Interpreting DNA Mixtures*


ABSTRACT: The interpretation of mixed DNA stains is explained in the context of likelihood ratios. The probabilities for the mixed-stain profile are evaluated under alternative explanations that specify the numbers of contributors and the profiles of any known contributors. Interpretations based simply on the frequencies with which random members of a population would not be excluded from a mixed-stain profile do not make use of all the information, and may overstate the strength of the evidence against included people. The effects of the numbers of contributors depends on whether all the alleles at a locus are present in the mixed stain. A general equation is given to allow likelihood ratios to be calculated, and includes the “2p” modification suggested by the 1996 NRC report. This modification is not always conservative. A computer program to perform calculations is available.

KEYWORDS: forensic science, mixed stains, DNA profiles, likelihood ratios

For a large variety of crimes, the use of genetic markers has proved invaluable in identifying perpetrators or exonerating falsely accused suspects. However, the interpretation of genetic profiles from biological samples can be complicated when the samples contain material from more than one person. This is especially common in rape cases: The sample may contain material from the victim or consensual sexual partners as well as from the perpetrator, or there may be multiple perpetrators. Mixed-stain evidence was discussed by Evett et al. (1991), and more recently by Aitken (1995). Their method, which is the same as the one we use, has been endorsed by the National Research Council (1996).

This approach differs from previous treatments (National Research Council, 1992) that calculate only the proportion of the population that is either included or excluded in a mixed stain. Such calculations ignore the profiles of specific people associated with the mixture, and so are analogous to the “random man not excluded” probabilities in paternity cases (e.g., Walker et al., 1983). Our calculations do consider specific profiles, and so are more like the “paternity index.” Moreover, the conventional calculations do not consider the number of contributors to the mixture, and therefore miss the interaction between the number of alleles and the number of contributors on the numerical strength of the evidence of a mixed stain. Only by comparing the probabilities of the evidentiary profile under alternative explanations is it possible to arrive at a complete analysis of mixtures. Indeed, we show by example that the conventional method can overstate the strength of evidence against a defendant not excluded from a mixed stain.

In this paper, we offer some rules to simplify the interpretation of mixed stains, and phrase these rules in terms of the probabilities of certain subsets of the alleles present in the mixed stain. The subsets are the alleles that come from people whose identity is not known with certainty. The rules we describe give an immediate way of deriving results such as those given as examples by the National Research Council (1996). Our results begin with the assumption that the number of contributors to the mixed stain is known, and then we point to methods that make conservative assumptions about this number. We also consider how to interpret mixed VNTR profiles, when there may be unseen alleles.

To keep the treatment as simple as possible, only discrete genetic systems will be treated. These include conventional blood group markers, binned RFLP markers, as well as the current PCR-based markers. We assume that a profile has been determined for an evidentiary sample, that there are indications of more than one contributor to the sample, and that the sample may contain the profiles of one or more known people. The task is to assign a numerical weight to this evidence, and this in turn requires that databases are available from which to estimate frequencies of components of the profile.

As a further simplification, we assume independence of alleles, within and between loci. We develop methods for handling single-locus profiles and then multiply results over loci. We also defer the issue of population structure and relatedness, both of which cause dependencies between the profiles of different people.

Likelihood Ratios

The convenient likelihood ratio framework for assigning weight to evidence will be used (e.g., Aitken, 1995; Balding and Nichols, 1994; Evett, 1983; Lindley, 1977; Weir, 1994). The likelihood ratio for comparing two alternative explanations, \( C \) and \( \overline{C} \), for the evidentiary profile is formed as the ratio of the probabilities of the profile under the alternatives. The likelihood ratio is written as \( L \) and the profile as \( E \), so

\[
L = \frac{Pr(E|C)}{Pr(E|\overline{C})}
\]

If a jury is being asked to make a choice between explanations \( C \) and \( \overline{C} \), it can be told “The profile \( E \) is \( L \) times more likely to have arisen under explanation \( C \) than under explanation \( \overline{C} \).”
**General Notation**

We have found it very helpful to introduce notation that makes calculation of the likelihood ratio somewhat automatic. The key is to specify who are the known contributors to the profile, and therefore which alleles in the profile may be from unknown contributors, under each explanation. The likelihood ratio becomes the ratio of the probabilities of the sample profile, taking into account both the known and unknown contributors. For example, suppose an evidentiary sample has four alleles $abcd$ at some locus and that it is known the sample contains DNA from the victim and the perpetrator. Suppose, further, the victim has type $ab$ and a suspect has type $cd$. The prosecution explanation $C$ would be that the known contributors are the victim and the suspect, whereas the defense explanation $C$ may be that the victim is a known contributor, but that alleles $cd$ came from an unknown contributor.

- The probability of the sample stain under $C$ is one, because all four alleles can be explained with certainty. There are no unknown contributors, and the probability is written as $P_0(\phi|abcd) = 1$. The subscript 0 indicates the number of unknown contributors, the quantity $\phi$ indicates the "empty" set of alleles that were from unknown contributors, and $abcd$ is the sample profile.

- The probability of the sample profile under explanation $C$ is written as $P_1(cd|abcd)$. Now the subscript 1 indicates the one unknown contributor, the set $cd$ is the alleles from this unknown person, and $abcd$ is still the whole profile.

- The likelihood ratio is

$$L = \frac{P_0(\phi|abcd)}{P_1(cd|abcd)} \quad (1)$$

In general, the evidentiary profile contains a set $E$ of alleles at some locus. There may be known contributors to the profile, having some or all of these alleles between them. Under at least one of the alternative explanations, there are unknown contributors. These people must carry the set $U$ of alleles in $E$ not carried by the known contributors, and they cannot carry any alleles not in $E$. If there are $x$ unknown contributors to the profile, then what is needed is the probability $P_x(U|E)$ that these $x$ people have the alleles in $U$ between them, but do not have any alleles not in $E$. When there are no unknown contributors, the symbol $P_0$ will be used.

We have found that this approach and notation allows us to approach any situation. For example, suppose a victim reports having been raped by two men, and one suspect is identified. Suppose, also, it is known that there have been no consensual partners. The evidentiary sample is assumed to contain DNA from the victim and the perpetrators, and is found to have alleles $abcd$. (Although preferential extraction, Gill et al., 1985, may minimize the DNA from the victim in the male fraction, it may not eliminate her DNA.) The victim is of type $ab$ and the suspect of type $cd$. Explanation $C$ is that there is one unknown contributor to the sample (the second perpetrator), but this person is not required to have any specific alleles. He cannot have alleles other than $abcd$. Under explanation $C$, there are two unknown contributors (both perpetrators) and these two must have contributed alleles $cd$ between them. The likelihood ratio is

$$L = \frac{P_1(\phi|abcd)}{P_2(cd|abcd)} \quad (2)$$

**Calculation of Probabilities**—We now illustrate how the $P_x(U|E)$ probabilities can be calculated. We also give a general formula, suitable for computer programs, and list the cases most likely needed in casework in Appendix 1.

**Profiles with One Allele**

If the profile has only alleles of type $a$ at a locus, then all contributors must be $aa$ homozygotes. The possibility of unseen alleles will be addressed later. The profile at other loci, or other factors, indicate that the evidentiary stain is a mixture. Assuming independence of alleles within loci, the probability that one unknown contributor has allele $a$, and no allele other than $a$, is

$$P_x(\phi|a) = p_a^x$$

where $p_a$ is the population frequency of allele $a$, and $p_a^x$ is the frequency of homozygotes in the population. If there are $x$ unknown contributors, they must all be $aa$ homozygotes and

$$P_x(a|a) = p_a^{2x}$$

If a known contributor to the profile is homozygous $aa$, then there are no unexplained alleles. The probability needed, for $x$ unknowns, is $P_x(\phi|a)$, where $\phi$ is the empty set. Although these unknown people do not need to provide a specific allele, in fact they can only be of type $aa$, so

$$P_x(\phi|a) = p_a^{2x}$$

This is a good place to point out that allele frequencies like $p_a$ refer to the “relevant” population: those people who might be considered potential contributors to the evidentiary sample. This population is defined by circumstances of the crime, rather than by attributes of the suspect. The circumstances may prescribe a particular ethnic group, in which case frequencies should be used for that group. Otherwise, frequencies from different ethnic groups should be considered. The ethnicity of the suspect is not material as we are not considering population structuring.

**Profiles with Two Alleles**

If the profile has alleles $ab$, then all contributors must be of genotype $aa$, $ab$ or $bb$. If the source of only allele $a$ is in question, the contributor of that allele must be of type $aa$ or $ab$. People of type $ac$, where $c$ is any other allele, are excluded because the sample profile does not contain $c$. Therefore the probability that one unknown contributor has allele $a$, and no allele other than $a$ or $b$, is

$$P_x(a|ab) = p_a^x + 2p_ap_b$$

$$= (p_a + p_b)^2 - p_b^2$$

The generalization to $x$ contributors is

$$P_x(a|ab) = (p_a + p_b)^{2x} - p_b^{2x}$$

This expression gives the probability of all combinations of $x$ people who have only alleles $a$ and $b$, and it excludes only the situation where all $x$ are $bb$ homozygotes because that set of people could not contribute allele $a$. When $x = 2$, for example, the possible
sets of people are \( aa, aa; aa, ab; aa, bb; ab, ab; ab, bb \). The probabilities of these sets are \( p_a^4, 4p_a^3p_b, 6p_a^2p_b^2, 4p_ap_b^3 \) which add to \( (p_a + p_b)^4 \). If unknown contributors already have alleles \( ab \), then unknowns need not have any specific alleles, but they cannot have alleles other than \( ab \), so

\[
P_x(ab) = (p_a + p_b)^2x
\]

When \( x = 2 \), the possible sets of people are \( aa, aa; aa, ab; aa, bb; ab, ab; ab, bb \). The probabilities of these sets add to \( (p_a + p_b)^4 \). If known contributors already have alleles \( ab \), then unknowns need not have any specific alleles, but they cannot have alleles other than \( ab \), so

\[
P_x(ab) = (p_a + p_b)^2x - p_a^2 - p_b^2
\]

This allows for all combinations of \( x \) people who have no alleles other than \( a \) or \( b \), except the cases where all \( x \) people are \( aa \) or all are \( bb \). These last two cases could not result in an \( ab \) profile for the mixed stain. When \( x = 2 \), the possible sets of people are \( aa, ab; aa, bb; ab, ab; ab, bb \). These sets have probabilities adding to \( (p_a + p_b)^4 \). Finally, if there are no known contributors, then both alleles \( ab \) must be carried by the unknown contributors

\[
P_x(ab) = (p_a + p_b)^2x \quad x > 1
\]

Profiles with Three Alleles

If the profile has alleles \( abc \), then all contributors must be of genotype \( aa, ab, bb, ac, bc, \) or \( cc \). If known contributors already have alleles \( abc \), then unknowns need not have any specific alleles but cannot have alleles other than \( abc \), so

\[
P_x(abc) = (p_a + p_b + p_c)^2x
\]

The probability that \( x \) unknown contributors have allele \( a \), and no allele other than \( a, b \), or \( c \), is

\[
P_x(a) = (p_a + p_b + p_c)^2x - (p_a + p_c)^2x
\]

When \( x = 1 \) this simplifies to \( p_a^2 + 2p_ap_b + 2p_ap_c \), corresponding to \( aa, ab, \) or \( ac \) genotypes.

The probability that \( x \) unknown contributors have alleles \( ab \), and no allele other than \( ab \) or \( c \), is

\[
P_x(ab) = (p_a + p_b + p_c)^2x - (p_a + p_c)^2x
\]

When \( x = 1 \), this simplifies to \( 2p_ap_b \) as only \( ab \) heterozygotes can contribute both \( a \) and \( b \).

Finally, if there are no known contributors, then all alleles \( abc \) must be carried by the unknown contributors, which means that \( x > 1 \), and

\[
P_x(abc) = (p_a + p_b + p_c)^2x - (p_a + p_b)^2x
\]

This expression is zero when \( x = 1 \), as it is not possible for one person to contribute all of alleles \( a, b \), and \( c \).

Profiles with Four Alleles

Similar arguments lead to the four-allele results:

\[
P_x(abcd) = (p_a + p_b + p_c + p_d)^2x
\]

Profiles with \( m \) Alleles

The general expressions for \( x \) contributors when there are three or four alleles seem fairly cumbersome, but they follow a pattern.
We noticed that a single equation can be given for the probability that \(x\) unknown contributors have at least the alleles in some set of alleles \(U\), but have no alleles not in the mixed-profile allele set \(E\). A formal proof of the equation was supplied by C.H. Brenner and is contained as Appendix 2 to this paper. The equation is

\[
P_x(U|E) = (T_0)^{2x} - \sum_j (T_{j})^{2x} + \sum_{j,k} (T_{j,k})^{2x} - \sum_{j,k,l} (T_{j,k,l})^{2x} + \cdots
\]

where \(T_0\) is the sum of probabilities of all alleles in \(E\), \(T_j\) is the sum of probabilities of all alleles in \(E\) except for the \(j\)th allele in \(U\), \(T_{j,k}\) is the sum of probabilities of all alleles in \(E\) except for the \(j\)th and \(k\)th alleles of \(U\), and so on. Such equations are useful for constructing computer programs, and we give instructions below for obtaining such a program. However, the explicit results we list in Appendix 1 will generally be sufficient.

As an illustration of the use of Appendix 1, suppose an evidentiary sample \(E\) has only three alleles \(abc\) and includes contributions from a victim and a perpetrator. If the victim has genotype \(ab\) and a suspect has profile \(aa\), then the explanation \(C\) including the suspect still requires unknown contributors for allele \(c\). The alternative explanation \(\overline{C}\) may still include the victim but excludes the suspect, and so includes some unknown contributors. If the circumstances of the crime specify that there were three contributors,

\[
L = \frac{P_x(abc)}{P_x(ab)} = \frac{p_a(2p_a + 2p_b + p_c)}{(p_a + p_b + p_c)^2 - (p_a + p_b)^2}\]

\[
= \frac{p_a}{p_a + p_b + p_c}\]  

\[
(4)
\]

### Number of Contributors

The results give so far depend on the number of contributors to the mixed sample. Sometimes the circumstances of the crime will dictate that number, and the forensic scientist must use that number in the numerical analysis. Other times, the number may not be known but the presence of more than two alleles at a locus implies that there was more than one contributor. However, if all the loci in the profile had no more than four alleles, but these loci had more than four alleles in the population, an argument could be made that explanations with other than two contributors may safely be ignored. Other situations that may indicate the number of contributors are when a mixed stain occurs amongst a collection of single-contributor stains.

Whenever there is doubt as to the number of contributors, there can be considerable variation in the likelihood ratio. The issue has been discussed in some detail by C.H. Brenner, R. Fimmers and M. Baur (submitted, personal communication from C.H. Brenner). For loci with more alleles than are present in the sample profile, the number of contributors increases. Conversely, smaller \(x\) values for the denominator will make the likelihood ratio smaller and will weaken the evidence against suspects included under \(C\) but not \(\overline{C}\).

When all the alleles at a locus are present in the evidentiary sample, however, an increased number of contributors makes the profile more likely. It becomes easier for large numbers of people to have all the alleles at the locus between them. Equation 3 now shows that \(P_x(U|E)\) → 1 as \(x\) increases since \(T_0 = 1\). The likelihood ratio is now smaller for larger \(x\) values.

These results on the effect of \(x\) on \(P_x(U|E)\) are strictly true only when \(x\) is large and may not hold for small \(x\). Care is needed, and it is advisable to obtain numerical values for several values of \(x\).

### Examples

We now show the effects of both the numbers of contributors and the profiles of specific people, by considering an example of a rape case (R. Cotton, personal communication). The Polymarker™ profile from a vaginal swab, as well as profiles from the victim and a suspect, are shown in Table 1, along with some allele frequencies used for illustrative purposes. Both the victim and the suspect profiles are included in the evidence profile.

If the two explanations for this evidence are:

- **C**: Contributors were the victim and a suspect,
- **\(\overline{C}\)**: Contributors were the victim and an unknown person.

then the evidence is certain under \(C\). The probabilities under \(\overline{C}\) are shown in Table 2, and the likelihood ratio is 8.9.

Evidence from a rape case may be collected from somewhere other than victim’s body, and explanation \(\overline{C}\) may then be that both contributors are unknown. With \(\overline{C}\) still that the evidence is from the victim and the suspect, the probabilities are as shown in Table 3. The likelihood ratio is now 46.6, and has increased because of the larger number of unknowns in the denominator. The evidence did not contain all the alleles at each locus.

Finally, consider the case where the evidence in Table 1 is known to be from two perpetrators, but only one (the “suspect” of Table 1) is known. The evidence is no longer regarded as being

### Table 1—Polymarker™ example.

<table>
<thead>
<tr>
<th>Profile</th>
<th>LDLR</th>
<th>GYPa</th>
<th>HBGG</th>
<th>D7S8</th>
<th>Gc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence</td>
<td>B</td>
<td>AB</td>
<td>AB</td>
<td>AB</td>
<td>ABC</td>
</tr>
<tr>
<td>Victim</td>
<td>B</td>
<td>AB</td>
<td>AB</td>
<td>AB</td>
<td>AC</td>
</tr>
<tr>
<td>Suspect</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>(p_a)</td>
<td>0.567</td>
<td>0.462</td>
<td>0.429</td>
<td>0.543</td>
<td>0.253</td>
</tr>
<tr>
<td>(p_b)</td>
<td>0.583</td>
<td>0.566</td>
<td>0.457</td>
<td>0.195</td>
<td>0.552</td>
</tr>
</tbody>
</table>

### Table 2—Calculations for rape case, one unknown.

| Pr(E|C) | Pr(E|\overline{C}) |
|------|------------------|
| LDLR | 1 \(P_x(BI) = p_b\) | 0.321 |
| GYPa | 1 \(P_x(\phi IAB) = (p_a + p_b)^2\) | 1.000 |
| HBGG | 1 \(P_x(\phi IAB) = (p_a + p_b)^2\) | 0.990 |
| D7S8 | 1 \(P_x(\phi IAB) = (p_a + p_b)^2\) | 1.000 |
| Gc   | 1 \(P_x(BIABC) = p_d(p_a + 2p_a + 2p_r)\) | 0.352 |
Profiles with Unseen Alleles

Adding Null Allele

An RFLP locus can give rise to a single band $c$ on an electrophoretic gel if it is homozygous $cc$, or if it is heterozygous $cn$ for $c$ and some unseen allele $n$. Such an unseen allele may need to be added to profiles of people for whom only one band is seen, and then added to mixture profiles to which that person is supposed to have contributed.

For example, suppose the evidentiary sample has profile $abc$, the victim has type $ab$ and a suspect has type $c$. If it is appropriate to assume the existence of unseen bands, the suspect's profile could be $cc$ or $cn$ and the sample profile $abc$ or $abcn$. In either case, the evidence is certain under explanation $C$ that the contributor was the victim and suspect. However, if $\bar{C}$ includes the victim but excludes the suspect, the unknown contributor of allele $c$ must possess allele $c$ and not possess an allele, other than $a$ or $b$, that would be "seen." The possible genotypes for this person are $ac$, $bc$, $cc$, or $nc$ and

$$L = \frac{1}{P_c(ab|bcn)}$$

Estimates of null frequencies have all been less than 0.05 (Chakraborty et al., 1994).

Although null alleles require care, the machinery previously established is adequate provided the null $n$ is added to the evidentiary profile. It need not appear in the set $U$ in the expression $P_x(U|E)$.

The "$2p$" Rule

For the interpretation of single-contributor stains, it is customary to estimate the frequency of single-band VNTR patterns as twice the frequency assigned to that band. If the frequency of allele $a$ is $p_a$, then the total frequency of individuals showing only that allele is $p_a + 2p_b + 2p_c$, where $n$ is the possible unseen allele. The customary rule is the conservative replacement of this expression by $2p_a$, and the National Research Council (1996) recommends the same procedure for mixed stains. The general expression for the probabilities $P_x(U|E)$ can be modified to accommodate this rule, and we begin with some examples. Once the possibility of single bands is allowed for alleles from known contributors, we allow it for all contributors (and all alleles).

For a profile with two alleles, $ab$, contributors must be $a^*$, $b^*$ or $ab$ where $a^*$ or $b^*$ means individuals showing only allele $a$ or $b$. They may be homozygotes or heterozygotes for an unseen allele, and are assigned a frequency of $2p_a$, $2p_b$. In the same notation as before

$$P_x(ab) = 2[p_a + 2p_ap_b]$$

with a generalization to $x$ contributors of

$$P_x(ab|bcn) = 2[p_a + p_b + p_ap_b - p_b]$$

For a profile with three alleles, $abc$, contributors must be $a^*$, $b^*$, $c^*$, $ab$, $ac$, or $bc$. For two contributors to have at least $ab$ but no more than $abc$ between them, the possibilities are $a^*$ with $b^*$ or $bc$; and $b^*$ with $ab$ or $ac$; and $ab$ with $c^*$, $ac$ or $bc$; and $ac$ with $bc$. Therefore

$$P_x(ab|abc) = 4[p_a + 2p_ap_b + 2p_ap_c] + 4p_ap_b(2p_a + p_b + 2p_c + 2p_p) + 8p_ap_b[p_c]^2$$

with a generalization to $x$ contributors of

\[
P_x(ab|abcn) = 4[p_a + p_b + p_c + p_ap_b + p_ap_c + p_ap_b + p_ap_c + p_ap_b + p_c]^2
\]

\[
-(p_a + p_b + p_c)^2 - (p_a + p_c + p_c)^2 + p_c^2
\]
### TABLE 5—Two-contributor examples from 1996 NRC Report. The contributors are the suspect and one unknown person under C, and two unknown people under C.

<table>
<thead>
<tr>
<th>E</th>
<th>Suspect</th>
<th>General</th>
<th>( p^2 )</th>
<th>Likelihood Ratio</th>
<th>2p</th>
</tr>
</thead>
<tbody>
<tr>
<td>abcd</td>
<td>ab</td>
<td>( \frac{P_1(abcd)}{P_2(abcd)} )</td>
<td>( \frac{1}{12p_1p_2} )</td>
<td>( \frac{1}{12p_1p_2} )</td>
<td></td>
</tr>
<tr>
<td>abc</td>
<td>bc</td>
<td>( \frac{P_1(ab)}{P_2(ab)} )</td>
<td>( \frac{p_a + 2p_b + 2p_c}{12p_1p_2(p_a + p_b + p_c)} )</td>
<td>( \frac{1}{4p_1p_2(3 + p_a + p_b + p_c)} )</td>
<td></td>
</tr>
<tr>
<td>abc</td>
<td>a</td>
<td>( \frac{P_1(abc)}{P_2(abc)} )</td>
<td>( \frac{1}{6p_1(p_a + p_b + p_c)} )</td>
<td>( \frac{1}{4p_1(3 + p_a + p_b + p_c)} )</td>
<td></td>
</tr>
<tr>
<td>ab</td>
<td>ab</td>
<td>( \frac{P_1(ab)}{P_2(ab)} )</td>
<td>( \frac{(p_a + p_b)^2}{2p_1p_2(2p_a^2 + 3p_b^2 + 2p_c^2)} )</td>
<td>( \frac{p_a + p_b + p_c}{2p_1p_2(2 + 2p_a + 2p_b + p_c)} )</td>
<td></td>
</tr>
<tr>
<td>ab</td>
<td>a</td>
<td>( \frac{P_1(ab)}{P_2(ab)} )</td>
<td>( \frac{2p_a + p_b}{2p_1(2p_a^2 + 3p_b^2 + 2p_c^2)} )</td>
<td>( \frac{1 + p_a}{2p_1(2 + 2p_a + p_b + p_c)} )</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 6—Calculations for a three-allele VNTR mixed profile. Under C, the known contributors have all three alleles abc. Under \( \overline{C} \) all contributors are unknown.

<table>
<thead>
<tr>
<th>No. unknowns under ( C )</th>
<th>No. of unknowns under ( C )</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>( p^2 )</td>
<td>( \frac{P_3(abc)}{P_2(abc)} )</td>
<td>1,620</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>( 2p )</td>
<td>( \frac{P_3(abc)}{P_2(abc)} )</td>
<td>158</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>( n )</td>
<td>( \frac{P_3(abc)}{P_2(abc)} )</td>
<td>3,380</td>
<td>226</td>
</tr>
<tr>
<td>3</td>
<td>( p^2 )</td>
<td>( \frac{P_3(abc)}{P_2(abc)} )</td>
<td>21,600</td>
<td>938</td>
</tr>
<tr>
<td></td>
<td>( 2p )</td>
<td>( \frac{P_3(abc)}{P_2(abc)} )</td>
<td>54</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>( n )</td>
<td>( \frac{P_3(abc)}{P_2(abc)} )</td>
<td>12,300</td>
<td>823</td>
</tr>
<tr>
<td>4</td>
<td>( p^2 )</td>
<td>( \frac{P_3(abc)}{P_2(abc)} )</td>
<td>396,000</td>
<td>17,200</td>
</tr>
<tr>
<td></td>
<td>( 2p )</td>
<td>( \frac{P_3(abc)}{P_2(abc)} )</td>
<td>8</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>( n )</td>
<td>( \frac{P_3(abc)}{P_2(abc)} )</td>
<td>108,000</td>
<td>7,250</td>
</tr>
</tbody>
</table>
$P_{x}(abc|abc) = 2\left(p_a + p_b + p_c + p_a p_b + p_a p_c + p_b p_c\right)^x$
\[\cdots\]  

The "2$p$" modification of Equation 3, for $x$ unknowns having at least the alleles in $U$ but no alleles not in $E$, is

$$P_x(U|E) = 2\left[(S_0) + \sum_j (S_{j}) + \sum_{j,k} (S_{j,k}) - \sum_{j,k,l} (S_{j,k,l})\right]$$

where $S_0$ is the sum of all the probabilities of the alleles in $E$ plus the sum of all the products of frequencies for alleles in $E$. Term $S_{j}$ is $S_0$ with all the terms involving the $j$th allele of $U$ removed, $S_{j,k}$ is $S_0$ with all the terms involving the $j$th and $k$th alleles in $U$ removed, and so on. To see why this result holds, note that $S_0$ is the square of the sum of the frequencies of all the alleles in $E$, with each squared frequency ($p^2$) replaced by twice that frequency ($2p$). The most common examples are shown in Appendix 1, and in Table 5 we confirm the results given by the National Research Council (1996). We adopt the language of that report in writing "2$p$" for calculations where unseen bands are not an issue, and "2$p$" where single band profiles are given twice the frequency of the band that is seen. It must be pointed out that the "2$p$" rule cannot be used for loci with few alleles when $S_0$ is greater than one.

A comparison of the unseen-allele and the "2$p$" approaches is afforded by a three-band profile $abc$ at D2S44 recovered from the center console of an automobile owned by the defendant in the case of People v. Simpson (Los Angeles County Case BA097211). This profile included the profile $ab$ of the defendant OS and the profile $ac$ of the victim RG (it did not include a second victim NB). Some frequencies for illustrative purposes are $p_a = 0.0316$, $p_b = 0.0842$, $p_c = 0.0926$. Three sets of calculations are laid out in Table 6: The possibility of unseen alleles could be ignored, an unseen allele (with frequency 0.05) could be allowed, or the "2$p$" rule could be used. In this instance, the court ordered that the number of contributors to the evidence profile be set to two, three or four. The "2$p$" rule is not always conservative, and we suggest caution in its use.

Discussion

The interpretation of mixed stains is straightforward in the likelihood ratio context. Alternative explanations for the mixed stain profiles need to be specified, and then compared on the basis of the probabilities of the profile under those explanations. Calculations that consider only single contributors are without a logical foundation. The 1992 report of the United States National Research Council recommended "If a suspect's pattern is found within the mixed pattern, the appropriate frequency to assign such a 'match' is the sum of frequencies of all the genotypes that are contained within (i.e., that are a subset of) the mixed pattern." We have shown by example that this method is not helpful, and may even be prejudicial. Indeed, the 1996 National Research Council report repudiates the 1992 recommendation: "This calculation is hard to justify, because it does not make use of some of the information available, namely, the genotype of the suspect. The correct procedure, we believe, was described by Evett et al. (1991)."

A program to calculate the likelihood ratio for mixed stains, for a range of numbers of unknown contributors under both $C$ and $\overline{C}$, can be obtained from the senior author. (e-mail address: weir@ncsu.stat.edu or World Wide Web URL: http://www2.ncsu.edu/~su.stat.edu or World Wide Web URL: http://www2.ncsu.edu//ncsu.CIL/stat__genetics/).

Acknowledgments

The wise counsel of Dr. Ian Evett, and the insights of Dr. Charles Brenner, are greatly appreciated.

References

## APPENDIX 1—SPECIAL CASES

### One allele

<table>
<thead>
<tr>
<th>Allele</th>
<th>$p^2$ rule</th>
<th>2p rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1(\phi(a))$</td>
<td>$(p_a + p_b)^2$</td>
<td>$2(p_a + p_b + p_dp_e)^2$</td>
</tr>
</tbody>
</table>

### Two alleles

<table>
<thead>
<tr>
<th>Allele</th>
<th>$p^2$ rule</th>
<th>2p rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1(\phi(ab))$</td>
<td>$(p_a + p_b)^2$</td>
<td>$2(p_a + p_b + p_dp_e)^2$</td>
</tr>
<tr>
<td>$P_2(\phi(ab))$</td>
<td>$(p_a^2 + 4p_ap_b + 6p_ap_b + p_b^2)$</td>
<td>$4p_a(1 + p_b)(p_a + p_b + 2p_b^2)$</td>
</tr>
<tr>
<td>$P_3(\phi(ab))$</td>
<td>$(p_a + p_b)^2$</td>
<td>$2[(p_a + p_b + p_dp_e)^2 - p_b^2]$</td>
</tr>
</tbody>
</table>

### Three alleles

<table>
<thead>
<tr>
<th>Allele</th>
<th>$p^2$ rule</th>
<th>2p rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1(\phi(abc))$</td>
<td>$p^2$</td>
<td>$(p_a + p_b + p_c)^2$</td>
</tr>
<tr>
<td>$P_2(\phi(abc))$</td>
<td>$2p$</td>
<td>$2(p_a + p_b + p_c + p_dp_e + p_dp_e)^2$</td>
</tr>
<tr>
<td>$P_3(\phi(abc))$</td>
<td>$p^2$</td>
<td>$(p_a + p_b + p_c - (p_b + p_c)^2)$</td>
</tr>
<tr>
<td>$P_4(\phi(abc))$</td>
<td>$2p$</td>
<td>$2[(p_a + p_b + p_c + p_dp_e + p_dp_e + p_dp_e - (p_b + p_c + p_dp_e)^2]$</td>
</tr>
</tbody>
</table>

### Four alleles

<table>
<thead>
<tr>
<th>Allele</th>
<th>$p^2$ rule</th>
<th>2p rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1(\phi(abcd))$</td>
<td>$p^2$</td>
<td>$(p_a + p_b + p_c + p_d)^2$</td>
</tr>
<tr>
<td>$P_2(\phi(abcd))$</td>
<td>$2p$</td>
<td>$2(p_a + p_b + p_c + p_d + p_dp_e + p_dp_e + p_dp_e + p_dp_e + p_dp_e)^2$</td>
</tr>
<tr>
<td>$P_3(\phi(abcd))$</td>
<td>$p^2$</td>
<td>$p_a(p_a + 2p_b + 2p_c)$</td>
</tr>
<tr>
<td>$P_4(\phi(abcd))$</td>
<td>$2p$</td>
<td>$2[(p_a + p_b + p_c + p_dp_e + p_dp_e + p_dp_e - (p_b + p_c + p_dp_e)^2]$</td>
</tr>
</tbody>
</table>
APPENDIX 2—PROOF OF A MIXED STAIN FORMULA OF WEIR

The genetic markers ("alleles") of an evidence stain may be identical to the alleles of a reference sample (such as a suspect for example). The likelihood ratio for the evidentiary strength favoring association is then simply the inverse of the profile frequency. However, the evidence stain is often complicated by the presence of additional alleles, variously from additional known or unknown suspects or victims. The likelihood ratio is clearly more complicated in such cases, but Weir et al. (1997) present in the foregoing paper without proof a general and elegant formula for the probabilities that occur as numerator or denominator. In this appendix, we give a proof.

Notation

Let E be the set of alleles observed in the evidence for some discrete-allele system. Of these some may be attributable to known parties; the remainder U ⊆ E are to be explained by x people with two alleles (not necessarily distinct) each. Explained means that U ⊆ X ⊆ E, where X is the set of all alleles in the x people. More generally for a subset S ⊆ U, we will say that people with alleles X exactly explain S if X ⊆ E and X \ S = S, or equivalently, putting W = \_AS, if UkW ⊆ X ⊆ E\W.

The set-notation symbols are to be understood according to the standard conventions: U ⊆ E means that the alleles U are among those of E, including the possibility that U = E. X \ S is the intersection—the set of alleles that are both in X and in S. ~ is the set difference—the set of alleles that are in U excluding those that are also in S. The cardinality (size, in number of alleles) of a set J is written |J|. The symbol ∋ denotes membership; j ∋ U means that j is an allele of the set U.

Following Weir et al. we write P_x(U|E) for the probability that x random people explain U. Let J, J ⊆ U be a set of alleles. We will be interested in sets of people who omit the set J. Let

\[ T_{\text{maj}} = \text{the total of the frequencies of the alleles in } E \setminus J. \] (1)
Theorem

Weir discovered that

\[ P_x(U|E) = T_0^a - \sum_{j \in U} T_j^a + \sum_{j,k \in U} T_{j,k}^a - \sum_{j,k,l \in U} T_{j,k,l}^a + \cdots \]  

(2)

but did not supply a proof.

Proof

A proof seems worthwhile. The general idea is clear enough—(2) is an instance of the principle of inclusion and exclusion (Hall, 1967). From the definition (1) and the assumption of a discrete allele system, \( T_{m,j}^a \) is the probability that \( x \) people's alleles are all in \( E \setminus J \). As the basis for the inclusion-exclusion analysis, we note that

\[ T_{m,j}^a = \sum_{W \subseteq U} P_x(U \setminus W | E \setminus W) \]  

(3)

because any set of people whose alleles are among \( E \setminus J \) exactly explains some one and only one subset, \( U \setminus W \), of \( U \setminus J \). The summation is taken over all sets of alleles \( W \) that satisfy \( J \subseteq W \subseteq U \).

Introduction of the sets \( W \), as a means of effectively classifying the various positive and negative contributions to the sum in (2), is the key idea in the proof.

Define

\[ Q_m = \sum_{J \subseteq W \subseteq U} T_{m,j}^a \]  

(4)

In this notation, (2) is expressed as

\[ P_x(U|E) = Q_0 - Q_1 + Q_2 - + \cdots \]  

(5)

Summing (3) for fixed \( m \) over all sets \( J \subseteq U \) of cardinality \( m \) we obtain from (4)

\[ Q_m = \sum_{J \subseteq W \subseteq U} P_x(U \setminus W | E \setminus W) \]  

(6)

\[ = \sum_{m=0}^{n} \sum_{W \subseteq U} P_x(U \setminus W | E \setminus W) \]  

(7)

\[ = \sum_{m=0}^{n} \sum_{W \subseteq U} P_x(U \setminus W | E \setminus W) \sum_{k=0}^{m} \binom{m}{k} \]  

(8)

\[ = \sum_{m=0}^{n} \sum_{W \subseteq U} P_x(U \setminus W | E \setminus W) \sum_{k=0}^{m} \binom{m}{k} \]  

(9)

On the right hand side of line (6), each \( W \) occurs many times, once for each \( J \) of which it is a superset. The object is to count how many times. Classifying the \( W \)'s according to their size \( k \) on line (7), we see on (8) that it is the same as the number of \( m \) allele subsets of a \( k \)-set, which is exactly the definition of the binomial symbol \( \binom{m}{k} \). Hence line (9).

To verify (5) form now the alternating sum over \( m \), where the sum runs to \( n = |U| \),

\[ Q_0 - Q_1 + Q_2 - + \cdots \]

\[ = \sum_{m=0}^{n} (-1)^m Q_m \]

\[ = \sum_{m=0}^{n} (-1)^m \sum_{k=0}^{m} \sum_{W \subseteq U \mid |W|=k} P_x(U \setminus W | E \setminus W) \binom{m}{k} \]  

(10)

\[ = \sum_{m=0}^{n} \sum_{W \subseteq U \mid |W|=k} P_x(U \setminus W | E \setminus W) \sum_{k=0}^{m} (-1)^m \binom{m}{k} \]  

(11)

As line (10) shows the same set \( W \) may occur in several \( Q_m \) terms. To compute the net contribution due to each \( W \), it is natural to reverse the order of summation so that the classification is on \( W \) first and then on \( m \), which is formula (11). To verify the transition from (10) to (11) note that the index sets of the double summations \( \sum_{m=0}^{n} \sum_{k=0}^{m} \) and \( \sum_{k=0}^{m} \sum_{j=0}^{m} \) range over the same pairs \((k, m)\)—namely the triangular array where \( 0 \leq m \leq k \leq n \).

Hence the net number of times that a contribution from each set \( W \) is included and excluded is given by the last factor in (11). That factor is simply unity when \( k = 0 \), and when \( k > 0 \) is it even simpler, for by the binomial theorem \( \sum_{m=0}^{n} (-1)^m \binom{m}{k} = (1 - t)^k = 0 \). So

\[ Q_0 - Q_1 + Q_2 - + \cdots = P_x(U|E) \]

\[ + \sum_{k=1}^{n} \sum_{W \subseteq U \mid |W|=k} P_x(U \setminus W | E \setminus W) \]  

\[ = P_x(U|E), \text{ Q.E.D.} \]

References


Additional information and reprint requests:
Charles H. Brenner, Ph.D.
2486 Hilgard Ave
Berkeley CA 94709