

U.S. Department of Justice
Federal Bureau of Investigation



**FORENSIC SCIENCE
COMMUNICATIONS**
July 2000 Volume 2 Number 3

Statistical and Population Genetics Issues Affecting the Evaluation of the Frequency of Occurrence of DNA Profiles Calculated From Pertinent Population Database(s)

DNA Advisory Board
February 23, 2000

CURRENT ISSUE

BACK ISSUES

**SEARCH
ALL ISSUES**

**JOURNAL
DESCRIPTION**

EDITORS

**INSTRUCTIONS
FOR AUTHORS**

**HANDBOOK OF
FORENSIC SERVICES**

LINKS

LAB HOME

**FBI
PUBLICATIONS**

Read about . . .

[Introduction](#)

[Source Attribution](#)

[Relatives](#)

[Mixtures](#)

[Database Search](#)

[Conclusion](#)

[References](#)

Introduction

When a comparison of DNA profiles derived from evidence and reference samples fails to exclude an individual(s) as a contributor(s) of the evidence sample, statistical assessment and/or probabilistic reasoning are used to evaluate the significance of the association. Proper statistical inference requires careful formulation of the question to be answered, including, in this instance, the requirements of the legal system. Inference must take into account how and what data were collected, which, in turn, determine how the data are analyzed and interpreted.

Previously, the DNA Advisory Board (DAB; June 21, 1996, New York, New York) endorsed the recommendations of the National Research Council's Report (1996; henceforth NRC II Report):

The DAB congratulates Professor Crow and his NRC [National Research Council] Committee for their superb report on the statistical and population genetics issues surrounding forensic DNA profiling. We wholeheartedly endorse the findings of the report in these substantive matters.

As the NRC II Report (1996) describes, there are alternate methods for assessing the probative value of DNA evidence. Rarely is there only one statistical approach to interpret and explain the evidence. The choice of approach is affected by the philosophy and experience of the user, the legal system, the practicality of the approach, the question(s) posed, available data, and/or assumptions. For forensic applications, it is important that the statistical conclusions be conveyed meaningfully. Simplistic or less rigorous approaches are often sought. Frequently, calculations such as the random match probability and probability of exclusion convey to the trier of fact the probative value of the evidence in a straightforward fashion. Simplified approaches are appropriate, as long as the analysis is conservative or does not provide false inferences. Likelihood ratio (LR) approaches compare mutually exclusive hypotheses and can be quite useful for evaluating the data. However, some LR calculations and interpretations can be complicated, and their significance to the case may not be apparent to the practitioner and the trier of fact.

Bayesian inference, which accounts for information other than the DNA evidence, also could be applied. Bayesian approaches sometimes require knowledge of circumstances beyond the domain of the DNA scientist and have not been addressed in U.S. criminal courts for DNA analysis. The DAB believes it is for the courts to decide whether or not Bayesian statistics are solely the responsibility of the trier of fact. The DAB recognizes that these different approaches can be applied, as long as the question to be answered and the assumptions underlying the analyses are clearly conveyed to the trier of fact.

We have been charged with clarifying issues that arise for the following special cases:

- Source attribution or identity;

- Cases where relatives may be involved;
- Interpretation of mixtures; and
- The significance of a match derived through a felon database search.

[Back to the top](#)

Source Attribution

According to *Webster's Third New International Dictionary* (Merriam-Webster 1961; henceforth Webster's Third), the term unique can convey several meanings, including *the only one, unusual, and some [circumstance] that is the only one of its kind*. Those who question the concept of assigning source attribution for DNA evidence often dwell on the former (e.g., Balding 1999). In their argument against source attribution, some critics say that it is difficult to establish, beyond doubt, that a DNA profile is carried by only one individual in the entire world. Within that context, their argument can be compelling, especially if the profile consists of a fairly small number of loci. Their conclusion, however, is problematic because source attribution should be evaluated within the *context defined by the case*, and the world's population rarely would be the appropriate context. Because source attribution can only be meaningful within the context of the instant case, *Webster's Third* definition of uniqueness comes closest to that required by the legal setting: a circumstance that is the only one of its kind.

By contrast to the world's population, examples of limiting, case-specific contexts are more common. Suppose, for example, the presence of a small group of individuals at the crime scene is stipulated. However, the identity of the single individual who sexually assaulted the victim is at issue. DNA evidence on the victim matches a DNA profile from only one of the named defendants. In this instance, it is simple to assign source because all other individuals are excluded. Now suppose the identities of some individuals at the crime scene are unknown, yet the DNA profile matches one of the defendants. Further suppose this defendant has no close relatives aside from parents. Source attribution is not challenging in this setting. While the answer depends on the number and kind of loci examined, in most instances the source can be assigned with a very high degree of scientific certainty. Suppose, instead, the defendant has multiple siblings, one of whom may have been the assailant and whose profiles are not available for some reason. Even then source can be assigned with a high degree of scientific certainty when a sufficient number of highly polymorphic loci are typed.

Inference regarding source attribution should always be based on the facts in the case. Arguments against source attribution based on premises having nothing to do with the case at hand should not be compelling.

Another set of questions arises when commentators fail to distinguish between source attribution and guilt. Some commentators, for example, set up the following scenario: Suppose inculcating DNA evidence appears to come from the defendant with high probability, yet all the non-DNA evidence is exculpatory (e.g., Balding 1999). In this instance, they say, source attribution is impossible. We do not agree. If, to a high degree of scientific certainty, the DNA evidence appears to come from the defendant, then the only reasonable conclusion is that the DNA did indeed come from the defendant. The trier of fact, however, has a different question to ponder: What value is source attribution if the preponderance of the evidence suggests the defendant cannot be the perpetrator? The trier of fact should seek other explanations for the data, some or all of which may exculpate the defendant.

As described above, the possible source of the DNA depends on the context of the case, and thus calculations for source attribution must reflect the appropriate reference population. If relatives are potential contributors, the calculations for source attribution must reflect that fact. If relatives are not potential contributors, the calculations for source attribution should be based on a defined population; that population could be as small as two unrelated individuals or an entire town, city, state, or country. The DNA analyst should take great care with evidence presentation, with two important facts in mind:

- Inference about source attribution is a probabilistic statement, and its degree of uncertainty is governed by the genetic information contained in the profile; and
- Inference about source attribution is distinct from inference regarding guilt.

One way to develop criteria to assess the question of source attribution is to let p_x

equal the random match probability for a given evidentiary profile X . The random match probability is calculated using the NRC II Report (1996) Formulae 4.1b and 4.4a for general population scenarios or Formula 4.10 under the assumption that the contributor and the accused could only come from one subgroup. The value θ is 0.01, except for estimates for isolated subgroups, where 0.03 is used. The rarity of the estimate is decreased by a factor of 10 (NRC II Report 1996).

Then $(1 - p_x)^N$

is the probability of not observing the particular profile in a population of N unrelated individuals. We require that this probability be greater than or equal to a $1 - \alpha$ confidence level

$$(1 - p_x)^N \geq 1 - \alpha$$

or

$$p_x \leq 1 - (1 - \alpha)^{1/N}.$$

Specifying a confidence level of 0.95 (0.99; i.e., an α of 0.05 or 0.01) will enable determination of the random match probability threshold to assert with 95% (99%) confidence that the particular evidentiary profile is unique, given a population of N unrelated individuals.

In practice, p_x is calculated for each of the major population groups residing in the

geographic area where the crime was committed (i.e., typically African American, Caucasian, and Hispanic). When there is no reason to believe a smaller population is relevant, the FBI, for example, has set N to 260 million, the approximate size of the U. S. population. For smaller, defined populations, N should be based on census values or other appropriate values determined by the facts of the case. The source attribution formula advocated here is simple and likely to be conservative, especially when N is larger than the size of the population that would inhabit a geographic area where a crime is committed.

[Back to the top](#)

Relatives

As described previously in the Source Attribution section, the possibility of a close relative (typically a brother) of the accused being in the pool of potential contributors of crime scene evidence should be considered in case-specific context. It is not appropriate to proffer that a close relative is a potential contributor of the evidence when there are no facts in evidence to suggest this instance is relevant. However, if a relative had access to a crime scene and there is reason to believe he/she could have been a contributor of the evidence, then the best action to take is to obtain a reference sample from the relative. After all, this scenario should be sufficient probable cause for obtaining a reference sample. Typing with the same battery of short tandem repeat (STR) loci will resolve the question of whether or not the relative carries the same DNA profile as the accused.

When a legitimate suspected relative cannot be typed, a probability statement can be provided. Given the accused DNA profile, the conditional probability that the relative

has the same DNA profile can be calculated. Examples of methods for estimating the probability of the same DNA profile in a close relative are described in the NRC II Report (1996) and Li and Sacks (1954).

[Back to the top](#)

Mixtures

Mixtures, which for our purposes are DNA samples derived from two or more contributors, are sometimes encountered in forensic biological evidence. The presence of a mixture is evident typically by the presence of three or more peaks, bands, dots, and/or notable differences in intensities of the alleles for at least one locus in the profile. In some situations, elucidation of a contributor profile is straightforward. An example would be the analysis of DNA from an intimate swab revealing a mixture consistent with the composition of the perpetrator and the victim. When intensity differences are sufficient to identify the major contributor in the mixed profile, it can be treated statistically as a single source sample. At times, when alleles are not masked, a minor contributor to the mixed profile may be elucidated. Almost always in a mixture interpretation, certain possible genotypes can be excluded. It may be difficult to be confident regarding the number of contributors in some complex mixtures of more than two individuals; however, the number of contributors often can be inferred by reviewing the data at all loci in a profile.

Interpretation of genotypes is complicated when the contributions of the donors is approximately equal (i.e., when a major contributor cannot be determined unequivocally) or when alleles overlap. Also, stochastic fluctuation during polymerase chain reaction (PCR) arising from low quantity of DNA template can make typing of a minor contributor complicated. When the contributors of a DNA mixture profile cannot be distinguished, two calculations convey the probative value of the evidence.

The first calculation is the probability of exclusion (PE; Devlin 1992 and references therein). The PE provides an estimate of the portion of the population that has a genotype composed of at least one allele not observed in the mixed profile. Knowledge of the accused and/or victim profiles is not used (or needed) in the calculation. The calculation is particularly useful in complex mixtures, because it requires no assumptions about the identity or number of contributors to a mixture. The probabilities derived are valid and for all practical purposes are conservative. However, the PE does not make use of all of the available genetic data.

The LR provides the odds ratio of two competing hypotheses, given the evidence (Evet and Weir 1998). For example, consider a case of sexual assault for which the victim reported there were two assailants. A mixture of two profiles is observed in the "male fraction," and the victim is excluded as a contributor of the observed mixed profile. Two men are arrested, and their combined profiles are consistent with the

mixture evidence. A likelihood calculation logically might compare the probability that the two accused individuals are the source of the DNA in the evidence versus two unknown (random men) are the source of the evidence. Various alternate hypotheses can be entertained as deemed appropriate, given the evidence. Calculation of a LR considers the identity and actual number of contributors to the observed DNA mixture. Certainly, LR makes better use of the available genetic data than does the PE.

Interpretation of DNA mixtures requires careful consideration of factors including, but not limited to, detectable alleles; variation of band, peak, or dot intensity; and the number of alleles. There are a number of references for guidance on calculating the PE or LR (Evetts and Weir 1998; NRC II Report 1996; PopStats in CODIS). The DAB finds either one or both PE or LR calculations acceptable and strongly recommends that one or both calculations be carried out whenever feasible and a mixture is indicated.

[*Back to the top*](#)

Database Search

As felon DNA databases develop in all 50 states, searches for matches between evidentiary and database profiles will become increasingly common. Two questions arise when a match is derived from a database search: (1) What is the rarity of the DNA profile? and (2) What is the probability of finding such a DNA profile in the database searched? These two questions address different issues. That the different questions produce different answers should be obvious. The former question addresses the random match probability, which is often of particular interest to the fact finder. Here we address the latter question, which is especially important when a profile found in a database search matches the DNA profile of an evidence sample.

When the DNA profile from a crime scene sample matches a single profile in a felon DNA database, the NRC II Report (1996) recommended the evaluation of question number 2 be based on the size of the database. They argued for this evaluation because the probability of identifying a DNA profile by chance increases with the size of the database. Thus this chance event must be taken into account when evaluating value of the matching profile found by a database search. Those who argue against NRC II's recommended treatment (e.g., Balding and Donnelly 1996; Evetts and Weir 1998; Evetts et al. in press) say the NRC II Report's formulation is wrong and undervalues the evidence. In fact, they argue that the weight of the evidence (defined in terms of a likelihood ratio) for a DNA database search exceeds the weight provided by the same evidence in a "probable cause" case — a case in which other evidence first implicates the suspect and then DNA evidence is developed.

When other evidence first implicates the suspect, the DNA evidence can be evaluated

using the probability p_x of randomly drawing the profile X from the (appropriate)

population, which expresses the degree of surprise that the suspect and evidentiary profiles match. Equivalently, we can express it as a LR for two competing hypotheses, namely the likelihood of the evidence when the data come from the same individual (H_s) versus the likelihood of the evidence when the data come from two

different individuals (H_d). The LR in this instance is

$$\text{Lik}(\text{Profile} | H_s) / \text{Lik}(\text{Profile} | H_d) = p_x / (p_x * p_x) = 1/p_x.$$

For the DNA database search, the NRC II Report recommended the calculation (defined in terms of a LR) to be evaluated as $1/(N p_x)$, where N is the size of the

database. While justification for this calculation is given in their report, it is often misunderstood. Stockmarr (1999) rederives this result in a way that should be more comprehensible. As a special case, assume only one profile in the database matches the evidentiary profile; we can consider that individual is a suspect. Now consider two competing hypotheses, namely the source is or is not in the database (H_{in} versus

$H_{not\ in}$). These likelihoods are relevant because we wish to identify whether the

suspect is likely to be the source of the sample (H_{in}) or if it is more likely he was

identified merely by chance ($H_{not\ in}$). What is the LR for these hypotheses?

$$\text{Lik}(\text{Profile} | H_{in}) / \text{Lik}(\text{Profile} | H_{not\ in}) = 1 / (N p_x).$$

Stockmarr (1999) argues this formulation is the appropriate treatment of the data, as did the NRC II Report (1996) before him. Both recognize an intuitive counter example. Suppose we had a DNA database of the entire world's population (size N), except one individual ($N - 1$). A DNA profile from a crime scene is found to match one and only one profile in the database, and its frequency is $1/N$. According to critics (e.g., Balding 1997), this example demonstrates the fallacious nature of the NRC II Report's proposed evaluation of the evidence for a database search, because the value of the evidence appears to be nil (the likelihood ratio is essentially one instead of a large number). Both Stockmarr (1999) and the NRC II Report recognize this interesting result; however, by treating the problem from a Bayesian perspective and

invoking prior probabilities that are a function of the size of the database, they argue the example is irrelevant. In essence, the prior probability of H rises as N rises. This *in*

approach is coherent, from the statistical perspective, but it may not be particularly helpful for the legal system. Without the use of prior probabilities, it should be apparent that the treatment of the database search recommended by the NRC II Report can be conservative when the database is extremely large.

It is important to consider the treatment proposed by Balding and Donnelly (1996) and recently endorsed by Evett, Foreman, and Weir (in press). By their line of reasoning, the LR is no different whether other evidence first implicates the suspect or the suspect is identified by a database search. In fact, they argue the true weight of the evidence is actually larger for the latter, albeit the increase is small unless N is large. This argument has some intuitive appeal, especially in light of the example given above, and it is true that their LR is unaffected by sampling.

Both camps appear to present rigorous arguments to support their positions. Indeed the proper treatment superficially appears to rest in the details of arcane mathematics (Balding and Donnelly 1996; NRC II Report 1996; Stockmarr 1999). We believe, however, there is a way to see which of the two treatments is better for the legal setting without resorting to mathematical details. Consider the following scenario:

A murder occurs, and the only evidence left at the crime scene is a cigarette butt. DNA analysis types five loci from the saliva on the cigarette butt. The probability of drawing the resulting profile X from a randomly selected individual is $p_x = 1/100,000$. A search of the

DNA database, which contains $N = 100,000$ profiles, reveals a single match. No other evidence can be found to link the "suspect," whose profile matches, to the murder.

If we follow Balding and Donnelly (1996), the message for the investigators is that the evidence is 100,000 times more likely if the suspect is the source than if he is not. Alternatively, by the NRC II Report (1996) recommendations, the evidence is not compelling because the likelihood the profile, a priori, is/is not in the database is the same. In probabilistic terms, it is not surprising to find a matching profile in the database of size 100,000 when the profile probability is 1/100,000. Curiously, the mathematics underlying both approaches are correct, despite the apparently divergent answers. It is the foundations of the formulations that differ, and they differ substantially.

At present there are about 20,000 known, variable STR loci in the human genome. Of these, forensic scientists use a little more than a dozen, which is sufficient for most forensic analyses. Although not strictly accurate, let us think of the selection of STR loci as random and return to our case. The forensic scientists who worked on the

cigarette butt could assay only five loci of the dozen they might type. Suppose they were to type five different loci and generate a new profile based on only these additional five loci? If our suspect were the true source of the sample, a match at those loci would be obtained; however, if he were not the source, a match would be highly unlikely. If the new (i.e., second) profile probability were again on the order of 1/100,000, someone else may have been selected. If our suspect is not the source, no one else in the database is, and yet we can easily imagine selecting a set of five loci (out of the thousands possible) to single out each individual therein. This seems like an unsatisfactory state in light of the LR espoused by Balding and Donnelly (1996).

Thus we are left with an interesting dilemma. Within a Bayesian context, the NRC II Report's LR and Balding and Donnelly's (1996) LR could be interpreted to yield a coherent evaluation of the evidence. Unfortunately, Bayesian logic has not been considered by the U.S. criminal legal system for DNA analysis. Clearly, what is required is a formulation of the LR that transparently conveys its import without resorting to Bayesian statistics. In this setting, the treatment of the database search recommended by the NRC II Report can be conservative, but only for the unlikely scenario of a very large N is it very conservative. Apparently the treatment of the database search recommended by Balding and Donnelly (1996) is not conservative when the number of loci genotyped is small and remains so until the number of loci becomes large enough to essentially ensure *uniqueness*. To put it another way, without the Bayesian framework, the Balding and Donnelly (1996) formulation is easily misinterpreted in a fashion unfavorable to the suspect. Stockmarr's (1999) formulation, which is a more formal exposition of what originally appeared in the NRC II Report (1996), communicates value of a database search far better, and it is always conservative. Thus, we continue to endorse the recommendation of the NRC II Report for the evaluation of DNA evidence from a database search.

[Back to the top](#)

Conclusion

Statistical analyses are sometimes thought to yield automatic rules for making a decision either to accept or reject a hypothesis. This attitude is false in any setting and should be especially avoided for forensic inference. One rarely rests his/her decisions wholly on any single statistical test or analysis. To the evidence of the test should be added data accumulated from the scientist's own past work and that of others (Snedecor and Cochran 1967). Thus, in this light, statistical analyses should be thought of as useful guides for interpreting the weight of the DNA evidence.

References

Balding, D. J. Errors and misunderstandings in the second NRC report, *Jurimetrics*

(1997) 37:603–607.

Balding, D. J. When can a DNA profile be regarded as unique?, *Science & Justice* (1999) 39:257–260.

Balding, D. J. and Donnelly, P. Evaluating DNA profile evidence when the suspect is identified through a database search, *Journal of Forensic Sciences* (1996) 41:603–607.

Devlin, B. Forensic inference from genetic markers, *Statistical Methods in Medical Research* (1992) 2:241–262.

Evett, I. W., Foreman, L. A. and Weir, B. S. *Biometrics*, in press.

Evett, I. W. and Weir, B. S. *Interpreting DNA Evidence*. Sinaue, Sunderland, Massachusetts, 1998.

Li, C. C. and Sacks, L. The derivation of joint distribution and correlation between relatives by the use of stochastic matrices, *Biometrics* (1954) 10:347–360.

Merriam-Webster, Incorporated. *Webster's Third New International Dictionary*. Merriam-Webster, Incorporated, Springfield, Massachusetts, 1961.

National Research Council Committee on DNA Forensic Science. *An Update: The Evaluation of Forensic DNA Evidence*. National Academy Press, Washington, D.C., 1996.

Snedecor, G. W. and Cochran, W. G. *Statistical Methods* (6th ed.). Iowa State University Press, Ames, 1967, p. 28.

Stockmarr, A. Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search, *Biometrics* (1999) 55:671–677.

[Back to the top](#)

FORENSIC SCIENCE COMMUNICATIONS JULY 2000 VOLUME 2 NUMBER 3

[CURRENT ISSUE](#)

[BACK ISSUES](#)

[SEARCH
ALL ISSUES](#)

[JOURNAL
DESCRIPTION](#)

[EDITORS](#)

[INSTRUCTIONS
FOR AUTHORS](#)

[LINKS](#)

[HANDBOOK OF
FORENSIC SERVICES](#)

[LAB HOME](#)

[FBI
PUBLICATIONS](#)