Ian W. Evett,[1] *D.Sc.; Lindsey A. Foreman,*[1] *Ph.D.; James A. Lambert,*[2] *M.Sc.; and Adrian Emes,*[1] *B.Sc.*

Using a Tree Diagram to Interpret a Mixed DNA Profile

REFERENCE: Evett IW, Foreman LA, Lambert JA, Emes AE. Using a tree diagram to interpret a mixed DNA profile. J Forensic Sci 1998;43(3):472–476.

ABSTRACT: A recent case is described where the evidence of bloodstaining on a knife suggested that it was a mixture from the two victims. Interpretation of the evidence in this problem necessitated the formulation of several sets of multiple hypotheses which were analyzed by means of a tree diagram. The problem was then greatly simplified to one of comparing the two alternative hypotheses of most interest. It was found that results were robust to variation in the expert's judgment regarding the possibility that a mixture of blood was present on the knife.

KEYWORDS: forensic science, DNA typing statistical interpretation, mixed DNA profile, likelihood ratio, tree diagram, multiple hypotheses.

A recent paper by Evett et al. (1) presented a quantitative method for taking account of peak areas when interpreting mixed DNA profiles. In this paper, we describe an interesting case in which a quantitative analysis of the peak areas did not appear to be necessary. The problem in this case was more one of determining which were the meaningful hypotheses to address. The solution was certainly not obvious at first sight, but we found that a relatively simple analysis based on a tree diagram clarified the issues remarkably.

Materials and Methods

Short tandem repeat (STR) profiling was carried out using multiplex polymerase chain reaction (PCR) according to the method described in [2, 3]. The multiplex detected the following STR loci: D8S1179 (4); D18S51 (5); HUMVWFA31/A (6); HUMTHO1 (7); HUMFIBRA (FGA) (8); D21S11 (9); and the Amelogenin sex test described in (10).

The DNA was analyzed on an ABD 377 GeneScanner and fragment sizes determined automatically by GeneScan software.

Case Summary and Profiling Results

For the purposes of this discussion, we simplify the circumstances of the actual case considerably to the following. Two sisters, Lisa and Pauline, had both been stabbed in the course of a violent attack on them by a single male (who was no blood relation). The exhibit of interest was a knife, found discarded near the house where the attack took place, and this bore bloodstaining. There were two main areas of staining which we refer to as stains 1 and 2. Blood samples were obtained for DNA analysis from both sisters and the profiling results are summarized in Table 1.

[1]The Forensic Science Service, Metropolitan Laboratory, 109 Lambeth Rd, London SEI 7LP, UK.

[2]The Forensic Science Service, Priory House, Gooch St North, Birmingham, B5 6QQ, UK.

Received 11 June 1997; accepted 17 Sept. 1997.

Formulation of Hypotheses

It would be tempting to infer that the best supported hypothesis is that stain 1 came from Lisa, who is homozygous 13 at D8, and that stain 2 came from Pauline, who has the required (8,13) genotype at that locus. However, the intensities of the peaks at D8 for stain 2 can be seen from Table 2 and Fig. 1 to be inconsistent with that view; that is, the peak area for allele 8 at D8 is about 10% of the area for allele 13. Inspection of the peak intensity information at the remaining loci in Fig. 1 suggests that the best supported combination of genotypes at D8 is (8,13) and (13,13). Therefore, the full range of alternative explanations for the origin of stain 2 appears to be:

Pauline alone
Lisa and Pauline
Pauline and an unknown person
Lisa and an unknown person
One unknown person
Two unknown persons

The question now arises of how these alternatives are to be weighed against each other. The peak area information appears to support hypotheses based on two contributors exhibiting genotypes (8,13) and (13,13) at D8. Other combinations, such as a mixture of (8,8) and (13,13), are possible but have little support and it was decided to ignore them for the purposes of this analysis. Furthermore, the interpretation of stain 2 should not be undertaken independently of that of stain 1. The composite analysis that we carried out is described below.

The analysis of any case is simplest if it is reasonable to consider only two explanations, or hypotheses, for the evidence. The weight of evidence is then encapsulated in the ratio of the probabilities of the evidence given each of the two explanations—*the likelihood ratio.* In this case, however, that does not seem possible and a more detailed analysis is necessary. Any hypothesis that we address must be a composite of three parts:

Origin of stain 1—For stain 1 there are two alternatives: either Lisa was the contributor or some unknown person.

Is stain 2 a mixture?—For stain 2, there are two alternative explanations for the intensity information: either the stain is a mixture of genotypes (8,13) and (13,13) at D8 (as discussed above) or it is not.

Origins of stain 2—For stain 2, we have seen that there are various alternative explanations for the observed genotype and these depend to some extent on the preceding hypotheses. For example, if the first sub-hypothesis is that stain 1 was from Lisa and the second sub-hypothesis is that stain 2 is a mixture, then

there are four possible alternative sub-hypotheses for the origins of stain 2: Lisa and Pauline; Lisa and an unknown person; Pauline and an unknown person, or two unknown persons.

Note that we restrict consideration to the case of two-person mixtures. Clearly it is possible in principle that three, four or more people contributed to the DNA on the knife, but these are increasingly poorly supported hypotheses which serve only to complicate the analysis. The entire scheme of combinations of sub-hypotheses, denoted by the H_i's, can best be shown using a tree diagram, as in Fig. 2. Note that all of the implied composite hypotheses have been numbered consecutively in the right-hand column. The U terms have been used to denote unknown people in the following way. U1 is a person with the same genotype as stain 1, that is, including homozygous genotype (13,13) at the D8 locus. U2 is a person with the same genotype as stain 2, that is, including genotype (8,13) at D8. It is also necessary to introduce a third unknown person, U3, for hypotheses 10 and 12 to allow for the possibility that stain 1 came from one person and stain 2 came from either *another* person and Pauline or from two *further* people; thus, U3 must have the same genotype as stain 1.

TABLE 1—*Summary of profiling results.*

	Observed Genotypes					
Sample	D18	D21	THO1	D8	FGA	VWA
Lisa	12, 15	59, 61	9, 9.3	13, 13	22, 24	15, 17
Pauline	12, 15	59, 61	9, 9.3	8, 13	22, 24	15, 17
Stain 1	12, 15	. . .	9, 9.3	13, 13	22, 24	15, 17
Stain 2	12, 15	. . .	9, 9.3	8, 13	22, 24	15, 17

TABLE 2—*Peak areas for stain 2 on knife at locus D8.*

Allele	Peak Area
8	313
13	2955

Logical Inference

The posterior probability of each hypothesis H_i, where $i = 1,2, \ldots, 14$, is given by Bayes' theorem

$$Pr(H_i|E) = \frac{Pr(E|G, H_i)Pr(H_i)}{\sum_i Pr(E|G, H_i)Pr(H_i)}$$

where E denotes the bloodstain evidence from the knife, $G = (G_L, G_P)$ denotes the profile evidence from Lisa and Pauline, and $Pr(H_i)$ denotes the prior probability of H_i given the non-DNA evidence. At first sight, this formulation presents a problem because it does not appear possible to separate the prior terms, which are the province of the court, from the likelihood terms $Pr(E|G,H_i)$, which are the province of the scientist. However, there is a solution as will be explained later. For the time being, the scientist is concerned with assigning values to the $Pr(E|G,H_i)$.

We have seen that E is in three parts which we define as follows: E_1, the genotype of stain 1; E_2 the peak area information for stain 2 at locus D8; and E_3, the genotype of stain 2. It is helpful to decompose each hypothesis into three parts, according to what it states in relation to each of these three facets of the evidence; so $H_i = \{H_{i1}, H_{i2}, H_{i3}\}$. As an illustration of this terminology, consider the composite hypothesis H_7 in Fig. 2:

H_{71}: stain 1 came from an unknown person
H_{72}: stain 2 is a mixture
H_{73}: stain 2 came from Lisa and Pauline

Then, for any of the likelihood terms, we have

$$Pr(E|G, H_i) = Pr(E_1, E_2, E_3|G, H_{i1}, H_{i2}, H_{i3})$$

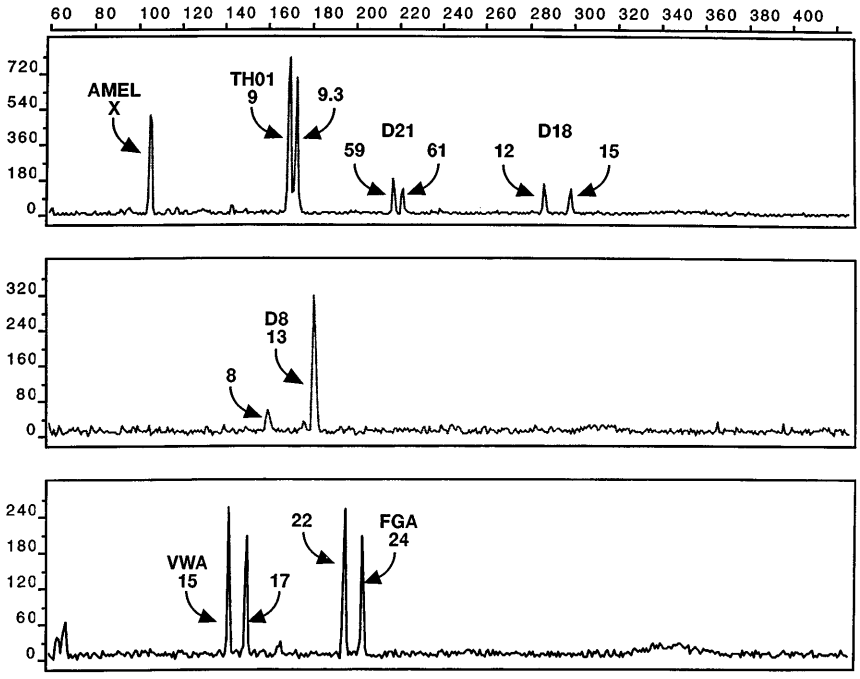

FIG. 1—*STR allele designations for bloodstain 2 on the knife.*

This can be decomposed using the multiplication rule for probability to give

$$Pr(E|G, H_i) = Pr(E_1|G, H_{i1}, H_{i2}, H_{i3})Pr(E_2|E_1, G, H_{i1}, H_{i2}, H_{i3}) \times Pr(E_3|E_1, E_2, G, H_{i1}, H_{i2}, H_{i3})$$

We may simplify by asserting the following:

$$Pr(E_1|G, H_{i1}, H_{i2}, H_{i3}) = Pr(E_1|G, H_{i1})$$

that is, the genotype of stain 1 depends only on the identity of the person from whom it came;

$$Pr(E_2|E_1, G, H_{i1}, H_{i2}, H_{i3}) = Pr(E_2|H_{i2})$$

that is, the peak area data depend only on whether or not stain 2 is a mixture;

$$Pr(E_3|E_1, E_2, G, H_{i1}, H_{i2}, H_{i3}) = Pr(E_3|G, H_{i2}, H_{i3})$$

that is, the genotype of stain 2 depends only on whether or not it is a mixture and the identity of the person, or persons, who contributed to it. Thus, it follows that

$$Pr(E|G, H_i) = Pr(E_1|G, H_{i1})Pr(E_2|H_{i2})Pr(E_3|G, H_{i2}, H_{i3})$$

It is now necessary to assign a value to each of the three components of the evidence given the relevant sub-hypotheses and this is demonstrated in Fig. 3.

First, using Caucasian allele proportions from Evett et al. (11) and adopting the approach of Balding and Nichols (12), we find that the match probabilities for the genotypes of stains 1 and 2 are 3×10^{-6} and 6×10^{-7} respectively, given *G*. The difference between the two reflects the fact that, among Caucasians, allele 8 at the D8 locus is rarer than allele 13. We also assume that, for any hypothesis which involves two unknowns, the two are unrelated both to each other and to Lisa and Pauline, that is, beyond being members of the same subpopulation, which is the assumption for the match probability calculation. It could be argued that if an unknown person has contributed to the bloodstaining on the knife, the fact that the two profiles of interest differ by only a single allele suggests that they are quite likely to have originated from two close blood relations. If a plausible alternative is that a close blood relative of Lisa and Pauline's was involved in the stabbing, who cannot be excluded, then this situation may be dealt with (12) and the strength of the evidence will be reduced. Next, we assume that genotyping is done without error so, if stain 1 is from Lisa, it is certain that it will have the same genotype as her and so on.

The most difficult step comes in considering the mixture hypotheses. Here we must emphasize a point we made in an earlier paper (1). It is very tempting for the scientist to address questions of the kind ''What is the probability that it is a mixture given the peak areas?'' However, we recall a general principle of forensic science: The scientist must consider the probability of the *evidence* given the hypothesis, not the other way around. So, in the present context, it is necessary to address the questions ''What is the probability of the peak area information given that it is a mixture?'' and ''What is the probability of the peak area information given that

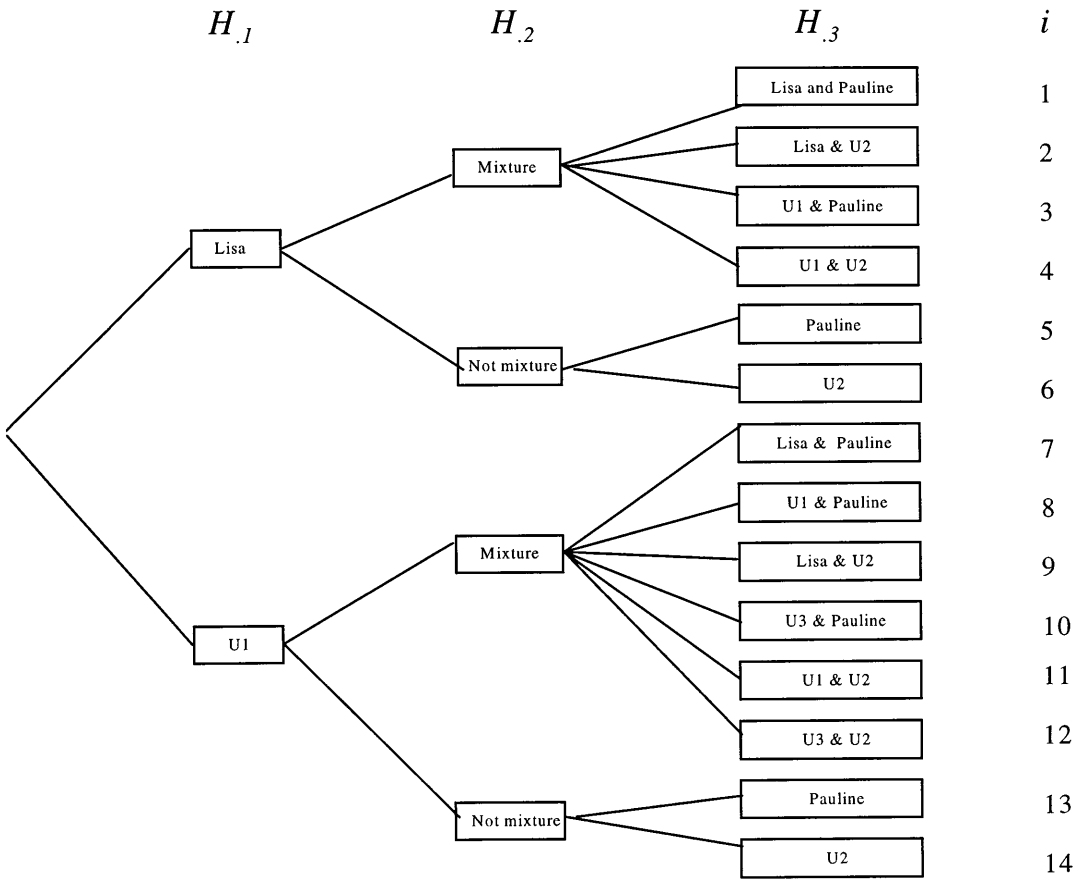

FIG. 2—*Tree diagram describing all possible hypotheses for the bloodstain evidence on the knife (E).*

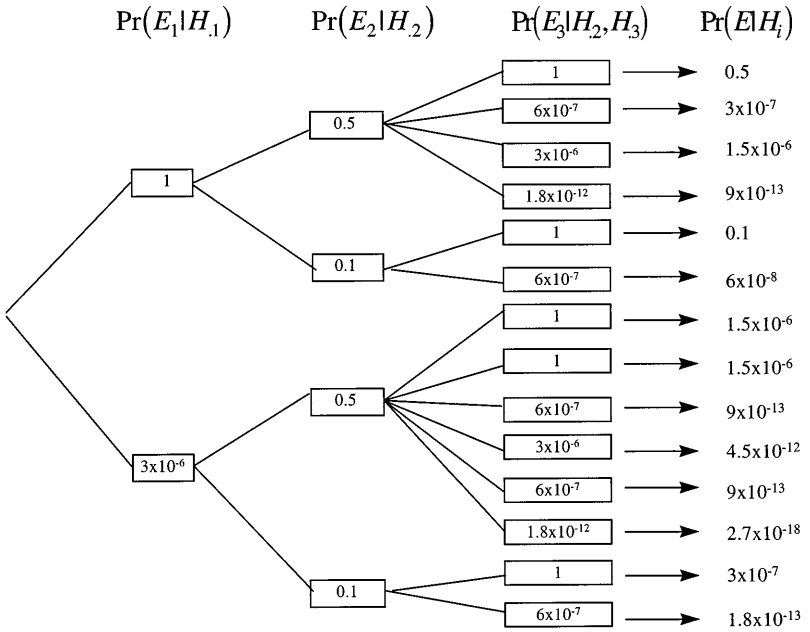

FIG. 3—*Tree diagram describing the probability of the evidence for each of the composite hypotheses in Fig. 2.*

it is not a mixture?'' We accept that these might be difficult questions even for an expert to answer. Later, however, we demonstrate that the interpretation is dependent only on the *ratio* of these probabilities and we believe that it *is* reasonable to ask an expert a question of the form ''How much more probable is the observed pattern of peak areas if the DNA is a mixture than if it is not?'' Purely for illustration, we proceed on the basis that the expert judges the answer to this question is of the order 5. The crucial feature here is that we later investigate the sensitivity of the final answer to this ratio and we show it is almost completely insensitive to the expert's assigned value. For illustration, we take the probability of the peak area information given that the DNA is a mixture to be 0.5 and given that it is not a mixture to be 0.1. We now have the means to complete the tree diagram in Fig. 3 which shows the probability of the evidence given the corresponding sub-hypotheses in Fig. 2.

The probability of the combined evidence given each composite hypothesis is then calculated by multiplying together the three terms along the respective branches leading to it. The results are shown in the last column. Consider, for example H_7. Referring to the sub-hypotheses in Fig. 2, $Pr(E_1|G,H_{71})$ is the probability of the observed genotype of stain 1 given that it came from some unknown person, which is 3×10^{-6}. Next, $Pr(E_2|H_{72})$ is the probability of the peak area data for stain 2 given that it is a mixture, for which the value 0.5 has been assumed. Finally, $Pr(E_3|G,H_{72}, H_{73})$ is the probability of the observed genotype of stain 2 given that it is a mixture of the DNA of Lisa and Pauline, which is 1. The probability of the combined evidence given hypothesis 7 is then the product of these three terms, that is, 1.5×10^{-6}.

Numerical Analysis

We note that hypotheses 1 and 5 are best supported. However, presenting a picture such as Fig. 3 as evidence to a court is unlikely to be illuminating. A radical simplification is possible if the number of hypotheses can be reduced to two and this can be done in the present case by formulating composite hypotheses as follows:

The knife bears DNA from both Lisa and Pauline
The knife bears DNA which is *not* from both Lisa and Pauline

These two hypotheses are mutually exclusive and exhaustive. The probability of the evidence given the first hypothesis is the sum of the values in Fig. 3 corresponding to hypotheses 1, 3, 5, and 7, provided the prior probabilities for each of these are taken to be the same. The probability of the evidence given the second hypothesis is the sum of the remaining terms in the final column of Fig. 3, again provided that the corresponding prior probabilities are the same. The problem has, at last, been reduced to one where a likelihood ratio can be calculated.

Recall, however, that we were tentative about assigning probabilities to the peak area evidence for stain 2. It is important to determine how sensitive the final likelihood ratio is to the ratio of these probabilities and this was done, in the first instance, by means of a spreadsheet calculation. The appearance of the spreadsheet was very similar to Fig. 3. The values in the column for $Pr(E_2|H_2)$ were calculated from two cells on the spreadsheet—one where a value for $Pr(E_2|$ mixture) could be entered and the other where the ratio $Pr(E_2|$ mixture)/$Pr(E_2|$ not mixture) was entered. In this way, all values for both terms could be explored, although we believe that the ratio would be the most meaningful quantity for an expert to estimate. The result of this analysis, which was initially of some surprise to us, was that the likelihood ratio of 277,778 was insensitive to six significant figures to all values of $Pr(E_2|$ mixture)/$Pr(E_2|$ not mixture), whatever the value of $Pr(E_2|$mixture). This is extremely comforting, particularly as it is the custom of the Forensic Science Service to report such likelihood ratios to two significant figures at most.

We then carried out an algebraic analysis to investigate the reason for the insensitivity.

Algebraic Analysis

Let p_L,p_P denote the match probabilities for Lisa's and Pauline's genotypes, respectively. We use $Pr(E_2|m)$ and $Pr(E_2|\overline{m})$ as shorthand for $Pr(E_2|$ mixture) and $Pr(E_2|$ not mixture), respectively, and

let $Q = Pr(E_2|m)/Pr(E_2|\overline{m})$. Then, from the structure of Figs. 2 and 3, we derive the following expression for the likelihood ratio, *LR*

$$\frac{Pr(E_2|m) + Pr(E_2|\overline{m}) + 2p_L Pr(E_2|m)}{Pr(E_2|m)(p_P + p_L + 3p_L p_P + p_L^2 p_P + p_L^2) + Pr(E_2|m)(p_P + p_L + p_L p_P)}$$

Dividing numerator and denominator by $Pr(E_2|\overline{m})$ gives us that the *LR* is

$$\frac{Q(1 + 2p_L) + 1}{Q(p_P + p_L + 3p_L p_P + p_L^2 p_P + p_L^2) + (p_P + p_L + p_L p_P)}$$

We now observe that p_L and p_P are very small compared to 1, so the numerator is, to a good approximation, $Q + 1$. We also ignore higher-order terms in the denominator to obtain

$$\text{LR} \approx \frac{Q + 1}{Q(p_P + p_L) + (p_P + p_L)} = \frac{1}{(p_P + p_L)}$$

So, to a good approximation (the spreadsheet analysis shows it is good to six significant figures), the LR comparing the two composite hypotheses is independent of the expert's opinion of the peak area evidence.

Discussion

When we first studied the data in this case we found the analysis for stain 2 puzzling. Pauline's DNA appeared to be present but how well was this hypothesis supported? After all, it was based only on a comparatively small peak in D8 and the answer seemed to depend critically on whether or not it was a mixture. Furthermore, if it were a mixture of Lisa and an unknown person, then we would expect to see three and four peak profiles at some of the other loci.

The bloodstain evidence was composed of three main parts: the profile type of stain 1, the peak intensity information for stain 2, and the profile type of stain 2. By formulating a set of composite hypotheses covering all possible explanations for this evidence, the problem was greatly clarified using a tree diagram. From this, we were able to compare the hypothesis of interest, that the knife exhibited bloodstaining from both Lisa and Pauline, against its alternative. The resulting LR value of approximately 280,000 indicated very strong evidence in support of this hypothesis. The LR utilized an expert's assessment of the relative likelihood of the peak intensity information for stain 2 if it did or did not originate from two sources. It emerged that the result was completely insensitive to this assessment. We investigated the analysis algebraically to establish the reason for this and found that the LR comparing the hypothesis that the knife bore blood from both Lisa and Pauline versus its complement was approximated simply by the inverse of the sum of the match probabilities of Lisa and Pauline's profiles. Thus, even though the structure of our case seems more complicated than usual, the analysis yields an analogous result to that obtained when testing for the presence of a *single* named contributor's DNA in a bloodstain originating from a single source; that is, the LR in this case reduces to the inverse of a match probability.

References

1. Evett IW, Gill PD, Lambert JA. Taking account of peak areas when interpreting mixed DNA profiles. J Forensic Sci 1998;43:62–9.
2. Sparkes R, Kimpton C, Watson S, Oldroyd N, Clayton T, Barnett L, et al. The validation of a 7-locus multiplex STR test for use in forensic casework. Int J Legal Med 1996;109:186–194.
3. Oldroyd NJ, Urquhart AJ, Kimpton CP, Millican ES, Watson SK, Frazier RRE, et al. Development and optimization of a highly discriminating multiplex PCR system suitable for forensic identification. In: Carracedo A, Brinkmann B, Bar W, Editors. Advances in Forensic Haemogenetics 6. Berlin Heidelberg, New York: Springer, 1996.
4. Co-operative Human Linkage Centre. Database. Accession no. 374.
5. Straub RE, Speer MC, Luo Y, Rojas K, Overhauser J, Ott J, et al. A microsatellite genetic linkage map of human chromosome 18. Genomics 1993;15:48–56.
6. Kimpton CP, Walton A, Gill P. A further tetranucleotide repeat polymorphism in the VWF gene. Hum Molec Genet 1992;1:287.
7. Polymeropoulos MH, Xiao Hi, Rath DS, Merril CR. Tetranucleotide polymorphism at the human tyrosine hydrolase gene (TH). Nucleic Acids Res 1991;19:3753.
8. Mills KA, Even D, Murray JC. Tetranucleotide repeat polymorphism at the human alpha fibrinogen locus (FGA). Hum Molec Genet 1992;1:779.
9. Sharma V, Litt M. Tetranucleotide repeat polymorphism at the D21S11 locus. Hum Molec Genet 1992;1:67.
10. Sullivan KM, Mannucci A, Kimpton CP, Gill P. A rapid and quantitative DNA sex test: fluorescence-based PCR analysis of X-Y homologous gene amelogenin. BioTechniques 1993;15:636–41.
11. Evett IW, Gill PD, Lambert JA, Oldroyd N, Frazier R, Watson S, et al. Statistical analysis of data for three British ethnic groups from a new STR multiplex. Int J Legal Med 1997;110:5–9.
12. Balding DJ, Nichols RA. DNA profile match probability calculations: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int 1994;64:125–140.

Additional information and reprint requests:
Dr. Ian Evett
Forensic Science Service
Metropolitan Laboratory
109 Lambeth Road
London SE1 7LP, UK

ERRATUM

Erratum/Correction of Evett IW, Foreman LA, Lambert JA, Emes A. Using a tree diagram to interpret a mixed DNA profile. J Forensic Sci 1998 May;43(3):472–76

Sir:

Since publication of the above referenced paper, we have noticed a couple of errors in our analysis. Please notice the following corrections.

1. *Figure 3*—For hypotheses 4 & 12, the numbers in the final two columns should be doubled; i.e. $Pr(E_3|H_{.2}, H_{.3})$ changes from 1.8×10^{-12} to 3.6×10^{-12}, $Pr(E|H_4)$ changes from 9×10^{-13} to 1.8×10^{-12} and $Pr(E|H_{12})$ changes from 27×10^{-18} to 5.4×10^{-18}. This is to take account of the 2 different ways that two unknown people can contribute profiles matching Lisa & Pauline in stain 2.
2. *Likelihood ratio*—The numerical (and, hence, algebraic) analyses described in the paper are flawed since the probability of the evidence, *E*, given the composite hypothesis *The knife bears DNA from both Lisa and Pauline* does not equal the sum of the values in the final column of Fig. 3 corresponding to hypotheses 1, 3, 5, and 7. Similarly, for the probability of *E* given the complementary hypothesis. In order to evaluate a likelihood ratio, we must focus on just 2 competing hypotheses, H_p for the numerator and H_d for the denominator. After discussion with the scientist, we can use the tree diagram of Fig. 2 and the probabilities specified in Fig. 3 to identify the most "suitable" hypotheses for comparison in the likelihood ratio. For example:

   - Choosing $H_p = H_1$ gives the maximum value of $Pr(E|H_i)$ for H_i which include both Lisa & Pauline. Choosing $H_d = H_{11}$ gives the maximum value of $Pr(E|H_i)$ for H_i which exclude both Lisa & Pauline. The resulting likelihood ratio is given by $1/(p_L p_P) = 6 \times 10^{11}$.
   - Alternatively, choosing $H_p = H_3$ or H_7 maximises $Pr(E|H_i)$ for H_i which include both Lisa & Pauline plus 1 unknown person, giving a reduced LR of $1/p_P = 1.67 \times 10^6$.

   In this way, a range of LR values can be identified corresponding to the comparison of plausible alternatives for H_p and H_d. In this particular case, all LR values in this range provided very strong support for the presence of blood from both Lisa & Pauline on the knife.

Editor's Note: Any and all future citations of the above-referenced paper should read: Evett IW, Foreman LA, Lambert JA, Emes A. Using a tree diagram to interpret a mixed DNA profile. [published erratum appears in J Forensic Sci 1999 Mar;44(2)] J Forensic Sci 1998;43:472–76.