the case in question.

The computation and use of the prior probabilities can be illustrated by considering a single initial suspect with one sibling, a pool of 100 potential suspects that includes that sibling, and a single-source evidence sample of unknown genotype. Without any knowledge of the genotype of the evidence sample, all 100 suspects are equally likely to be the source of the evidence sample. Thus, *P*(*sib*), the probability that the evidence sample originated from the sibling of the single initial suspect, is 1/100. Similarly, the probability that the source of the sample was one of the 99 members of the suspect pool who are not related to the single initial suspect, *P*(*random*), is simply 99/100. Table 6 gives the posterior probability *P*(*sib* | *E*) when the LR = 100.

---

22. DEPARTMENT OF ECON. & SOC. AFF., UN POPULATION DIVISION, Abortion Policies: A Global Review (2001), *available at* http://www.un.org/esa/population/publications/abortion/doc/ Notes.doc.

Table 6. Effect of Population Size
(Alternate Suspect Pool Size) on Probabilities

The numerator, *P*(*sib*), and denominator, *P*(*random*), of the likelihood ratio, as well as the probability that an evidence sample originated from the sibling of an initial suspect, *P*(*sib* | *E*), are shown for the case where *P*(*E* | *sib*) = 0.01 and *P*(*E* | *random*) = 0.0001.

Population Size	*P*(*sib*)	*P*(*random*)	*P*(*sib* | *E*)
10	10%	90%	91.7%
100	1%	99%	50.3%
1000	0.1%	99.9%	9.1%
10,000	0.01%	99.99%	1.0%
100,000	0.001%	99.999%	0.1%

Table 6 presents posterior probabilities that demonstrate the decrease in correct classification associated with a growing pool of alternate suspects. By definition, the corresponding posterior odds are:

$$PO = \frac{P(sib \mid E)}{P(random \mid E)}. \tag{16}$$

Each row of Table 7 presents the mean results over five randomized datasets simulated in the specified population size with a fixed PO threshold of 1. For example, given an alternate suspect pool the size of a small town (10,000 individuals), Table 4 demonstrates that using a PO threshold of 1 introduces a 62% (100% – ~38%) probability of failing to place a sibling of the single initial suspect under suspicion. To achieve the same error rates, the PO threshold must be decreased as the size of the alternate suspect pool increases. This is demonstrated in Table 8, where the false positive rate has been fixed at approximately 5%.

Table 7. Effect of Population Size (Alternate Suspect Pool Size)
on Classification Rates Using a PO Threshold of 1

Population Size	Actual Perpetrator Is			
	Sibling		Unrelated	
	Sibling is Correctly Investigated	Guilty Sibling Is Not Investigated	No Sibling Is Investigated	Sibling Is Mistakenly Investigated
10^2	75.37%	24.63%	99.96000%	0.042581%
10^4	38.06%	61.94%	99.99960%	0.000392%
10^6	10.86%	89.14%	99.99996%	0.000044%
10^8	1.77%	98.23%	>99.99999%	$<10^{-6}$%
10^{10}	0.17%	99.83%	>99.99999%	$<10^{-6}$%

**Table 8. Effect of Using Various PO Thresholds
to Maintain a False Positive Rate of ≈ 5%**

Population Size	LR Threshold	Actual Perpetrator Is			
		Sibling		Unrelated	
		Sibling is Correctly Investigated	Guilty Sibling Is Not Investigated	No Sibling Is Investigated	Sibling Is Mistakenly Investigated
10^2	1.62E-03	98.57%	1.43%	95.00%	4.99%
10^4	1.60E-05	98.57%	1.43%	95.00%	4.99%
10^6	1.60E-07	98.57%	1.43%	95.00%	4.99%
10^8	1.60E-09	98.57%	1.43%	95.00%	4.99%
10^{10}	1.60E-11	98.57%	1.43%	95.00%	4.99%

A key simplification in this work is that the alternate suspect pool is comprised of individuals completely unrelated to the initial suspect.[23]

IV. DISCUSSION

Perfect matches from government-maintained databases are resulting in a growing number of convictions,[24] but their utility to identify suspects who have *yet to be genotyped* is just beginning to be appreciated.[25] The analysis here provides an objective framework for law enforcement agencies to determine what level of similarity between an evidence sample and a single nonmatching initial suspect warrants investigation of a close relative of an initial suspect. Two important parameters, the size of the reasonable alternative suspect pool and the tolerance for false positives or negatives, are beyond the scope of forensic scientists and are left to be determined on a jurisdictional (and even case-by-case) basis.

As originally suggested by Sjerps and Kloosterman, the observation that a single initial suspect shares rare alleles with an evidence sample casts more suspicion on their close relatives than an observation that they share relatively common alleles.[26] In the case of relatively rare alleles and only a single initial suspect (not a large number of initial suspects such as those maintained in convicted-offender databases) as well as a small alternative suspect pool, sharing as few as five out of a possible 26 alleles might constitute sufficient grounds for the investigation of a sibling. In the case of common alleles, however, as many as 15 alleles may need to be shared. Consequently, it is not possible to arrive at a single metric such as "number of shared alleles" that is independent of allele

---

23. A more comprehensive Bayesian model can be found in other sources. David J. Balding & Peter Donnelly, *Inference in Forensic Identification*, 158 J. ROYAL STAT. SOC. A 21, 21–22 (1995); John Buckleton & Christopher M. Triggs, *Relatedness and DNA: Are We Taking It Seriously Enough?*, 152 FORENSIC SCI, INT'L 115, 116 (2005).

24. Federal Bureau of Investigation, *supra* note 3.

25. Willing, *supra* note 6.

26. Sjerps & Kloosterman, *supra* note 11, at 176–77.

frequencies, the number of initial suspects considered, and the number of potential alternative suspects for the purposes of determining that the investigation of a sibling is warranted. A determination with specific formulae such as those provided here can be performed using only information from the evidence sample and the alleles common to an initial suspect. Alleles possessed by the initial suspect that are not found in the evidence sample (as well as the size of the database searched) play no role in this evaluation beyond excluding them as a possible contributor of the evidence sample.[27]

There has been debate over the appropriate way to describe the significance of a DNA match between a suspect and an evidence sample where the suspect was originally identified because of similarities between the suspect's profile and the evidence sample. The illustrations used here avoid the complications of this debate by effectively presuming a database size of "1" (considering that there was only one initial suspect who was genotyped and found to be similar but not to match an evidence sample). Moreover, the formulae provided here do not consider the complications associated with ambiguities such as those arising from mixtures, mutation, technical artifacts, and degradation-inhibition in the DNA profile obtained from an evidence sample.[28] One simple means of avoiding many of these complications would be to determine likelihood ratios by using only information from those loci where the genotype of a single contributor can be unambiguously determined.

---

27. NRC I, *supra* note 4; NRC II, *supra* note 4; Balding, *supra* note 4; Balding & Donnelly, *supra* note 4; EVETT & WEIR, *supra* note 4.

28. *Part 1*, *supra* note 5, at 21–25; *Part 2*, *supra* note 5, at 25–27.