James M. Curran,[1] *Ph.D.; Christopher M. Triggs,*[2] *Ph.D.; John Buckleton,*[3] *Ph.D.; and B. S. Weir,*[1] *Ph.D.*

Interpreting DNA Mixtures in Structured Populations

REFERENCE: Curran JM, Triggs CM, Buckleton J, Weir B.S. Interpreting DNA mixtures in structured populations. J Forensic Sci 1999;44(5):987–995.

ABSTRACT: DNA profiles from multiple-contributor samples are interpreted by comparing the probabilities of the profiles under alternative propositions. The propositions may specify some known contributors to the sample and may also specify a number of unknown contributors. The probability of the alleles carried by the set of people, known or unknown, depends on the allelic frequencies and also upon any relationships among the people. Membership of the same subpopulation implies a relationship from a shared evolutionary history, and this effect has been incorporated into the probabilities. This acknowledgment of the effects of population structure requires account to be taken of all people in a subpopulation who are typed, whether or not they contributed to the sample.

KEYWORDS: forensic science, DNA typing, interpretation, mixed DNA profiles, population structure, likelihood ratios

The interpretation of DNA profiles from more than one contributor is one of the most challenging tasks facing forensic scientists. Part of the complexity is due to the very large number of combinations of genotypes that must be considered in some situations, although a body of theory for a coherent treatment of mixed stains is now available (1–3). For a defendant who is not excluded from a mixed stain this theory avoids the potential prejudice that can follow from simplistic "random man not excluded" arguments.

In some cases, the typing technology may allow complexity to be avoided. When fragments are detected in ways that allow semiquantitation of the amount of DNA for each allele it may be possible to determine which alleles are from the same contributor. Examples include fluorescently-labeled length variants detected by lasers, or silver staining to detect band intensity on a gel. There can still be doubt, however, especially when different people contribute more or less equally to the mixture and such problems increase with the number of contributors. As long as a quantitative assessment of the evidentiary strength of DNA mixtures is required, we believe that there will be a need for analyses that consider all possible sets of genotypes that would lead to the mixture profile.

Our previous treatment (3) assumed independence of all the alleles in the mixed profile. This means independence within individuals, implying Hardy-Weinberg and linkage equilibrium, as well as independence between individuals, meaning that the contributors are unrelated. Although these assumptions may be adequate in many situations, they ignore the low-level dependence among alleles within the same population due to evolutionary forces. Two people within the same population must have common ancestors at some point in the past, the point being closer for smaller populations, and this imposes a dependence between their alleles. A necessary corollary to this evolutionary relationship is the low degree of inbreeding among offspring of two parents from the same population. It is this logic that leads to the necessity of working with conditional profile probabilities rather than the profile probabilities themselves, and it is what led to Recommendation 4.2 of the second NRC report (2). Instead of determining the probability of finding a profile in a random member of a population, it is necessary to determine the probability of finding the profile given that the profile has been seen once already. Conditional probabilities take explicit account of allelic dependencies.

In this paper we extend our previous treatment to allow for the dependencies among all the alleles carried by the contributors to the mixture. Initially we will assume that all contributors belong to the same population, as this is likely to maximize the effects we are considering. We will also adopt the relatively simple formulation for the probabilities of sets of alleles advocated by Balding and Nichols (4). Less restrictive treatments (5) would be unwieldy. Although we do not expect the population structure effects that we are considering will be substantial, we believe that they should be considered for mixed DNA stains to the same extent that they are considered for single stains.

Likelihood Ratios

Likelihood ratios have been recognized by authors of several recent books as the appropriate way of interpreting evidence (6–11). At a trial there will be alternative hypotheses or propositions about who contributed to this evidence: the prosecution will have proposition H_p, and here we will suppose there is a single alternative proposition H_d. The likelihood ratio LR is

$$\mathrm{LR} = \frac{\Pr(\text{Evidence} \mid H_p)}{\Pr(\text{Evidence} \mid H_d)}$$

The DNA evidence E for mixed-stain cases is the set of alleles found among all the people who have either been typed directly or whose type is inferred because they are considered to have contributed to the stain. Previously (3) we took E to mean only the alleles in the stain, but the addition now of the alleles from people who may have been typed even though they are hypothesized not

[1] Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh, NC.
[2] Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand.
[3] ESR, Private Bag 92021, Auckland, New Zealand.
Received 3 Feb. 1998; and in revised form 24 Aug. 1998; accepted 15 Dec. 1998.

to have contributed to the stain is necessary to allow for the effects of population structure.

We will make a distinction between the genetic profile, which is simply a listing of the distinct alleles in the mixture, and the statistical profile which is a list of all $2n$ alleles when there are n contributors. These two profiles will be different whenever some contributors are homozygous, or when some contributors share alleles. We will ignore the possibility of null alleles so that only homozygous individuals contribute a single allele to a genetic profile.

We will use much of our previous notation (3), and repeat our observation that the interpretation of a mixed stain genetic profile requires a specification of the known contributors to the profile and of the number of unknown contributors. We will derive results for single loci and then multiply likelihood ratios over loci.

As an example, suppose the evidentiary sample in a single-perpetrator rape case shows three alleles *a, b, c* at some locus. The sample was recovered from the victim's person, she was found to be of type *ab* and a suspect was found to be of type *c*. The prosecution proposition is likely to be H_p: "The victim and the suspect were the only contributors to the sample," and a likely alternative proposition is H_d: "The victim and some unknown man were the only contributors to the sample." The usual solution (2,3) for this situation is

$$\mathrm{LR} = \frac{1}{p_c(2p_a + 2p_b + p_c)} \tag{1}$$

where the p's are the allele frequencies. We now derive this result from the perspective of this paper, first with population structure ignored.

Under proposition H_p only the victim and suspect are involved and they have both been typed. The DNA evidence is therefore the genotype pair (*ab, cc*). We write the probability of this pair as $\Pr(ab, cc) = 2\Pr(abcc)$. The approach we are taking assigns probabilities to sets of alleles without regard to the arrangement of alleles among individuals, but we do need a factor of "2" for the heterozygous victim. Had the victim been *ab* and the suspect *bc* we would have required the probability $4\Pr(abbc)$ since there are then two heterozygotes. When population structure is ignored, as it was previously (3), the probability of a set of alleles is just the product of frequencies of the separate alleles, so $\Pr(abcc) = p_a p_b p_c^2$. The numerator of LR is, therefore,

$$\Pr(E \mid H_p) = 2p_a p_b p_c^2 \tag{2}$$

Note that, because the victim and suspect are both known individuals, there is no need to consider the 2! orders of these two people as was erroneously done in in the first printing of (7).

Under proposition H_d there are three people to consider: the suspect of genotype *cc* who did not contribute to the sample, and the victim of type *ab* plus the perpetrator of unknown genotype who both did contribute to the sample. Examination of the profiles of the sample and the victim shows that the unknown man must have allele *c* and may also have alleles *a, b* or *c*. There are a total of six alleles in *E*, and the probability is $\Pr(ab, cc, ac) + \Pr(ab, cc, bc) + \Pr(ab, cc, cc)$ or $4\Pr(aabccc) + 4\Pr(abbccc) + 2\Pr(abcccc)$. The denominator of LR is

$$\Pr(E \mid H_d) = 4p_a^2 p_b p_c^3 + 4p_a p_b^2 p_c^3 + 2p_a p_b p_c^4 \tag{3}$$

The factors of 2 or 4 are because of the one or two heterozygotes. Dividing Eq 2 by Eq 3 leads to the previously known result given in Eq 1.

It will be helpful to modify this example before proceeding further. Suppose the profiles are the same as just discussed, except that now the sample is not from the victim's person (e.g., it may be from discarded clothing) and the alternative to H_p is specified as H_d: "Two unknown people were the contributors to the sample." Under this proposition, there are four people involved: the victim and suspect, neither of whom contributed to the sample, and two unknown people who were the contributors. These last two people must have alleles *abc* between them but cannot have any other alleles. The possible combinations of genotypes for the unknown people are (*aa, bc*), (*ab, ac*), (*ab, bc*), (*ab, cc*), (*bb, ac*), (*ac, ab*), (*ac, bb*), (*ac, bc*), (*bc, aa*), (*bc, ab*), (*bc, ac*), and (*cc, ab*). These 12 combinations represent three distinct sets of alleles: *aabc, abbc, abcc,* and each set has a coefficient of 12 which is the number of ways of arranging the four alleles into two different genotypes. The coefficient includes the effects of the two orders of alleles within heterozygotes as well as the two orders of different genotypes such as *aa, bc* and *bc, aa.* The probabilities of all eight alleles among the four people involved are obtained by multiplying the probabilities $12\Pr(aabc)$, $12\Pr(abbc)$, $12\Pr(abcc)$ by the probability $\Pr(ab, cc) = 2\Pr(abcc)$ of the victim and suspect, and can be written as $24\Pr(aaabbccc) + 24\Pr(aabbbccc) + 24\Pr(aabbcccc)$ so that

$$\Pr(E \mid H_d) = 24p_a^3 p_b^2 p_c^3 + 24p_a^2 p_b^3 p_c^3 + 24p_a^2 p_b^2 p_c^4 \tag{4}$$

Dividing Eq 2 by Eq 4 gives the LR for this situation as

$$\mathrm{LR} = \frac{\Pr(E \mid H_p)}{\Pr(E \mid H_d)} = \frac{1}{12p_a p_b p_c(p_a + p_b + p_c)} \tag{5}$$

as has been given before (3).

We now modify the solutions in Eqs 1 and 5 to accommodate the situation where all people, the victim, the suspect and (under H_d) the unknown person(s), belong to the same subpopulation. Probabilities for the genotype(s) of the unknown person(s) must take into account the knowledge that two people in this subpopulation have been found to have genotypes *ab* and *cc*.

For both scenarios, H_p is that only the victim and suspect were the contributors to the sample. We will show that the required term $\Pr(abcc)$ is given by

$$\Pr(abcc) = \frac{[(1-\theta)p_a][(1-\theta)p_b][(1-\theta)p_c][(1-\theta)p_c + \theta]}{(1-\theta)(1)(1+\theta)(1+2\theta)}$$

where θ is the coancestry coefficient in the subpopulation to which the victim and suspect both belong.

For the denominator in the first scenario, which is that the victim and an unknown person contributed to the sample but the suspect did not, there are three people and six alleles to consider. We will show, for example, that

$$\Pr(aabccc) = \frac{\begin{array}{r}[(1-\theta)p_a][(1-\theta)p_a + \theta_a][(1-\theta)p_b] \\ \times [(1-\theta)p_c][(1-\theta)p_c + \theta][(1-\theta)p_c + 2\theta]\end{array}}{(1-\theta)(1)(1+\theta)(1+2\theta)(1+3\theta)(1+4\theta)}$$

where θ is the coancestry coefficient for the subpopulation to which all three people belong. These expressions lead to

$$\mathrm{LR} = \frac{(1+3\theta)(1+4\theta)}{[(1-\theta)p_c + 2\theta][(1-\theta)(2p_a + 2p_b + p_c) + 7\theta]} \tag{6}$$

which reduces correctly to Eq 1 when $\theta = 0$.

For the second scenario, where both contributors to the sample

are unknown under H_d, we need terms such as Pr(*aaabbccc*), and we will show that

$$\Pr(aaabbccc) = \frac{X_a X_b X_c}{Y}$$

where

$$X_a = [(1-\theta)p_a][(1-\theta)p_a + \theta][(1-\theta)p_a + 2\theta]$$
$$X_b = [(1-\theta)p_b][(1-\theta)p_b + \theta]$$
$$X_c = [(1-\theta)p_c][(1-\theta)p_c + \theta][(1-\theta)p_c + 2\theta]$$
$$Y = (1-\theta)(1)(1+\theta)(1+2\theta)(1+3\theta)(1+4\theta)(1+5\theta) \times (1+6\theta)$$

so that LR becomes

$$\mathrm{LR} = \frac{(1+3\theta)(1+4\theta)(1+5\theta)(1+6\theta)}{12[(1-\theta)p_a + \theta][(1-\theta)p_b + \theta][(1-\theta)p_c + 2\theta] \times [(1-\theta)(p_a + p_b + p_c) + 7\theta]} \quad (7)$$

and this reduces to Eq 5 when $\theta = 0$.

What is the numerical effect of using Eq 6 instead of Eq 1? When allele frequencies are all relatively small at 0.1 and θ has the relatively high value of 0.03, the LR drops from 20 to 12.33. Multiplying values from Eq 6 over several loci can give quite large LR values, but they will be less than those from Eq 1 in which population structure is ignored.

The approach we have just illustrated is as follows. Alternative propositions are needed that specify the numbers of contributors to the evidentiary sample. Some of these contributors will be known and typed people, and some will be unknown people. Those contributors, together with any typed people who are known (under the proposition) not to be contributors, contain among them a set of alleles whose probability can be written down as the product of the separate allele proportions or as a more complicated function that incorporates the population structure parameter θ. There is also a factor of 2 for each known heterozygote, and a term for the number of ways of arranging all $2x$ alleles from x unknown people into pairs. There may be different sets of alleles from unknown people under some propositions, and the probabilities for these sets must be added together. The likelihood ratio is the ratio of probabilities under alternative propositions. As additional examples, we list the results for each of the common cases described in (7) in the Appendix.

Although it is possible to follow the above line of argument for any situation, we prefer to work with a general approach amenable to automatic (computer-based) calculation as we did previously (3). This will relieve the forensic scientist of the need for lengthy calculations in the same way that computer programs such as POPSTATS can be used for other DNA calculations. We will lay out the logic behind this general approach even though we anticipate the routine use of computer packages.

In order to do this we need to break the problem into two parts; we list the alleles, with their multiplicities, carried by the unknown contributors under H_p or H_d, and then we determine the probabilities of the allele sets. The two probabilities lead to the likelihood ratio. It is our use of the theory in (4) that allows us to concentrate on alleles rather than genotypes.

Notation

Much of the complexity in dealing with mixtures can be removed by a mnemonic notation, as laid out in Table 1. We find it very helpful to label the alleles at a locus **A** by the letters A_i. There are sets of alleles (not necessarily distinct—the statistical profiles) that occur in the crime sample (C). For a particular proposition there are alleles (T) carried by typed people declared to be contributors and alleles (U) carried by unknown contributors to the sample, and there are alleles (V) carried by any people declared not to have contributed to the sample. There are corresponding sets of distinct alleles—the genetic profiles—and these sets are indicated by a g subscript. Note that the same person may be declared to be a contributor to the sample under one proposition, and declared not to be contributor under another proposition.

TABLE 1—*Notation for mixture calculations.*

Alleles in the profile of the evidence sample.	
C	The set of alleles in the evidence profile.
C_g	The set of distinct alleles in the evidence profile.
n_C	The known number of contributors to C.
h_C	The unknown number of heterozygous contributors.
c	The known number of distinct alleles in C_g.
c_i	The unknown number of copies of allele A_i in C. $1 \le c_i \le 2n_C$, $\sum_{i=1}^{c} c_i = 2n_C$
Alleles from typed people that H declares to be contributors.	
T	The set of alleles carried by the declared contributors to C.
T_g	The set of distinct alleles carried by the declared contributors.
n_T	The known number of declared contributors to C.
h_T	The known number of heterozygous declared contributors.
t	The known number of distinct alleles in T_g carried by n_T declared contributors.
t_i	The known number of copies of allele A_i in T. $0 \le t_i \le 2n_T$, $\sum_{i=1}^{c} t_i = 2n_T$.
Alleles from unknown people that H declares to be contributors.	
U	The sets of alleles carried by the unknown contributors to C.
x	The specified number of unknown contributors to C: $n_C = n_T + x$.
$c - t$	The known number of alleles that are required to be in U.
r	The known number of alleles in U that can be any allele in C_g, $r = 2x - (c - t)$.
n_x	The number of different sets of alleles U, $n_x = (c + r - 1)!/[(c-1)!r!]$.
r_i	The unknown number of copies of A_i among the r unconstrained alleles in U. $0 \le r_i \le r$, $\sum_{i=1}^{c} r_i = r$.
u_i	The unknown number of copies of A_i in U: $c_i = t_i + u_i$, $\sum_{i=1}^{c} u_i = 2x$. If A_i is in C_g but not in T_g: $u_i = r_i + 1$. If A_i is in C_g and also in T_g: $u_i = r_i$.
Alleles from typed people that H declares to be non-contributors.	
V	The set of alleles carried by typed people declared not to be contributors to C.
n_V	The known number of people declared not to be contributors to C.
h_V	The known number of heterozygous declared non-contributors.
v_i	The known number of copies of A_i in V: $\sum_i v_i = 2n_V$.

Allele Sets

The alleles in the evidence profile are carried by typed people declared to be contributors or unknown people, so that C is the combination (union) of sets T and U. For a given proposition, the probability of the evidence profile depends also on the alleles carried by people who have been typed but are declared by that proposition not to have contributed to the profile. For a proposition in which there are x unknown contributors, we write the probability as $P_x(T, U, V)$ in an extension of our previous notation (3). Note, however, that the present probability is for all the alleles in the sets T, U, V whereas the probability in (3) was for only the alleles in U conditional on those in T. In the total set of $2n_C + 2n_V = 2n_T + 2n_U + 2n_V$ alleles, we see from Table 1 that allele A_i occurs $c_i +$

$v_i = t_i + u_i + v_i$ times. We add the probabilities over all possible $n_x = (c + r - 1)!/[(c - 1)!r!]$ distinct sets of u_i. As listed in Table 1, c is the number of distinct alleles in C_g and r is the number of alleles carried by unknown people that can be any one of these c alleles.

Generating the n_x sets U is a two-stage process. Some of the alleles in each set must be present: these are the alleles in the set C_g that are not in set T_g. Other alleles are not under this constraint because they already occur in T_g, and there are r_i copies of A_i alleles in this unconstrained set. It is a straightforward computing task to let r_1 range over the integers 0, 1, . . ., r, then let r_2 range over the integers 0, 1, . . ., $r - r_1$, then let r_3 range over the integers 0, 1, . . ., $r - r_1 - r_2$, and so on. The final count r_c is obtained by subtracting the sum of $r_1, r_2, \ldots, r_{c-1}$ from r. The total number of A_i alleles in set U is $\sum_{i=1}^{c} u_i = 2x$ where $u_i = r_i$ for those alleles in both C_g and T_g, and $u_i = r_i + 1$ for alleles in C_g but not in T_g.

For any ordering of the $2x = \sum_i u_i$ alleles in U, successive pairs of alleles can be taken to represent genotypes and there are $(2x)!/(\prod_{i=1}^{c} u_i!)$ possible orderings. This is the number of possible sets of unknown genotypes that have each allelic set U. Although it is the genotypes that correspond to the x unknown people, it is the set of $2x$ alleles that we use to determine the probability, in combination with the $2n_T + 2n_V$ alleles among the known people. Because the n_T typed people all have specified genotypes, we consider not all possible orderings of the $2n_T$ alleles but just a factor of 2 for each heterozygote. Similarly, we need a factor of 2 for each heterozygote among the set of n_V non-contributors (this corrects erroneous statements in (7)).

For the single-perpetrator rape example above, now writing alleles a, b, c as A_1, A_2, A_3, the evidence sample set is $C_g = (A_1, A_2, A_3)$ and $c = 3$. Under H_d (the victim and one unknown person contributed to the mixed stain) the set from known people is $T = (A_1, A_2)$ and $n_T = 1$, $t = 2$. The set from the unknown person must contain A_3 since $c - t = 1$, $x = 1$, $r = 1$, but can also contain any of the three alleles in set C_g: i.e. there are $n_x = 3$ different sets of alleles from the unknown person. We also considered the situation where H_d is that the evidence stain was from two unknown people, $x = 2$ and no known contributors, $n_T = t = 0$. Now U must contain all three alleles A_1, A_2, A_3, $c - t = 3$, and the $r = 1$ other allele can be any of these three. There are $n_x = 3$ different sets U. The counts of alleles A_1, A_2, A_3 in these sets are, therefore, (2,1,1), (1,2,1), (1,1,2) and each of these can be ordered in $4!/(2!1!1!) = 12$ ways.

Allele Dependencies

We now consider how to attach probabilities to the sets of alleles discussed in the last section. We suppose that a state of evolutionary equilibrium has been established, so that the probabilities of sets of alleles can be found from the Dirichlet distribution (13). This distribution depends on allele proportions and the coancestry coefficient. The statement that the relationship between pairs of alleles in a subpopulation can be quantified by the coancestry coefficient θ has several interpretations (12). Here we will take it to mean that the probability that two alleles taken at random from the subpopulation are both of type A_i is $p_i^2 + \theta p_i(1 - p_i)$, where p_i is the allele frequency of A_i *averaged over subpopulations.* When allele frequencies over populations follow the Dirichlet distribution, the probability of a set of frequencies $\{p_i\}$ for alleles A_i is given by

$$\Pr(\{p_i\}) = \frac{\Gamma(\gamma.)}{\prod_i \Gamma(\gamma_i)} \prod_i (p_i)^{\gamma_i - 1}$$

where

$$\gamma_i = (1 - \theta)p_i/\theta,\ \gamma. = \sum_i \gamma_i = (1 - \theta)/\theta$$

and Γ is the gamma function with the property $\Gamma(x + 1) = x\Gamma(x)$. The great advantage of this Dirichlet distribution is that it allows the probability of any set of alleles to be found very simply. If the set has m_i copies of A_i, then the probability is

$$\Pr\left(\prod_i A_i^{m_i}\right) = \frac{\Gamma(\gamma.)}{\Gamma(m. + \gamma.)} \prod_i \frac{\Gamma(m_i + \gamma_i)}{\Gamma(\gamma_i)} \qquad (8)$$

where $m. = \sum_i m_i$. This is the result upon which Eqs 4.10 in the 1996 NRC report (2) are based (4).

In our mixed-stain situation, there are $t_i + u_i + v_i$ copies of allele A_i, and the required probability is

$$P_x(T, U, V) = \sum_{\substack{0 \le r_i \le r \\ \sum_{i=1}^{c} r_i = r}} \frac{(2x)!2^{h_T + h_V}}{\prod_{i=1}^{c} u_i!} \times \frac{\Gamma(\gamma.)}{\Gamma(\gamma. + 2x + 2n_T + 2n_V)} \prod_{i=1}^{c} \frac{\Gamma(\gamma_i + t_i + u_i + v_i)}{\Gamma(\gamma_i)} \qquad (9)$$

Summing over the $\{r_i\}$ values accounts for all n_x sets U.

Although this is a very compact expression, implementing it in a computer program is easier after some expansion. From the properties of the gamma function $\Gamma(\cdot)$ and the definition of γ_i

$$\frac{\Gamma(\gamma.)}{\Gamma(\gamma. + 2x + 2n_T + 2n_V)} = \frac{\theta^{2x + 2n_T + 2n_V}}{\prod_{j=0}^{2x + 2n_T + 2n_V - 1} [(1 - \theta) + j\theta]}$$

$$\frac{\Gamma(\gamma_i + t_i + u_i + v_i)}{\Gamma(\gamma_i)} = \frac{\prod_{j=0}^{t_i + u_i + v_i - 1} [(1 - \theta)p_i + j\theta]}{\theta^{t_i + u_i + v_i}}$$

We can also make the summation over $\{r_i\}$ values more explicit by showing the range of values of each r_i. Equation 9 becomes

$$P_x(T, U, V) = \sum_{r_1=0}^{r} \sum_{r_2=0}^{r - r_1} \cdots \sum_{r_{c-1}=0}^{r - r_1 - \ldots - r_{c-2}} \frac{(2x)!2^{h_T + h_V}}{\prod_{i=1}^{c} u_i!} \times \frac{\prod_{i=1}^{c} \prod_{j=0}^{t_i + u_i + v_i - 1} [(1 - \theta)p_i + j\theta]}{\prod_{j=0}^{2x + 2n_T + 2n_V - 1} [(1 - \theta) + j\theta]} \qquad (10)$$

Likelihood ratios are formed as the ratios of two such probabilities, and we note that people declared to be contributors under one proposition may be declared to be non-contributors under the other. In other words, every person typed is declared to be either a contributor or a non-contributor. The number of people typed, and the alleles they carry among them, are the same for every proposition. For this reason, $n_T + n_V$, $h_T + h_V$ and $u_i + v_i$ will be the same in the probabilities for each proposition. The term $2^{h_T + h_V}$ will cancel out of the likelihood ratio, as will some of the terms in the products in the numerator and denominator of the right hand side of Eq 10.

If population structure is ignored, and θ is set to zero, Eq 10 reduces to

$$P_x(T, U, V) = \sum_{r_1=0}^{r} \sum_{r_2=0}^{r - r_1} \cdots \sum_{r_{c-1}=0}^{r - r_1 - \ldots - r_{c-2}} \frac{(2x)!2^{h_T + h_V}}{\prod_{i=1}^{c} u_i!} \prod_{i=1}^{c} p_i^{t_i + u_i + v_i}$$

This is equivalent to Eq 3 in our earlier treatment (3) and may be in a form more convenient for computation. Because of cancelation of terms in the likelihood ratio, it can be seen that n_T, n_V, h_T, h_V, t_i, v_i are not used when $\theta = 0$. In this case the value of LR depends only on the numbers and frequencies of the alleles carried by unknown contributors. There is no need to consider the genotypes of typed people, whether or not they contribute to the evidence sample. This is different to the situation where population structure is taken into account—then the genotypes of all typed people are needed.

In the degenerate case where there are no typed people, contributors or non-contributors, $t_i = v_i = 0$, then $u_i = r_i$ and the sum for $\theta = 0$ is just a multinomial expansion:

$$P_x(U) = \left(\sum_{i=1}^{c} p_i\right)^{2x}$$

Examples

We now consider an example where the evidence sample $C_g = (A_1A_2A_3A_4)$ ($c = 4$) is known to be from two perpetrators but only one suspect, of type A_1A_2, has been apprehended. Proposition H_p is that this suspect and one unknown person were the contributors, so $T = (A_1A_2)$ ($n_T = 1$, $t = 2$) and U has only one possibility ($n_x = 1$): the two alleles A_3A_4. There are no known non-contributors, so $V = \phi$, $n_V = 0$ where ϕ denotes the empty set. The probability under H_p is

$$P_1(\{A_1A_2\}, \{A_3A_4\}, \{\phi\}) = \frac{2!2^1}{1!1!}\frac{\Gamma(\gamma.)}{\Gamma(\gamma. + 4)}\prod_{i=1}^{4}\frac{\Gamma(\gamma_i + 1)}{\Gamma(\gamma_i)}$$

$$= \frac{4(1-\theta)^3 p_1 p_2 p_3 p_4}{(1+\theta)(1+2\theta)}$$

Proposition H_d is that there are no known contributors, $T = \phi$, $n_T = 0$, there is one person known not be a contributor, $V = (A_1A_2)$, $n_V = 1$, and there are two unknown contributors who must carry all four alleles between them. Once again, there is only one possible set $U = (A_1A_2A_3A_4)$, $n_x = 1$ and the probability is

$$P_2(\{\phi\}, \{A_1A_2A_3A_4\}, \{A_1A_2\}) = \frac{4!2^1}{1!1!1!1!}\frac{\Gamma(\gamma.)}{\Gamma(\gamma. + 6)}\prod_{i=1}^{2}\frac{\Gamma(\gamma_i + 2)}{\Gamma(\gamma_i)}\prod_{i=3}^{4}\frac{\Gamma(\gamma_i + 1)}{\Gamma(\gamma_i)}$$

$$= \frac{48(1-\theta)^3 p_1 p_2 p_3 p_4[(1-\theta)p_1 + \theta][(1-\theta)p_2 + \theta]}{(1+\theta)(1+2\theta)(1+3\theta)(1+4\theta)}$$

The likelihood ratio for this example is, therefore,

$$\mathrm{LR} = \frac{(1+3\theta)(1+4\theta)}{12[(1-\theta)p_1 + \theta][(1-\theta)p_2 + \theta]}$$

which reduces to $1/(12p_1p_2)$ when $\theta = 0$ as has been given previously (1,3).

A more complicated example is for a rape committed by three men. Suppose that the evidence sample has alleles (A_1, A_2, A_3, A_4), the victim is of type A_1A_2 and a single suspect has type A_3A_3. Then two alternative propositions are; H_p: "The victim, the suspect and two unknown men contributed to the sample," and H_d: "The victim and three unknown men contributed to the sample."

The evidence genetic profile has $c = 4$ alleles $C_g = (A_1, A_2, A_3, A_4)$. Under proposition H_p there are $t = 3$ distinct alleles $T_g = (A_1, A_2, A_3)$ from two known contributors and no alleles from people known not to be contributors, $V = \phi$. For $x = 2$ unknown contributors, the number of sets of $r = 3$ alleles these people can carry in addition to the A_4 allele they must have among them is $n_2 = 6!/(3!3!) = 20$. The counts u_1, u_2, u_3, u_4 for all four alleles A_1, A_2, A_3, A_4 among the two unknown men, together with the multiplicities $[4!2^1]/[u_1!u_2!\,u_3!u_4!]$, are

0,0,0,4:2	0,0,1,3:8	0,0,2,2:12	0,0,3,1:8	0,1,0,3:8
0,1,1,2:24	0,1,2,1:24	0,2,0,2:12	0,2,1,1:24	0,3,0,1:8
1,0,0,3:8	1,0,1,2:24	1,0,2,1:24	1,1,0,2:24	1,1,1,1:48
1,2,0,1:24	2,0,0,2:12	2,0,1,1:24	2,1,0,1:24	3,0,0,1:8

Under proposition H_d there are $t = 2$ alleles, $T = (A_1, A_2)$, from a known contributor (the victim) and two alleles $V = A_3, A_3$ from a person (the suspect) known not to be a contributor. For $x = 3$ unknown contributors, the number of sets of $r = 4$ alleles these people can carry in addition to the A_3, A_4 alleles they must have among them is $n_3 = 7!/(4!3!) = 35$. The counts u_1, u_2, u_3, u_4 for A_1, A_2, A_3, A_4, with coefficients $[6!2^1]/(u_1!u_2!u_3!u_4!)$, for the 35 possible sets are:

0,0,1,5:12	0,0,2,4:30	0,0,3,3:40	0,0,4,2:30	0,0,5,1:12
0,1,1,4:60	0,1,2,3:120	0,1,3,2:120	0,1,4,1:60	0,2,1,3:120
0,2,2,2:180	0,2,3,1:120	0,3,1,2:120	0,3,2,1:12	0,4,1,1:60
1,0,1,4:60	1,0,2,3:120	1,0,3,2:120	1,0,4,1:60	1,1,1,3:240
1,1,2,2:360	1,1,3,1:240	1,2,1,2:360	1,2,2,1:360	1,3,1,1:240
2,0,1,3:120	2,0,2,2:180	2,0,3,1:120	2,1,1,2:360	2,1,2,1:360
2,2,1,1:360	3,0,1,2:120	3,0,2,1:120	3,1,1,1:240	4,0,1,1:60

For each proposition, the multiplicities are multiplied by the appropriate Dirichlet probabilities and the 20 or 35 terms added together. Obviously this is a task better suited for a computer.

Multiple Subpopulations

So far we have considered the situation where all people involved in the evidence interpretation have been in the same subpopulation. Other situations are likely, especially when victim and suspect belong to different racial groups. The same sets of alleles are involved as before, but now the probabilities need to be calculated separately for the alleles within each subpopulation.

We begin by returning to our first example of a single-perpetrator rape where the victim was of type A_1, A_2, the suspect was of type A_3A_3 and the evidence sample was $A_1A_2A_3$. If there was reason to believe that the perpetrator was of the same racial type as the suspect, but of a different type from the victim, then the victim's alleles need to be separated from those of the suspect and, under H_d, from the unknown perpetrator. Suppose that the victim belonged to racial group 1, with coancestry θ_1 for her subpopulation and allele frequencies p_1, p_2, for A_1, A_2. Suppose also that the suspect and perpetrator belong to racial group 2, with coancestry coefficient θ_2 for their subpopulation and allele frequencies q_1, q_2, q_3 for alleles A_1, A_2, A_3. Suppose, further, that there is zero coancestry between alleles in different racial groups so that alleles in groups 1 and 2 can be treated independently.

Under H_p, the probability is

$$P_0(\{A_1A_2A_3A_3\}, \{\phi\}, \{\phi\}) = 2(1-\theta_1)p_1p_2 \times q_3[(1-\theta_2)q_3 + \theta_2]$$

since the pair A_1A_2 from group 1 and the pair A_3A_3 from group 2 are treated separately. Under H_d, one of the three components of $P_1(\{A_1A_2\}, U, \{A_3A_3\})$ is

$$P_1(\{A_1A_2\}, \{A_1A_3\}, \{A_3, A_3\}) = 2(1-\theta_1)p_1p_2 \times \frac{2(1-\theta_2)q_2q_3[(1-\theta_2)q_3 + \theta_2][(1-\theta_2)q_3 + 2\theta_2]}{(1+\theta_2)(1+2\theta_2)}$$

since the pair A_1A_2 from group 1 and the two pairs A_1A_3, A_3A_3 from group 2 are treated separately. Equation 6 is replaced by

$$\mathrm{LR} = \frac{(1+\theta_2)(1+2\theta_2)}{[(1-\theta_2)q_3 + 2\theta_2][(1-\theta_2)(2q_1 + 2q_2 + q_3) + 3\theta_2]}$$

The general Eq 10 can be modified to allow for different subpopulations. However, when any of the three sets *T, U, V* contains alleles from different subpopulations, as was the case in the example just considered, it will be necessary to introduce further notation. Each of the counts t_i, u_i, v_i would need to be split into a component for each subpopulation, and the multiplicity coefficients would also need to be derived separately for each subpopulation.

Discussion

We offer this treatment of the effects of population structure on DNA mixture calculations to complement two previous treatments—the effects of population structure on single stains (2,4) and the interpretation of mixed stains without population structure (1,3). Our study therefore closes a gap in current DNA forensic interpretation.

Our treatment is based firmly on the use of likelihood ratios and the accompanying need for conditional probabilities. There is no alternative when the evidence is less than certain under the proposition H_p. Conditional probabilities are necessary to incorporate the known genetic nature of DNA profiles. The full meaning of profiles cannot be found without accounting for the role of evolution in shaping the probabilities of sets of profiles. The novel feature of this study lies in accounting for the information contained in the profiles of people who are declared not to have contributed to the evidence profile. This has arisen for the situation of a suspect, who is not excluded from the evidence profile, being declared not to be a contributor under proposition H_d.

The arguments made for incorporating non-contributors can be extended. Several people may be typed during the course of an investigation. Even if they are excluded as being contributors, they provide information for the probability calculations when they can be considered to belong to the same subpopulation as (some of) people not excluded. They make their contribution to the calculation via allelic set *V*.

Our treatment has assumed a specific number of unknown contributors, but we realize that this number is very likely not to be known. Although some general statements about conservative assumptions can be made (3), such as assuming large numbers of unknown people for loci with few alleles and small numbers of unknown people for loci with many alleles, we prefer not to formulate rules. Instead we recommend the calculation of likelihood ratios under plausible ranges of numbers, and the reporting of the more conservative results.

We have not allowed for unseen, or "null," alleles as has been done previously (2,3) because the move away from RFLP technology in forensic science has diminished the need for such a treatment. We have not considered other typing-system features such as intensity or peak height differences as these have been discussed elsewhere. However, we do consider that the approach described here is sufficiently flexible to allow the interpretation of many different mixed-stain DNA profiles.

Software for conducting the calculations described in this paper can be obtained directly from the World Wide Web page www.stat.ncsu.edu (click on "Statistical Genetics") or by sending email to weir@stat.ncsu.edu.

Acknowledgments

This work was supported in part by a postdoctoral fellowship from the New Zealand Foundation for Research in Science and Technology to JMC, and by NIH grant GM 45344 to North Carolina State University. John Storey wrote computer software to implement the calculations in the paper.

References

1. Evett IW, Buffery C, Wilcott G, Stoney D. A guide to interpreting single locus profiles of DNA mixtures in forensic cases. J Forensic Sci Soc 1991; 31:41–7.
2. National Research Council. DNA technology in forensic science. Washington, DC: National Academy Press 1992.
3. Weir BS, Triggs CM, Starling LI, Walsh KAJ, Buckleton J. Interpreting DNA mixtures. J Forensic Sci 1997;42:213–22.
4. 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;64:125–40.
5. Weir BS. The effects of inbreeding on forensic calculations. Ann Rev Genet 1994;28:597–621.
6. Aitken CGG. Statistics and the evaluation of evidence for forensic scientists. New York: Wiley 1995.
7. Evett IW, Weir BS. Interpreting DNA evidence: Statistical genetics for forensic science. Sunderland, MA; Sinauer 1998.
8. Faigman DL, Kaye DH, Saks MJ, Sanders J. Modern scientific evidence: The law and science of expert testimony. St. Paul, MN; West 1997.
9. Robertson B, Vignaux GA. Interpreting evidence: evaluating forensic science in the courtroom. Chichester, UK; Wiley 1995.
10. Royall R. Statistical evidence: A likelihood paradigm. London; Chapman and Hall 1997.
11. Schum DA. Evidential foundations of probabilistic reasoning. New York; Wiley 1994.
12. Weir BS. The coancestry coefficient in forensic science. Proc 8th Int Symp Human Identification. Madison, WI; Promega 1998.
13. Wright S. The genetical structure of populations. Ann Eugen 15:323–54.

Additional information and reprint requests:
Bruce S. Weir, Ph.D.
North Carolina State University
Dept of Statistics
PO Box 8203
Raleigh, NC 27695-8203

APPENDIX

In this Appendix we show the effects of population structure for each of the six common situations described in Chapter 7 of (7). A

diagram for the profiles in each case is shown in Fig. 1, and in each case setting $\theta = 0$ reduces the result to the one given in (7).

Case 1: Four-Allele Mixture, Heterozygous Victim, and Heterozygous Suspect

The victim is of type A_3A_4, the suspect is of type A_1A_2, and the crime sample of type $A_1A_2A_3A_4$. The two propositions are

H_p: The victim and the suspect contributed to the stain.
H_d: The victim and an unknown person contributed to the stain.

The evidence sample is $C = C_g = (A_1A_2A_3A_4)$ and $c = 4$.

Under H_p, the alleles from known contributors are $T = T_g = A_1A_2A_3A_4$ and $n_T = 2$, $h_T = 2$, $t = 4$. There are no alleles from unknown contributors or from people declared not to be contributors, so $n_V = h_V = 0$.

Under H_d, the alleles from known contributors are $T = T_g = A_3A_4$ and $n_T = 1$, $h_T = 1$, $t = 2$. The alleles from unknown contributors are constrained to be $U = A_1A_2$ and $x = 1$, $r = 0$. The alleles from people declared not to be contributors are $V = A_1A_2$ and $n_V = 1$, $h_V = 1$.

The required probabilities are

$$H_p\text{: } P_0(\{A_1A_2A_3A_4\}, \phi, \phi) = \frac{2^2(1-\theta)^4 p_1p_2p_3p_4}{(1-\theta)(1+\theta)(1+2\theta)}$$

$$H_d\text{: } P_1(\{A_3A_4\}, \{A_1A_2\}, \{A_1A_2\}) = \frac{2^2 2!(1-\theta)^4 p_1p_2p_3p_4[(1-\theta)p_1+\theta][(1-\theta)p_2+\theta]}{1!1!(1-\theta)(1+\theta)(1+2\theta)(1+3\theta)(1+4\theta)}$$

and the likelihood ratio is

$$\text{LR} = \frac{(1+3\theta)(1+4\theta)}{2[(1-\theta)p_1+\theta][(1-\theta)p_2+\theta]}$$

Case 2: Three-Allele Mixture, Homozygous Victim, and Heterozygous Suspect

The victim is of type A_3, the suspect is of type A_1A_2, and the crime sample of type $A_1A_2A_3$. The two propositions are

H_p: The victim and the suspect contributed to the stain.
H_d: The victim and an unknown person contributed to the stain.

The evidence sample is $C = (A_1A_2A_3A_3)$, so $C_g = (A_1A_2A_3)$ and $c = 3$.

Under H_p, the alleles from known contributors are $T_g = A_1A_2A_3$ and $n_T = 2$, $h_T = 1$, $t = 3$. There are no alleles from unknown contributors or from people declared not to be contributors, so $n_V = h_V = 0$.

Under H_d, the allele from known contributors is $T_g = A_3$ and $n_T = 1$, $h_T = 0$, $t = 1$. The alleles from the unknown contributor are constrained to include A_1A_2, and $x = 1$, $r = 0$. The alleles from the person declared not to be a contributor are $V = A_1A_2$ and $n_V = 1$, $h_V = 1$.

The required probabilities are

$$H_p\text{: } P_0(\{A_1A_2A_3A_3\}, \phi, \phi) = \frac{2^1(1-\theta)^3 p_1p_2p_3[(1-\theta)p_3+\theta]}{(1-\theta)(1+\theta)(1+2\theta)}$$

$$H_d\text{: } P_1(\{A_3A_3\}, \{A_1A_2\}, \{A_1A_2\}) = \frac{2^1 2!(1-\theta)^3 p_1p_2p_3[(1-\theta)p_1+\theta] \times [(1-\theta)p_2+\theta][(1-\theta)p_3+\theta]}{1!1!(1-\theta)(1+\theta)(1+2\theta)(1+3\theta)(1+4\theta)}$$

and the likelihood ratio is

$$\text{LR} = \frac{(1+3\theta)(1+4\theta)}{2[(1-\theta)p_1+\theta][(1-\theta)p_2+\theta]}$$

as it was for Case 1.

Case 3: Three-Allele Mixture, Heterozygous Victim, and Homozygous Suspect

The victim is of type A_2A_3, the suspect is of type A_1, and the crime sample of type $A_1A_2A_3$. The two propositions are

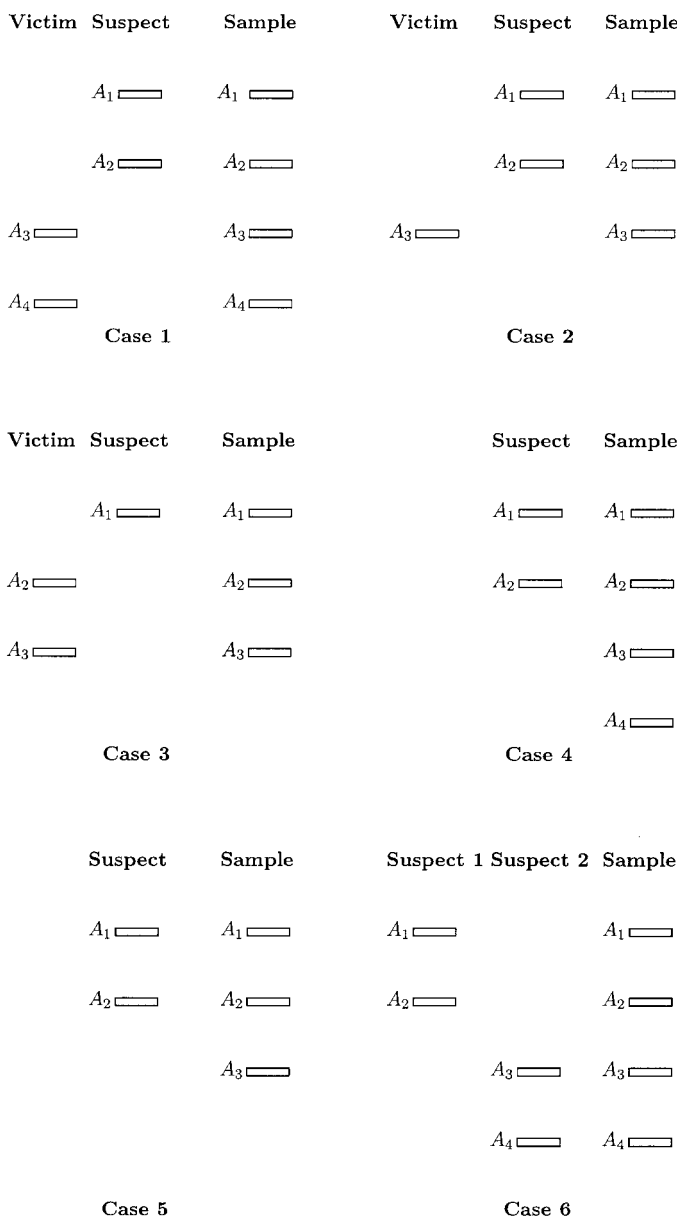

FIG. 1

H_p: The victim and the suspect contributed to the stain.
H_d: The victim and an unknown person contributed to the stain.

The evidence sample is $C = (A_1A_1A_2A_3)$, so $C_g = (A_1A_2A_3)$ and $c = 3$.

Under H_p, the alleles from known contributors are $T_g = A_1A_2A_3$ and $n_T = 2$, $h_T = 1$, $t = 3$. There are no alleles from unknown contributors or from people declared not to be contributors, so $n_V = h_V = 0$.

Under H_d, the alleles from known contributors are $T_g = A_2A_3$ and $n_T = 1$, $h_T = 1$, $t = 2$. The alleles from the unknown contributor are constrained to include A_1, and $x = 1$, $r = 1$. The unknown contributor may also carry alleles A_1, A_2 or A_3. The alleles from the person declared not to be a contributor are $V = A_1$, so $n_V = 1$, $h_V = 0$.

The required probabilities are

$$H_p\text{: } P_0(\{A_1A_1A_2A_3\}, \phi, \phi) = \frac{2^1(1-\theta)^3 p_1p_2p_3[(1-\theta)p_1 + \theta]}{(1-\theta)(1+\theta)(1+2\theta)}$$

$$H_d\text{: } P_1(\{A_2A_3\}, \{A_1A_?\}, \{A_1A_1\})$$
$$= \frac{2^12!(1-\theta)^3 p_1p_2p_3[(1-\theta)p_1 + \theta] \times [(1-\theta)p_1 + 2\theta][(1-\theta)p_1 + 3\theta]}{2!(1-\theta)(1+\theta)(1+2\theta)(1+3\theta)(1+4\theta)}$$
$$+ \frac{2^12!(1-\theta)^3 p_1p_2p_3[(1-\theta)p_1 + \theta][(1-\theta)p_1 + 2\theta][(1-\theta)p_2 + \theta]}{1!1!(1-\theta)(1+\theta)(1+2\theta)(1+3\theta)(1+4\theta)}$$
$$+ \frac{2^12!(1-\theta)^3 p_1p_2p_3[(1-\theta)p_1 + \theta] \times [(1-\theta)p_1 + 2\theta][(1-\theta)p_3 + \theta]}{1!1!(1-\theta)(1+\theta)(1+2\theta)(1+3\theta)(1+4\theta)}$$

and the likelihood ratio is

$$\text{LR} = \frac{(1+3\theta)(1+4\theta)}{[(1-\theta)p_1 + 2\theta][(1-\theta)(p_1 + 2p_2 + 2p_3) + 7\theta]}$$

Case 4: Four-Allele Mixture, Heterozygous Suspect, and One Unknown

The suspect is of type A_1A_2, and the crime sample of type $A_1A_2A_3A_4$. The two propositions are

H_p: The suspect and an unknown person contributed to the stain.
H_d: Two unknown people contributed to the stain.

The evidence sample is $C = C_g = (A_1A_2A_3A_4)$ and $c = 4$.

Under H_p, the alleles from known contributors are $T = T_g = A_1A_2$ and $n_T = 1$, $h_T = 1$, $t = 2$. There are two alleles A_3A_4 from unknown contributors, but no alleles from people declared not to be contributors, so $n_V = h_V = 0$.

Under H_d, there are no alleles from known contributors are $T = T_g = \phi$ and $n_T = 0$, $h_T = 0$, $t = 0$. The alleles from unknown contributors are constrained to be $U = A_1A_2A_3A_4$ and $x = 2$, $r = 0$. The alleles from people declared not to be contributors are $V = A_1A_2$ and $n_V = 1$, $h_V = 1$.

The required probabilities are

$$H_p\text{: } P_1(\{A_1A_2\}, \{A_3A_4\}, \phi) = \frac{2^12!(1-\theta)^4 p_1p_2p_3p_4}{1!1!(1-\theta)(1+\theta)(1+2\theta)}$$

$$H_d\text{: } P_2(\phi, \{A_1A_2A_3A_4\}, \{A_1A_2\})$$
$$= \frac{2^14!(1-\theta)^4 p_1p_2p_3p_4[(1-\theta)p_1 + \theta][(1-\theta)p_2 + \theta]}{1!1!1!1!(1-\theta)(1+\theta)(1+2\theta)(1+3\theta)(1+4\theta)}$$

and the likelihood ratio is

$$\text{LR} = \frac{(1+3\theta)(1+4\theta)}{12[(1-\theta)p_1 + \theta][(1-\theta)p_2 + \theta]}$$

Case 5: Three-Allele Mixture, Heterozygous Suspect, and One Unknown

The suspect is of type A_1A_2, and the crime sample of type $A_1A_2A_3$. The two propositions are

H_p: The suspect and one unknown person contributed to the stain.
H_d: Two unknown people contributed to the stain.

The evidence sample is $C = (A_1A_2A_2A_3A_3)$, so $C_g = (A_1A_2A_3)$ and $c = 3$.

Under H_p, the alleles from known contributors are $T_g = A_1A_2$ and $n_T = 1$, $h_T = 1$, $t = 2$. The alleles from unknown contributors are constrained to include A_3 and may also include A_1, A_2 or A_3. There are no alleles from people declared not to be contributors, so $n_V = h_V = 0$.

Under H_d, there are no alleles from known contributors, so $n_T = 0$, $h_T = 0$, $t = 0$. The alleles from the unknown contributor are constrained to include A_1, and $x = 1$, $r = 1$. The unknown contributor may also carry alleles A_1, A_2 or A_3. The alleles from the person declared not to be a contributor are $V = A_1A_2$ and $n_V = 1$, $h_V = 1$.

The required probabilities are

$$H_p\text{: } P_1(\{A_1A_2\}, \{A_3A_?\}, \phi)$$
$$= \frac{2^12!(1-\theta)^3 p_1p_2p_3[(1-\theta)p_1 + \theta]}{1!1!(1-\theta)(1+\theta)(1+2\theta)}$$
$$+ \frac{2^12!(1-\theta)^3 p_1p_2p_3[(1-\theta)p_2 + \theta]}{1!1!(1-\theta)(1+\theta)(1+2\theta)}$$
$$+ \frac{2^12!(1-\theta)^3 p_1p_2p_3[(1-\theta)p_3 + \theta]}{2!(1-\theta)(1+\theta)(1+2\theta)}$$

$$H_d\text{: } P_2(\phi, \{A_1A_2A_3?\}, \{A_1A_2\})$$
$$= \frac{2^14!(1-\theta)^3 p_1p_2p_3[(1-\theta)p_1 + \theta] \times [(1-\theta)p_1 + 2\theta][(1-\theta)p_2 + \theta]}{2!1!1!(1-\theta)(1+\theta)(1+2\theta)(1+3\theta)(1+4\theta)}$$
$$+ \frac{2^14!(1-\theta)^3 p_1p_2p_3[(1-\theta)p_1 + \theta] \times [(1-\theta)p_2 + 2\theta][(1-\theta)p_2 + 2\theta]}{1!2!1!(1-\theta)(1+\theta)(1+2\theta)(1+3\theta)(1+4\theta)}$$
$$+ \frac{2^14!(1-\theta)^3 p_1p_2p_3[(1-\theta)p_1 + \theta] \times [(1-\theta)p_2 + 2\theta][(1-\theta)p_3 + \theta]}{1!1!2!(1-\theta)(1+\theta)(1+2\theta)(1+3\theta)(1+4\theta)}$$

and the likelihood ratio is

$$\text{LR} = \frac{(1+3\theta)(1+4\theta)[(1-\theta)(2p_1 + 2p_2 + p_3) + 5\theta]}{12[(1-\theta)p_1 + \theta][(1-\theta)p_2 + \theta] \times [(1-\theta)(p_1 + p_2 + p_3) + 5\theta]}$$

Case 6: Four-Allele Mixture, Two Heterozygous Suspects

The suspects are of type A_1A_2 and A_3A_4, and the crime sample is of type $A_1A_2A_3A_4$. The two propositions may be

H_p: The two suspects contributed to the stain.
H_d: Two unknown people contributed to the stain.

The evidence sample is $C = C_g = (A_1A_2A_3A_4)$ and $c = 4$.

Under H_p, the alleles from known contributors are $T = T_g = A_1A_2A_3A_4$ and $n_T = 2$, $h_T = 2$, $t = 4$. There are no alleles from unknown contributors or from people declared not to be contributors, so $n_V = h_V = 0$.

Under H_d, there are no alleles from known contributors, $T = T_g = \phi$ and $n_T = h_T = t = 0$. The alleles from unknown contributors are constrained to be $U = A_1A_2A_3A_4$ and $x = 2$, $r = 0$. The alleles from people declared not to be contributors are $V = A_1A_2A_3A_4$ and $n_V = 2$, $h_V = 2$.

The required probabilities are

$$H_p\text{: } P_0(\{A_1A_2A_3A_4\}, \phi, \phi) = \frac{2^2(1-\theta)^4 p_1p_2p_3p_4}{(1-\theta)(1+\theta)(1+2\theta)}$$

$$H_d\text{: } P_2(\phi, \{A_1A_2A_3A_4\}, \{A_1A_2A_3A_4\}) = Q$$

where

$$Q = \frac{\begin{array}{r}2^2 4!(1-\theta)^4 p_1p_2p_3p_4[(1-\theta)p_1+\theta][(1-\theta)p_2+\theta] \\ \times[(1-\theta)p_3+\theta][(1-\theta)p_4+\theta]\end{array}}{\begin{array}{r}1!1!1!1!(1-\theta)(1+\theta)(1+2\theta) \\ \times(1+3\theta)(1+4\theta)(1+5\theta)(1+6\theta)\end{array}}$$

and the likelihood ratio is

$$\mathrm{LR} = \frac{(1+3\theta)(1+4\theta)(1+5\theta)(1+6\theta)}{\begin{array}{r}24[(1-\theta)p_1+\theta][(1-\theta)p_2+\theta] \\ \times[(1-\theta)p_3+\theta][(1-\theta)p_4+\theta]\end{array}}$$