Incorporating Uncertainty into Likelihood Ratios for DNA Evidence

Bruce Weir, University of Washington
John Buckleton, ESR New Zealand
James Curran, University of Auckland

Supported in part by NIJ award 2011-DN-BX-K541

Simple Forensic Scenario

Biological evidence collected at a crime scene has genetic profile G_C.

A suspect has been identified and then found to have a matching profile G_S.

US forensic scientists have preferred to consider the probability $\Pr(G_C)$ that a person selected randomly from a population has the crime-scene type. This has been referred to as the match probability but, more properly, it is the *profile probability.*

Likelihood Ratio Approach

A better approach is to formulate alternative hypotheses for the matching profiles $E : G_C, G_S$

H_p : The suspect is the source of the evidence.
H_d : The suspect is not the source of the evidence.

and then form the likelihood ratio

$$\text{LR} = \frac{\Pr(E|H_p}{\Pr(E|H_d)}$$

Likelihood Ratio Approach

The likelihood ratio can be manipulated to

$$\mathsf{LR} \;=\; \frac{\mathsf{Pr}(G_C|G_S, H_p)}{\mathsf{Pr}(G_C|G_S, H_d)}$$

The numerator is generally set to 1 when $G_C = G_S$ or set to 0 if $G_C \neq G_S$ (but see later ...) so that

$$\mathsf{LR} \;=\; \frac{1}{\mathsf{Pr}(G_C|G_S, H_d)}$$

and $\mathsf{Pr}(G_C|G_S, H_d)$ is the *Match Probability*, the chance an unknown person has the crime-scene profile given that the suspect has the profile.

Simple Example

Suppose $G_C = G_S = ab$, a heterozygote at one genetic marker. The conventional expression for the profile probability assumes Hardy-Weinberg equilibrium: $\Pr(G_C) = 2p_a p_b$.

A more general expression allows for population structure, or uncertainty in the allelic frequencies p_a, p_b. Having observed $G_S = ab$ then provides information about $\Pr(G_C)$:

$$\Pr(G_C|G_S) = \frac{2[\theta + (1-\theta)p_a][\theta + (1-\theta)p_b]}{(1+\theta)(1+2\theta)} \geq 2p_a p_b$$

Other expressions allow the unknown source to be, for example, the brother of the suspect. Then

$$\Pr(G_C|G_S) = \frac{1}{4}(1 + p_a + p_b + 2p_a p_b) \geq \frac{1}{4}$$

Uncertainty in LR

The LR values rest on sample allele frequencies $\tilde{p}_a, \tilde{p}_b$ and so have some sampling uncertainty. "How large is your frequency database?" is a question that a lawyer may ask.

A more perceptive question might be "Are your estimates of p_a, p_b based on a sample from the relevant population?" The answer should, almost always, be "No."

A sample of size n from not exactly the relevant population has variance

$$\begin{aligned} \mathrm{Var}(\tilde{p}_a) &= p_a(1-p_a)\left(\theta + \frac{1-\theta}{n}\right) \\ \mathrm{Cov}(\tilde{p}_a, \tilde{p}_b) &= -p_a p_b \left(\theta + \frac{1-\theta}{n}\right) \end{aligned}$$

Uncertainty in LR

Assuming independence over several genetic markers l, the LR_l values for each locus can be multiplied together. Adding the $\ln(\mathsf{LR}_l)$ values lets us invoke normality and provide confidence intervals on the log scale:

$$\ln(\mathsf{LR}) \ \pm \ z_{\alpha/2}\sqrt{\mathsf{Var}(\ln(\mathsf{LR})}$$

and on the original scale

$$\frac{1}{C}\mathsf{LR} \ , \ C \times \mathsf{LR}$$

where

$$C \ = \ e^{[z_{\alpha/2}\sqrt{\mathsf{Var}(\ln(\mathsf{LR})}]}$$

These C values are generally in the range $10 \sim 1,000$.

Uncertainty in LR

These ideas were extended to evidence with multiple contributors:

Beecham, G.W. and B.S. Weir. 2011. Confidence intervals for DNA evidence likelihood ratios. Journal of Forensic Sciences 56 Supplement 1:S166–S171.

Uncertainty in profiles

As technology has progressed, DNA profiles are being obtained from smaller amounts of material and alleles "drop out" or "drop in." Both phenomena can be demonstrated in the laboratory and rates of occurrence estimated from well-designed studies.

The consequence is that there no longer needs to be a match between the profiles obtained from the crime-scene stain and the suspect. It is no longer that case that $\Pr(G_C|G_S, H_p) = 1$ or $\Pr(G_C|G_S, H_p) = 0$.

The following development is from:
Gill P, Gusmão L, Haned H, Mayr WR, Morling N, Parson W, Prieto L, Prinz M, Schneider H, Schneider PM, Weir BS. 2012. DNA commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods. Forensic Science International: Genetics (in press).

Typing thresholds

When the DNA quantity is sufficient to generate peaks above the (arbitrary) stochastic threshold (maybe 150 rfu) and the two alleles are a balanced heterozygote, a match between the donor and the crime stain is usually seen. As the template DNA level decreases, the signal level decreases and the heterozygote balance deteriorates. This occurs because of stochastic or random effects that have previously been well characterized. Allele drop-out is an extreme example of heterozygote imbalance, where one allele falls below the limit of detection threshold (typically 50 rfu). The inevitable consequence of allele drop-out is that a partial profile is generated. This means that the crime-stain DNA profile may not match the DNA profile of the hypothesized contributor.

Simple Example with Drop-out

Suppose the crime-scene sample can be assumed to be from a single contributor and has profile $G_C = a$ at a single locus and the signal is below the 150 rfu threshold. A suspect has been identified and has profile $G_S = ab$.

Under H_p : the suspect is the source of the sample, the b allele must have dropped out of the crime-scene profile and the a allele must not have dropped out. If D is the drop-out probability, assumed the same for all alleles,

$$\Pr(G_C|G_S, H_p) \;=\; (1-D)D$$

Simple Example with Drop-out

Under H_d : the suspect is not the source of the sample, the unknown donor of the sample may have been homozygous of type aa and it is not the case that both alleles dropped out. Alternatively, the donor may have been of type ax, where x is any allele other than a. The a allele did not drop out and the x allele dropped out. If D_2 is the probability that two alleles of the same type drop out, then

$$\Pr(G_C|G_S, H_d) \;=\; (1-D_2)p_a^2 + 2p_a(1-p_a)(1-D)D$$

The likelihood ratio is

$$\text{LR} \;=\; \frac{(1-D)D}{(1-D_2)p_a^2 + 2p_a(1-p_a)(1-D)D}$$

Allelic drop-in

If a crime-scene sample is contaminated with DNA from another donor, additional alleles are likely to be seen at all or most of the loci. If only one or two alleles, not from an hypothesized donor, are seen among all loci then drop-in may have occurred. The drop-in probability is written as C.

In the previous example of $G_C = a, G_S = ab$, then there must not have been any drop-in under H_p:

$$\Pr(G_C|G_S, H_p) \;=\; (1-D)D(1-C)$$

There must also not have been drop-in under H_d when the donor was aa or ax but now the donor may also have been xx, both of which dropped out and a dropped in or the donor may have been xy, both of which dropped out and a dropped in:

$$\begin{aligned}\Pr(G_C|G_S, H_d) \;=\; & (1-D_2)p_a^2(1-C) + 2p_a(1-p_a)(1-D)D(1-C) \\ & + p_x^2 D_2 C p_a + 2p_x p_y D^2 C p_a\end{aligned}$$

Not quite as simple an example

Now suppose the crime-scene sample can be assumed to be from a single contributor and has profile $G_C = ac$ at a single locus and the signal is below the 150 rfu threshold. A suspect has been identified and has profile $G_S = ab$.

Under H_p : the suspect is the source of the sample, the b allele must have dropped out of the crime-scene profile, the a allele must not have dropped out and the c allele must have dropped in:

$$\Pr(G_C|G_S, H_p) \;=\; (1-D)DCp_c$$

Not quite as simple an example

There are several possibilities for the genotype of the unknown donor under H_d:

aa: not the case that both dropped out and c dropped in.
cc: not the case that both dropped out and a dropped in.
ac: neither allele dropped out, no allele dropped in.
ax: a did not drop out, x did drop out, c dropped in.
cx: c did not drop out, x did drop out, a dropped in.
xx: both alleles dropped out, both a and c dropped in.
xy: both alleles dropped out, both a and c dropped in.

$$\begin{aligned}\Pr(G_C|G_S,H_p) &= p_a^2(1-D_2)Cp_c + p_c^2(1-D_2)Cp_a + 2p_ap_c(1-D)^2(1-C) \\ &\quad + 2p_ap_x(1-D)DCp_c + 2p_cp_x(1-D)DCp_a \\ &\quad + p_x^2D_2C^2p_ap_c + 2p_xp_yD^2C^2p_ap_c\end{aligned}$$