Match Likelihood Ratio for Uncertain Genotypes

Mark W. Perlin[*,1], PhD, MD, PhD
Joseph B. Kadane[2], PhD
Robin W. Cotton[3], PhD

[1]Cybergenetics, Pittsburgh, PA, USA
[2]Statistics Department, Carnegie Mellon University, Pittsburgh, PA
[3]Biomedical Forensic Sciences, Boston University School of Medicine, Boston, MA

19 July 2009

Presented at the Seventh International Conference on Forensic Inference and Statistics, Lausanne, Switzerland, 21 August 2008.

Submitted for publication

* *Corresponding author contact information:*

Dr. Mark W. Perlin
Cybergenetics
160 North Craig Street, Suite 210
Pittsburgh, PA 15213 USA
412.683.3004
412.683.3005 FAX
perlin@cybgen.com

Abstract

Genetic data are not necessarily fully informative, leading to uncertainty in an inferred genotype. The posterior genotype probability distribution incorporates the identification information present in the data. To compare uncertain genotypes, we introduce here a match likelihood ratio (MLR), a simple generalization of the likelihood ratio standardly used to understand the import of genetic evidence in forensic applications. The MLR gives the relative probability of a match between questioned evidence and a suspect, with respect to a match between the evidence and a relevant population. Coancestry can be naturally incorporated. We present illustrative examples, and provide a detailed analysis and comparison for a two person DNA mixture. We describe MLR's computation efficiencies when making multiple genotype comparisons, and show how MLR was used to explain evidence in court. As statistical computing of forensic DNA inferences becomes more commonplace, the MLR may help in quantifying match identification information.

Keywords: likelihood ratio, DNA, evidence, match, genotype, probability, coancestry

Table of Contents

Introduction

DNA is a powerful identification methodology that helps solve crimes (Butler 2005). Biological specimens are collected as crime scene evidence. Scientists generate short tandem repeat (STR) laboratory data (Edwards, Civitello et al. 1991) from these specimens, and then infer DNA genotypes from the observed STR data peaks. By comparing these evidence genotypes against the genotypes of possible suspects (including DNA databases), matches between genotypes can help identify individuals who contributed their DNA to the evidence.

When the genetic data are not fully informative, there is uncertainty in the inferred genotype. Computer systems have been developed that infer uncertain genotypes from kinship (Hilden 1970; Heuch and Li 1972) and DNA mixture (Mortera, Dawid et al. 2003; Perlin 2003; Bill, Gill et al. 2005; Wang, Xue et al. 2006) data. These systems may produce multiple genotype values at a locus, and assign probabilities or likelihoods to the possible values.

When comparing an uncertain DNA mixture genotype with a reference, many approaches use only the most likely genotype values. Some just select a single genotype value having a maximum likelihood for reference comparison (Perlin and Szabady 2001; Wang, Xue et al. 2006; Cowell, Lauritzen et al. 2007). Other approaches select a subset of highly likely genotype values (SWGDAM 2000; Bill, Gill et al. 2005; Curran 2008). An alternative DNA mixture approach is to use a likelihood ratio (LR) of a joint likelihood for the mixture data relative to the marginal likelihoods (Evett, Gill et al. 1998), integrating over all feasible genotype values. A LR method (Good 1950; Roeder 1994) is attractive because it uses effectively the match information present in the data (Gill, Brenner et al. 2006).

Some computer systems use Bayesian inference to obtain a posterior genotype probability mass function (pmf) that summarizes the data identification information (Perlin 2003; Curran 2008). The posterior genotype pmf generally contains more information than a single maximum likelihood genotype value, or a subset of highly likely values (O'Hagan and Forster 2004). A LR formulation based on the posterior genotype would not need to revisit the original data. This paper introduces a *match likelihood ratio* (MLR) that directly compares uncertain genotypes in order to provide a match rarity statistic.

We begin by describing the uncertain genotypes that will be used in the MLR. We then show how genotype match can be represented as a probability of genotype equality, and how to compute that probability, touching briefly on coancestry considerations. We next define the match likelihood ratio as a ratio of two genotype match probabilities, and show that the MLR is indeed a LR. We briefly describe some illustrative examples to show how MLR works. We then analyze one DNA mixture example in some detail, computing a MLR for a published posterior distribution genotype, and comparing that result with a less informative genotype. We consider how MLR accelerates LR computation with investigative DNA databases. We also report on how we used MLR to help explain DNA mixture evidence in court. We conclude with a discussion of how the MLR can be used in forensic practice.

Uncertain Genotypes

A simple, uncomplicated single source DNA reference sample usually yields an unambiguous genotype at each genetic locus. However, crime scene evidence is often more complex. Evidence can have low levels of DNA, damaged DNA or contain mixtures of DNA from multiple contributors. The result is that there may be considerable uncertainty in the genotypes inferred from the data. The genetic uncertainty at a locus can be described by associating a probability with each genotype value. The genotype's probability distribution characterizes the information learned from examining the DNA data using a particular interpretation method.

Starting from questioned sample data d_Q, a forensic scientist can apply a DNA interpretation method and infer a questioned genotype Q (Table 1). A population allele data set d_R can be used to infer a relevant population genotype R that gives relative frequencies of genotype occurrence. From a possible suspect's data d_S, a known suspect genotype S can be inferred. We assume that all of these data are independent of each other.

An individual's genotype at a genetic locus has some allele pair value. Let X be a fixed finite set of all such allele pair possibilities. When there is uncertainty in an inferred genotype, the genotypes Q, R, and S become random variables on set X. These genotypes have respective pmfs $q(x)$, $r(x)$ and $s(x)$, where x is a genotype value in X (Table 1).

Uncertain genotypes regularly appear in forensic DNA practice, although their description may be intertwined with some specific match statistic. For example, the probability of inclusion statistic[1] (SWGDAM 2000; Budowle, Onorato et al. 2009) can be viewed as a method of inferring genotype Q from DNA mixture data as a list of allele pairs, and then matching its uniform genotype pmf with a unique suspect genotype S (see "Illustrative Examples," number 3). Observed genotype frequencies represent a sample from a relevant population described in random variable R. The relevant population may be an ethnic subgroup (National Research Council 1996), or one that does not contain the suspect (Balding 2005). Genotype R may also describe a distribution for some other alternative hypothesis.

Genotype Match

A match event between the two genotypes Q and S occurs when Q=S. We are interested in the probability of this match event, or $\Pr\{Q=S\}$. We observe that the probability of two genotypes sharing a common value is the sum of joint probability events over all the disjoint genotype values x in the value set X

$$\Pr\{Q=S\} = \sum_{x\in X} \Pr\{Q=x \,\&\, S=x\}$$

Each joint probability term can be factored using conditional probability to form the sum of products

$$\Pr\{Q=S\} = \sum_{x\in X} \Pr\{Q=x \mid S=x\} \cdot \Pr\{S=x\}$$

[1] This inclusion statistic goes by several names, including "Combined Probability of Inclusion" (CPI) and "Random Man Not Excluded" (RMNE).

To avoid examiner bias (National Research Council 2009), we assume that an objective analyst inferred the questioned genotype Q without any knowledge of the suspect's genotype S. Therefore, the probability of the questioned genotype at any value is independent of the suspect's genotype, or, $\Pr\{Q = x \mid S = x\} = \Pr\{Q = x\}$, and so

$$\Pr\{Q = S\} = \sum_{x \in X} \Pr\{Q = x\} \cdot \Pr\{S = x\}$$

The analyst's genotype inferences for Q and S included their respective pmfs $q(x)$ and $s(x)$. Therefore, the match probability between the questioned genotype and the suspect is the sum of genotype probability products

$$\Pr\{Q = S\} = \sum_{x \in X} q(x) \cdot s(x) \tag{1}$$

By similar reasoning with genotypes Q and R, we also conclude that the match probability between the questioned genotype and some relevant population is the sum of genotype probability products

$$\Pr\{Q = R\} = \sum_{x \in X} q(x) \cdot r(x) \tag{2}$$

DNA samples having identical genotypes may share a common ancestry (Balding and Nichols 1994), and so are not necessarily independent. The *coancestry coefficient* θ is the probability that a randomly selected allele shared by two genotypes is identical by descent. We can write the theta-dependent genotype match probability between Q and S (and similarly for Q and R) as the sum of products

$$\sum_{x \in X} q(x) \cdot s(x) \cdot \mu_{QS}(\theta, x) \tag{3}$$

We define the *coancestry measure* $\mu_{QS}(\theta, x)$ as the ratio of the joint posterior genotype probability to the product of the marginal posterior genotype probabilities. We can calculate this measure by rearranging the posterior probabilities with Bayes Theorem (Feller 1968), and then substituting in the standard Dirichlet representation of population allele frequencies (Evett and Weir 1998).

Match Likelihood Ratio

We define the match likelihood ratio (MLR) as the probability of a match between genotypes Q and S relative to the probability of a match between Q and R.

(4) $$MLR = \frac{\Pr\{Q = S\}}{\Pr\{Q = R\}}$$

This match rarity statistic can be reported by reading from formula (4) the statement:

> "a match between the genotypes of the evidence and the suspect is (some number) times more probable than a match between those of the evidence and a random person,"

or, more colloquially,

> "a match between the evidence and the suspect is (some number) times more likely than a match between the evidence and a random person."

We need to show that the MLR is actually a LR. Bayes Theorem tells us that the evidence information can be summarized in a LR (Lindley 2006). The LR compares the probability of the evidence (E), conditioned on a hypothesis (H) and background knowledge (K), to the evidence probability conditioned on the negation of H (~H), along with K. The genotypes

Q, R and S, along with their pmfs, provide the background information K. By ignoring the prior odds ratio, the LR focuses on how well the hypothesis explains the evidence (Aitken and Taroni 2004).

The symbols E, H and K denote propositions, where "a proposition is defined to be a statement where it is meaningful to assert that it is true or that it is false" (Good 1950, p. 1). Forensic DNA interpretation has customarily used the symbol E to denote an evidence proposition about some function of the observed STR data peaks (Evett and Weir 1998), thereby forming a data likelihood ratio (DLR). However, a valid LR construction (Good 1950, ch. 6) need not adhere to this particular DLR convention. Since our approach compares genotypes (whose inference has already summarized the STR data), it is more natural in this situation to have the evidence proposition E describe a (true or false) match event between two genotypes.

The evidence proposition E that we are concerned with here is an observed event Q=U that there is a match between the inferred questioned genotype Q and a genotype U belonging to an unknown person. The standard prosecutor's hypothesis H is that the unknown person is the suspect, hence genotype U is the genotype S. The alternative hypothesis ~H (e.g., propounded by the defense) is that the unknown is not the suspect, but instead some other person randomly selected from a relevant population, and so genotype U is the population genotype R. Therefore the standard LR based on this genotype match evidence compares the alternative hypotheses as

$$\begin{aligned} LR &= \frac{\Pr\{Q=U \mid H, K\}}{\Pr\{Q=U \mid \sim H, K\}} \\ &= \frac{\Pr\{Q=U \mid U \ is \ S\}}{\Pr\{Q=U \mid U \ is \ R\}} \end{aligned}$$

After substituting in the appropriate conditioned genotypes, we obtain the MLR defined in (4)

$$= \frac{\Pr\{Q = S\}}{\Pr\{Q = R\}} = MLR$$

which establishes that the MLR is the standard forensic identity LR.

Combining the MLR probability ratio of (4), together with the sum of product formulas (1) and (2), we obtain the MLR sum of products ratio evaluation form

$$MLR = \frac{\sum_{x \in X} q(x) \cdot s(x)}{\sum_{x \in X} q(x) \cdot r(x)} \tag{5}$$

Accounting for coancestry using (3) would give a theta-adjusted MLR as

$$MLR(\theta) = \frac{\sum_{x \in X} q(x) \cdot s(x) \cdot \mu_{QS}(x, \theta)}{\sum_{x \in X} q(x) \cdot r(x) \cdot \mu_{QR}(x, \theta)}$$

Illustrative Examples

We illustrate the application of the MLR by using the sum of probability products ratio (5) to compute some useful DNA match statistics.

1. Single source. When the questioned evidence and the suspect both come from clean single source DNA, their respective independent genotypes Q and S each have a unique allele pair value. Therefore pmfs $q(x)$ and $s(x)$ both have a probability of 1 at their respective genotype values, and are equal to 0 at all other values. When Q and S agree on the same genotype value

x_0, the MLR numerator is 1 and the denominator reduces to the population genotype frequency $r(x_0)$. The match LR statistic (5) thus reproduces the usual random match probability $\frac{1}{r(x_0)}$.

2. Kinship. Suppose that the questioned evidence has a unique genotype Q, but that the suspect genotype S is inferred genetically from his or her mother and father. Then S has a probability distribution $s(x)$, with possible Mendelian values of 1/4, 1/2 or 1 at each locus. (More informative (and commensurately complex) S genotype pmfs would also incorporate the likelihood of offspring (Sisson 2007), but these are not introduced here.) Unique genotype Q can be compared with this uncertain genotype S to find a match probability $\Pr\{Q=S\}$. The computed MLR from equation (5) normalizes this factor by a match probability $\Pr\{Q=R\}$ between Q and the relevant population genotype R to determine the match rarity.

3. Mixtures. Now suppose that suspect genotype S is unique, but that questioned genotype Q is from a DNA mixture with n visible allele peaks. There are $N=\frac{n(n+1)}{2}$ unordered pairings of the n alleles. Thus one might infer a list of these N possible genotype values x, assigning each one the same uniform probability $q(x)=\frac{1}{N}$, and all other values probability zero. Suppose that there is a matching suspect genotype S (having a known allele pair) at one of these genotype values. With allele frequencies p_1, p_2, …, p_n, the genotype R has probability p_i^2 for homozygote values, and $2p_ip_j$ for heterozygote values. By substituting the genotype pmfs $q(x)$, $r(x)$ and $s(x)$ into the MLR expression (5), we derive the standard probability of inclusion match statistic (CPI, RMNE) that many workers use to interpret DNA mixtures (SWGDAM 2000).

$$\begin{aligned} MLR &= \frac{\sum\limits_{x\in X} q(x)\cdot s(x)}{\sum\limits_{x\in X} q(x)\cdot r(x)} \\ &= \frac{\frac{1}{N}\cdot 1}{\frac{1}{N}\cdot\left(p_1{}^2 + 2p_1p_2 + \ldots + p_n{}^2\right)} \\ &= \frac{1}{\left(p_1 + p_2 + \ldots + p_n\right)^2} \end{aligned}$$

4. Missing persons. We can compare an evidence genotype Q (mixture example 3) with an inferred Mendelian reference genotype S (kinship example 2). Substituting their genotype pmfs, along with the pmf of relevant population genotype R, into the MLR equation (5) determines match rarity as a LR. This approach is useful in mass disasters, where damaged DNA remains produce an uncertain questioned genotype Q (Perlin 2007) that can be compared with a missing person genotype S reconstructed from family genotypes (Heuch and Li 1972).

Note that for the single source (example 1) and the inclusion probability distributions appearing in simple kinship and DNA mixture analysis (examples 2 and 3), the MLR sum of products ratio reduces to a familiar reciprocal of a sum of population genotype probabilities. This symmetrical form may not occur with uncertain genotype comparisons (example 4) that have unequal genotype probabilities. With such nonuniform genotype pmfs, the MLR is calculated using formula (5).

Mixture Example

It would be instructive to see how MLR is used with a posterior genotype pmf for a two person DNA mixture, and then compare the genotype match information with that obtained using a maximization approach. In this mixture example, questioned major contributor genotype Q was inferred in a MCMC computation from a hierarchical Bayesian model (Curran 2008, Figure 3, blue bars) using quantitative STR peak height mixture data (Wang, Xue et al. 2006, Table 10). We constructed a relevant population genotype R using a standard Caucasian allele frequency database (Budowle, Moretti et al. 1999). The suspect genotype S was known (Wang, Xue et al. 2006, Table 10).

Curran discussed the genotype ambiguity of STR locus D13S317 on this data set, so we illustrate the MLR approach on his inferred genotype Q at this locus. Proceeding from left to right (Table 2a), the first table column gives the allele pair genotype values that appear in the posterior distribution. Column $q(x)$ shows the posterior probability values of genotype Q at D13S317, column $r(x)$ shows the population probabilities of genotype R, and column $s(x)$ shows the unambiguous suspect genotype S with allele pair [11, 12]. Each term in the numerator match probability $\Pr\{Q=S\}$ appears in column $q(x)\cdot s(x)$, which sums to 0.670. The terms in the denominator match probability $\Pr\{Q=R\}$ appear in column $q(x)\cdot r(x)$; these add up to 0.165. The LR is the ratio of these two match probabilities, which equals 4.054. The weight of evidence (base 10 logarithm of the LR) information at D13S317 with genotype Q is therefore 0.608.

Alternatively, a maximizing approach can produce a genotype Q' from the list of allele pairs contained in D13S317's 99% highest posterior probability set[2]. By considering each allele pair in this unordered set to be equally likely, we form a new genotype Q' that has the uniform probabilities shown in column $q'(x)$ (Table 2b). Columns $r(x)$ and $s(x)$ are unchanged from Table 2a. Assessing the MLR of Q' relative to that of genotype Q, the numerator match probability $\Pr\{Q'=S\}$ is halved to 0.333, and the denominator match probability $\Pr\{Q'=R\}$ is the slightly smaller 0.131. The LR for uniform genotype Q' is reduced to 2.539, with a lower logarithmic information value of 0.405.

We compared the LR's and $\log_{10}$(LR)'s of inferred genotypes Q and Q' (Table 3). At every locus, the posterior distribution genotype Q (unequal inferred q(x) probabilities) has a more informative LR than the uniform genotype Q' (equal q'(x) probabilities). By the product rule (i.e., locus independence), the joint LR for Q is $10^{15.75}$, whereas the joint LR for Q' is $10^{8.70}$. The reason for this seven order of magnitude LR improvement is that the full posterior genotype pmf inferred from the quantitative data is more informative than a set of equally probable allele pairs.

[2] One might instead consider using a single maximum probability allele pair value. However, making such a definite genotype value assignment risks producing an entirely uninformative joint LR of zero when a misclassification occurs (Cowell, Lauritzen et al. 2007).

Computational Considerations

The MLR approach decomposes the LR computation into two steps (Figure 1a). The first step sums over every genotype possibility x in the context of the evidence data (and other parameters) using Bayes theorem to infer a posterior probability distribution q(x) for genotype Q. Then, MLR combines the three genotypes Q, R and S by summing over the products of their probability distributions to form the LR. The conventional DLR instead computes the LR in a single step (for both numerator and denominator) that sums over genotype values (Evett, Gill et al. 1998) (Figure 1b). That is, the MLR uses an explicit genotype representation Q that has been partially evaluated (Futamura 1971) from these data, whereas the DLR does not preserve a genotype object in an intermediate step.

There can be conceptual utility in forming and preserving the genotype Q and its probability function q(x). The genotype is a natural representation of genetic identity, since it corresponds directly to an individual's DNA type. Also, its probability distribution captures our knowledge (and uncertainty) about unknown allele pair values. Some people find it helpful to visualize genotype value combinations (e.g., DNA mixtures) and compare these patterns with the observed data. Pedagogically, we often use these genotype concepts and pictures when educating students, judges and juries. Importantly, though, explicit representation of genotypes (as random variables) can confer significant computational advantages in certain situations.

Genotype inference can be computationally expensive when using a faithful hierarchical Bayesian model that accurately accounts for quantitative STR peak data. The associated genotype MCMC summation (and integration over other variables) for DNA mixture problems typically entails hours or days of computer time (Perlin 2005; Cowell, Mortera et al. 2008; Curran 2008). With the DLR, this genotype summation cost is incurred anew every time a comparison is made with a different suspect genotype S. However, the MLR approach exacts this cost only once, since the evidence genotype Q is preserved. The inferred pmf q(x) can therefore be reused in each subsequent (virtually instantaneous sum of products) suspect comparison.

DNA databases enable "cold hit" comparisons between crime scene evidence and suspect genotypes. Providing a LR score for every scene-to-suspect match can quantify database match information. The most informative LR is obtained when modeling the original quantitative peak height evidence (Balding and Buckleton 2009), which can be preserved for the i^th^ case in an MCMC inferred genotype Q_i having Bayesian pmf $q_i(x)$. Comparing the stored scene genotype Q_i with a set of J suspect genotypes $\left\{S_j\right\}$ is a very fast computation using MLR equation (5). However, the DLR computation does not preserve any genotype Q_i, so its costly MCMC integration must be repeated J times over the same case i quantitative data, once for each suspect j. While MLR naturally supports highly informative DNA database LR determination (Perlin 2005), a redundant DLR approach would be computationally prohibitive for typical database sizes (e.g., where J is a million or more convicted offenders).

When identifying victim remains in a mass disaster, there can be uncertainty in both the victim remains genotypes $\left\{Q_i\right\}$, as well as in the missing person genotypes $\left\{S_j\right\}$. In our work on reanalyzing the World Trade Center (WTC) disaster DNA data (Perlin 2007), each of the I genotypes Q_i was typically inferred by a joint Bayesian analysis of data d_{Q_i}, comprising multiple samples from damaged remains. Similarly, a subject genotype S_j could be inferred from data d_{S_j}, comprising low level DNA or mixture personal effect samples, and kinship family references. Whereas a full DLR comparison of all victim remains data $\left\{d_{Q_i}\right\}$ with all missing person data $\left\{d_{S_j}\right\}$ would entail $I \cdot J$ (multiplicative) LR computations, our MLR approach to the WTC reanalysis involved only $I + J$ (additive) genotype inferences that were afterwards compared rapidly using MLR equation (5) to obtain the LRs. Moreover, the cost of the MCMC inference was reduced, since DLR's joint consideration of the d_{Q_i} and d_{S_j} data (e.g., by generalizing the quantitative data LR (Evett, Gill et al. 1998, eq. 5)) is a more computationally expensive integration than MLR's separate inferences of genotype Q_i from just data d_{Q_i}, and of genotype S_j from just data d_{S_j}.

Court Case

The likelihood component of a total probability model describes how well the model accounts for observed data. A more accurate likelihood function can elicit greater identification information from the data (Gill, Brenner et al. 2006), hence infer a more informative genotype pmf. In Bayesian inference (O'Hagan and Forster 2004), complete modeling of the quantitative data can infer a genotype that preserves all of the data's identification information. The MLR provides a mechanism for automatically translating this evidence genotype (probability) representation into a LR, when making a comparison with a suspect relative to a population. We recently served as scientific experts in a criminal trial[3] that highlighted several points along the information spectrum of (infinitely many possible) likelihood functions, and demonstrated how MLR can help explain match information in court.

Dentist John Yelenic was murdered in his home in Southwestern Pennsylvania. Pennsylvania State Trooper Kevin Foley, cohabitating boyfriend of the victim's estranged wife, was accused of the homicide. The primary physical evidence was a two-person DNA mixture extracted from the victim's fingernails, containing the victim (93% of total DNA) and a second minor unknown contributor (7%). Interpreting the mixture evidence using a probability of inclusion (PI) method, the original FBI laboratory reported a DNA match statistic LR_{PI} of 13 thousand, considerably less than the million to one level that juries find persuasive (Koehler 2001). The prosecution therefore retained independent outside experts (Drs. Cotton and Perlin) to perform more informative interpretations of the DNA mixture evidence.

Dr. Cotton’s obligate allele (OA) analysis listed all allele pairs at a locus that contained an evidence allele other than the victims, yielding a LR_{OA} of 23 million. Dr. Perlin conducted a quantitative modeling (QM) of the mixture data using Cybergenetics TrueAllele® computer system, finding a LR_{QM} of 189 billion. TrueAllele genotype inference uses MCMC to explore a Bayesian model (Perlin 2003) with a multivariate normal (peak height) data likelihood function (Perlin and Szabady 2001), and generally accepted hierarchical mixture weight modeling (Curran 2008) with additional variables for stutter (Perlin, Lancia et al. 1995) and relative amplification (Ng 1998) PCR artifacts. Unlike the original laboratory's PI approach, the OA and QM methods

[3] Commonwealth of Pennsylvania vs. Kevin J. Foley, Indiana County, No. 1170, Crim 2009.

both assumed that the victim contributed DNA to his own fingernail sample (observed in the data as a 93% major component). In all three methods, LR comparison was made to suspect Foley, relative to a Caucasian reference population (Budowle, Moretti et al. 1999).

The judge admitted the OA and QM methods into evidence after hearing the outside experts testify on the general acceptance of these LR approaches in the relevant scientific community of forensic inference and statistics. At the trial, the defense cross-examination questions focused on why there were three different DNA match statistics (LR_{PI}, LR_{OA}, LR_{QM}) having very different magnitudes (10^4, 10^7, 10^{11}) for the same data. The prosecution experts compared genotype patterns with peak data to educate the jury about DNA mixture interpretation. Each interpretation method used progressively increasing amounts of the data to infer a genotype pmf:

- PI did not use either the victim profile or quantitative peak heights;
- OA did use the victim profile, but not peak heights; and
- QM used both the victim profile and the quantitative data.

The experts explained that using more of the data generally produces a more informative genotype.

MLR is a natural way to translate genotype possibilities (relative to a suspect and a population) into LR match information. We explained to the jury that (the MLR formulation of) the LR is the probability of a specific match between genotypes Q and S, relative to that of a random match between Q and R. Using a spreadsheet that presented the MLR calculation (similar to Table 2), the jurors saw how multiplying, adding and dividing genotype pmfs would compute a LR at a locus. We presented the LR contribution at each locus, comparing the three interpretation methods (analogous to Table 3); this bar chart showed how more informative genotypes produced a higher LR at certain loci. We gave the plain English statement of the LR (see MLR equation (4) paragraph), which does not mention conditional probabilities and can be applied equally well to all three methods. Although Trooper Foley testified that he was innocent, the DNA fingernail evidence indicated otherwise. The jury convicted him of first-degree murder.

Discussion

We have introduced a likelihood ratio approach for inferring match strength when there are uncertain genotypes. The key idea is to form a LR that compares the probability of a specific genotype match relative to that of a nonspecific match. The MLR assesses identification hypotheses for an observed match event, which works well with posterior genotype probability distributions, and provides information that is equivalent to the usual data event DLR. The MLR preserves the data's identification information by using the entire posterior genotype probability distribution, rather than a limited subset.

Much of the power of DNA evidence comes from making "cold hit" comparisons to offender databases (Gill and Werrett 1990; Niezgoda and Brown 1995). These databases compare a set of evidence genotypes $\left\{Q_i\right\}$ with a set of likely suspect genotypes $\left\{S_j\right\}$. In particular, comparisons are made between genotypes, without any use of the underlying genetic data. The MLR supports this DNA database paradigm by working directly with (possibly uncertain) genotypes, and efficiently computing a LR weight of evidence for every reported match.

The MLR transforms an inferred questioned evidence genotype (along with suspect and population genotypes) into a single information measurement number. This summarization is useful for validating a genotype inference method (or a laboratory procedure), since the observed LR distribution can characterize the information efficacy (distribution mean) and reproducibility (within-case variance) (Perlin 2006). Similarly, the information yield of different DNA laboratory and genotype inference methods can be compared through their LR values on representative specimens. When reporting a DNA match, the MLR summarizes identification rarity, preserving all of the data information contained in the posterior genotype pmf.

To reduce examiner bias, an objective approach is to (i) first infer (and commit to) a questioned genotype from the evidence data, (ii) only afterwards make any match comparison with a suspect genotype, and (iii) then report a LR rarity statistic with respect to a relevant population (Berry 1991; Tobin and Thompson 2006). The MLR supports this inference sequence, since equation (4) can compare any questioned genotype Q with any suspect genotype S, and determine the LR with respect to a population genotype R. Indeed, the MLR is able to "match" DNA materials only after all the genotypes Q, R and S have been determined.

There are many genotype inference methods for mixtures and other complex DNA data. The genotypes that result from applying these diverse methods to the same data can produce LR match information ranging over ten orders of magnitude (Butler and Kline 2005). Statistical computing infers genotypes that tend to preserve more identification information and provide greater consistency. The MLR can accept genotype input that has been inferred using any of these methods, and preserve all of the match information contained in the genotype pmf.

The MLR accommodates ongoing scientific improvements in genotype inference. Hierarchical Bayesian modeling can be continuously refined to incorporate more aspects of the STR data process and its uncertainty (e.g., PCR stutter, relative amplification, degraded DNA, marker balance, many unknown contributors, low level DNA). Moreover, a model can combine independent DNA sample data using a joint likelihood function that multiplies together the separate likelihoods, as we routinely do in the TrueAllele system with low level DNA mixtures. Regardless of the model specification, the inferred output is a genotype pmf that can be easily compared with other genotypes using MLR.

The MLR approach has application beyond DNA evidence. The Bayesian framework of first inferring a type, and then using the type's pmf in MLR equation (5) to compute a match rarity LR, is entirely general. We have mapped this framework onto other forensic subdisciplines, such as fire debris, firearms/toolmarks, blood spatter and fingerprints. For example, using integrals in place of sums, one can derive the standard LR formula for glass evidence (Lindley 1977) as a MLR of normally distributed types.

Forensic science has been criticized for lacking a sound statistical basis for reporting matches and their rarity (National Research Council 2009). While DNA evidence has been relatively unscathed, the continuing debate over DNA mixture interpretation (Gill, Brenner et al. 2006; Budowle, Onorato et al. 2009) and low level DNA (Balding and Buckleton 2009; Budowle, Eisenberg et al. 2009) shows that DNA is not entirely immune to such challenges.

Some have proposed that it is not even possible to give other non-DNA subdisciplines a rigorous statistical basis (Budowle, Bottrell et al. 2009). The MLR framework suggests otherwise. Bayesian inference permits the probabilistic inference of forensic types, and MLR enables their comparison to ascertain match rarity.

Acknowledgements

The authors would like to thank the anonymous reviewers whose comments and questions helped improve the clarity of the manuscript. Conversations with David Kaye and Bill Thompson at the ICFIS'08 meeting helped refine the MLR plain language statement and identify how the MLR can remove examiner bias. Bill Allan extracted the posterior genotype probabilities from the bar graph in Figure 3 of Curran that we used in computing the MLR mixture example. Careful readings and suggestions for improvement by Meredith Clarke and Erin Turo were also very helpful. Matt Legler, Cara Spencer and Jessica Staab helped in preparing the manuscript.

References

Aitken, C. G. and F. Taroni (2004). Statistics and the Evaluation of Evidence for Forensic Scientists. Chicester, UK, John Wiley & Sons.

Balding, D. J. (2005). Weight-of-Evidence for Forensic DNA Profiles. New York, John Wiley & Sons.

Balding, D. J. and J. Buckleton (2009). "Interpreting low template DNA profiles." Forensic Science International: Genetics(in press).

Balding, D. J. and R. A. Nichols (1994). "DNA profile match calculation: how to allow for population stratification, relatedness, database selection and single bands." Forensic Sci Int **64**: 125-140.

Berry, D. A. (1991). "Inferences using DNA profiling in forensic identification and paternity cases." Statist. Sci. **6**(2): 175-205.

Bill, M. R., P. Gill, J. Curran, T. Clayton, R. Pinchin, M. Healy and J. Buckleton (2005). "PENDULUM - A guideline based approach to the interpretation of STR mixtures." Forensic Science International **148**(2-3): 181-189.

Budowle, B., M. Bottrell, S. Bunch, R. Fram, D. Harrison, S. Meagher, C. Oien, P. Peterson, D. Seiger, M. Smith, M. Smrz, G. Soltis and R. Stacey (2009). "A perspective on errors, bias, and interpretation in the forensic sciences and direction for continuing advancement." J Forensic Sci **54**(4): 798-809.

Budowle, B., A. Eisenberg and A. van Daal (2009). "Validity of low copy number typing and applications to forensic science." Croatian Medical Journal **50**(3): 207-17.

Budowle, B., T. Moretti, A. Baumstark, D. Defenbaugh and K. Keys (1999). "Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians." J Forensic Sci **44**(6): 1277–1286.

Budowle, B., A. J. Onorato, T. F. Callaghan, A. D. Manna, A. M. Gross, R. A. Guerrieri, J. C. Luttman and D. L. McClure (2009). "Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework." J Forensic Sci **54**(4): 810-21.

Butler, J. M. (2005). Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers. New York, Academic Press.

Butler, J. M. and M. C. Kline (2005). NIST Mixture Interpretation Interlaboratory Study 2005 (MIX05). Promega 16th International Symposium on Human Identification, Grapevine, TX.

Cowell, R., J. Mortera and S. L. Lauritzen (2008). Probabilistic modelling of pairs of two and three-person DNA mixtures (Talk). The Seventh International Conference on Forensic Inference and Statistics, Lausanne, Switzerland.

Cowell, R. G., S. L. Lauritzen and J. Mortera (2007). "A gamma bayesian network for DNA mixture analysis." Bayesian Analysis **2**(333-48).

Curran, J. (2008). "A MCMC method for resolving two person mixtures." Science & Justice **48**(4): 168-177.

Edwards, A., A. Civitello, H. Hammond and C. Caskey (1991). "DNA typing and genetic mapping with trimeric and tetrameric tandem repeats." Am. J. Hum. Genet. **49**: 746-756.

Evett, I. W., P. Gill and J. A. Lambert (1998). "Taking account of peak areas when interpreting mixed DNA profiles." J. Forensic Sci. **43**(1): 62-69.

Evett, I. W. and B. S. Weir (1998). Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists. Sunderland, MA, Sinauer Assoc.

Feller, W. (1968). An Introduction to Probability Theory and Its Applications. New York, John Wiley & Sons.

Futamura, Y. (1971). "Partial evaluation of computation process - an approach to a compiler-compiler." Comp. Sys. Cont. **2**(5): 45-50.

Gill, P., C. H. Brenner, J. S. Buckleton, A. Carracedo, M. Krawczak, W. R. Mayr, N. Morling, M. Prinz, P. M. Schneider and B. S. Weir (2006). "DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures." Forensic Science International **160**: 90-101.

Gill, P. and D. Werrett (1990). "Interpretation of DNA profiles using a computerised database." Electrophoresis **11**: 444-448.

Good, I. J. (1950). Probability and the weighing of evidence. London, Griffin.

Heuch, I. and F. Li (1972). "PEDIG - A computer program for calculation of genotype probabilities, using phenotypic information." Clinical Genetics **3**: 501-504.

Hilden, J. (1970). "GENEX - An algebraic approach to pedigree probability calculus." Clinical Genetics **1**: 319-348.

Koehler, J. J. (2001). "When are people persuaded by DNA match statistics?" Law and Human Behavior **25**(5): 493-513.

Lindley, D. V. (1977). "A problem in forensic science." Biometrika **64**(2): 207-213.

Lindley, D. V. (2006). Understanding Uncertainty. Hoboken, NJ, John Wiley & Sons.

Mortera, J., A. P. Dawid and S. L. Lauritzen (2003). "Probabilistic expert systems for DNA mixture profiling." Theoretical Population Biology **63**: 191-205.

National Research Council (1996). Evaluation of Forensic DNA Evidence: Update on Evaluating DNA Evidence. Washington, DC, National Academies Press.

National Research Council (2009). Strengthening Forensic Science in the United States: A Path Forward. Washington, DC, National Academies Press.

Ng, S.-K. (1998). Automating computational molecular genetics: solving the microsatellite genotyping problem. Computer Science, Carnegie Mellon University.

Niezgoda, S. J. and B. Brown (1995). The FBI Laboratory's COmbined DNA Index System Program. Sixth International Symposium on Human Identification.

O'Hagan, A. and J. Forster (2004). Bayesian Inference. New York, John Wiley & Sons.

Perlin, M. W. (2003). Simple reporting of complex DNA evidence: automated computer interpretation. Promega's Fourteenth International Symposium on Human Identification, Phoenix, AZ.

Perlin, M. W. (2005). Real-time DNA investigation. Promega's Sixteenth International Symposium on Human Identification, Dallas, TX.

Perlin, M. W. (2006). Scientific validation of mixture interpretation methods. Promega's Seventeenth International Symposium on Human Identification, Nashville, TN.

Perlin, M. W. (2007). Identifying human remains using TrueAllele® technology. Forensic Investigation and Management of Mass Disasters. M. I. Okoye and C. H. Wecht. Tucson, AZ, Lawyers & Judges Publishing Co**:** 31-38.

Perlin, M. W., G. Lancia and S.-K. Ng (1995). "Toward fully automated genotyping: genotyping microsatellite markers by deconvolution." Am. J. Hum. Genet. **57**(5): 1199-1210.

Perlin, M. W. and B. Szabady (2001). "Linear mixture analysis: a mathematical approach to resolving mixed DNA samples." Journal of Forensic Sciences **46**(6): 1372-1377.

Roeder, K. (1994). "DNA fingerprinting: a review of the controversy." Statist. Sci. **9**(2): 222-456.

Sisson, S. A. (2007). "Genetics: genetics and stochastic simulation do mix!" The American Statistician **61**(2): 112-119.

SWGDAM (2000). "Short Tandem Repeat (STR) interpretation guidelines (Scientific Working Group on DNA Analysis Methods)." Forensic Sci Commun (FBI) **2**(3).

Tobin, W. A. and W. C. Thompson (2006). "Evaluating and challenging forensic identification evidence." The Champion **30**(6): 12-21.

Wang, T., N. Xue and J. D. Birdwell (2006). "Least-square deconvolution: A framework for interpreting short tandem repeat mixtures." Journal of Forensic Sciences **51**(6): 1284-1297.

Tables

Table 1. For each of three genotype classes (questioned evidence, relevant population and suspect profile), the notation for its associated data, genotype and probability mass function is shown.

Table 2. **(a)** The genotype probabilities and match likelihood ratio calculation are shown for the posterior distribution genotype Q inferred by Curran at locus D13S317 that has unequal q(x) probabilities. **(b)** The MLR calculation is shown for a uniform genotype Q' inferred as a subset of D13S317 allele pair values that have equal q'(x) probabilities.

Table 3. The LR and log(LR) values are shown at every locus for inferred genotypes Q and Q'. The joint weight of evidence is the sum of the locus log(LR) values.

Table 1

	data	**genotype**	**pmf**
Questioned evidence	d_Q	Q	q(x)
Relevant population	d_R	R	r(x)
Suspect profile	d_S	S	s(x)

Table 2

(a)

		Genotype Probability Distributions			Match Probabilities	
allele pair		**Q**	**R**	**S**	**Pr(Q=S)**	**Pr(Q=R)**
x		q(x)	r(x)	s(x)	q(x)s(x)	q(x)r(x)
11	11	0.300	0.102			0.031
11	12	0.670	0.197	1	0.670	0.132
12	12	0.030	0.095			0.003
				Pr(Q=S)	0.670	
				Pr(Q=R)		0.165
				Likelihood Ratio		4.054

(b)

		Genotype Probability Distributions			Match Probabilities	
allele pair		**Q'**	**R**	**S**	**Pr(Q'=S)**	**Pr(Q'=R)**
x		q'(x)	r(x)	s(x)	q'(x)s(x)	q'(x)r(x)
11	11	0.333	0.102			0.034
11	12	0.333	0.197	1	0.333	0.066
12	12	0.333	0.095			0.032
				Pr(Q'=S)	0.333	
				Pr(Q'=R)		0.131
				Likelihood Ratio		2.539

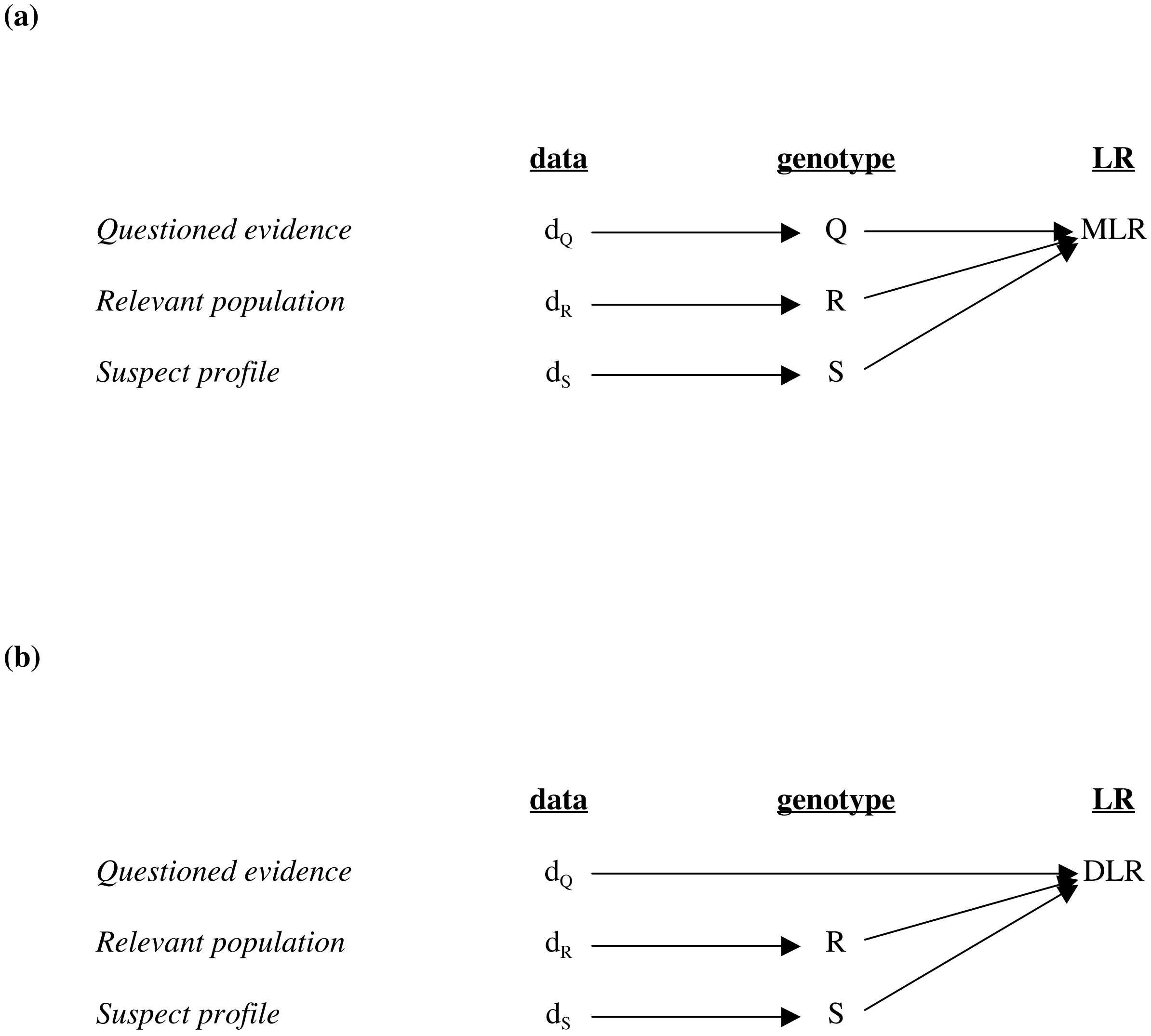
Table 3

	LR		**log(LR)**	
locus	**Q**	**Q'**	**Q**	**Q'**
CSF1PO	3.089	1.658	0.490	0.220
D13S317	4.054	2.539	0.608	0.405
D16S539	10.962	5.006	1.040	0.699
D18S51	30.640	5.310	1.486	0.725
D21S11	11.951	4.845	1.077	0.685
D3S1358	11.901	3.712	1.076	0.570
D5S818	7.668	4.000	0.885	0.602
D7S820	10.172	3.603	1.007	0.557
D8S1179	111.981	4.034	2.049	0.606
FGA	14.732	5.553	1.168	0.745
TH01	68.311	17.566	1.834	1.245
TPOX	21.507	2.984	1.333	0.475
vWA	50.046	14.552	1.699	1.163
		joint log(LR)	15.753	8.695

Figures

Figure 1. **(a)** The MLR is computed from three genotypes, each of which is inferred independently from their respective data. **(b)** The DLR is computed by summing over all genotype possibilities for the questioned evidence, and does not use a posterior genotype pmf Q. While both approaches use a likelihood function to compare genotypes with data, with MLR this comparison is done when inferring a genotype pmf, while with DLR this is done through genotype summation.

Figure 1


(a)
data
genotype
LR
Questioned evidence
Relevant population
Suspect profile
d_Q
d_R
d_S
Q
R
S
MLR
(b)
data
genotype
LR
Questioned evidence
Relevant population
Suspect profile
d_Q
d_R
d_S
R
S
DLR