



From Match Strength to Match Accuracy: Answering the Ultimate DNA Question

Mark W. Perlin, PhD, MD, PhD, William P. Allan, MS, Jennifer M. Bracamontes, MS, Matthew M. Legler

Cybergenetics, Pittsburgh, PA

Abstract

Forensic DNA interpretation has traditionally focused on match strength. After inferring an unknown genotype from evidence, we compare that genotype with a reference profile and report a likelihood ratio (LR) that measures the strength of the comparison. This approach has become a cornerstone of modern forensic science and is now widely used for complex DNA evidence.

Yet an important question remains largely unanswered. Before comparing an inferred genotype with a suspect, victim, database profile, or relative, how reliable is that genotype itself? How well can it distinguish a true contributor from a noncontributor? This presentation introduces a new framework for answering that ultimate question.

Previous research had developed match strength distributions derived from uncertain genotypes [1]. These paired distributions describe how an inferred genotype behaves across all possible match and nonmatch comparisons (**Fig. 1**). Rather than focusing on a single LR value, they characterize the overall identification performance of the genotype.

The new framework applies receiver operating characteristic (ROC) analysis in a new way to measure genotype discrimination accuracy [2]. This novel high-resolution ROC analysis combines the genotype's match and nonmatch LR distributions into a simple and intuitive probability measure of discrimination accuracy.

The area under the ROC curve (AUC) can be interpreted as the probability that a true contributor will receive stronger support than a noncontributor (**Fig. 2**). In practical terms, the AUC measures how accurately an inferred genotype can distinguish the people who contributed DNA from those who did not.

This classification perspective provides a new way to evaluate forensic DNA evidence. Instead of asking only how strongly a genotype matches a particular individual, we can ask how accurately that genotype can make match decisions in general. The answer is a single probability number that quantifies the identification capability of the evidence itself.

The approach has practical applications throughout forensic DNA interpretation. It can be used to assess genotype quality, compare analytical methods, evaluate the impact of thresholds (**Fig. 3**), measure information loss, and understand why some evidence profiles are more informative than others. It also provides a quantitative framework for validation studies and quality assessment by directly measuring the discrimination performance of inferred genotypes.

Beyond these applications, the framework suggests a broader shift in how forensic DNA evidence can be understood. Traditionally, the inferred genotype has been viewed primarily as an intermediate result used to calculate LR values. In the new framework, the genotype becomes the central scientific result. From that genotype, one can derive match-strength distributions, estimate error rates, measure discrimination accuracy, and evaluate identification performance.

The likelihood ratio evaluates a person against a genotype. ROC discrimination accuracy evaluates the genotype itself. These measures are complementary, but they answer different questions. The former quantifies the strength of a particular comparison. The latter quantifies the reliability of the underlying genotyping data and inference.

For decades, forensic DNA has focused on measuring match strength. This new work introduces a practical way to measure match accuracy. By quantifying how well an inferred genotype can distinguish contributors from noncontributors, discrimination probability provides a direct answer to a fundamental forensic question: how reliable is the DNA evidence as a tool for identification?

Heliyon Highlights

- 1. Low-LR DNA evidence can always be reported using LR distribution error rates.**
- 2. ROC analysis of a single genotype can measure its LR discrimination capability.**
- 3. Inclusionary and exclusionary LRs and error rates are mathematically symmetrical.**
- 4. Data and reporting thresholds unnecessarily reduce DNA identification information.**
- 5. Probabilistic genotyping can be rigorously validated on casework evidence items.**

Methods

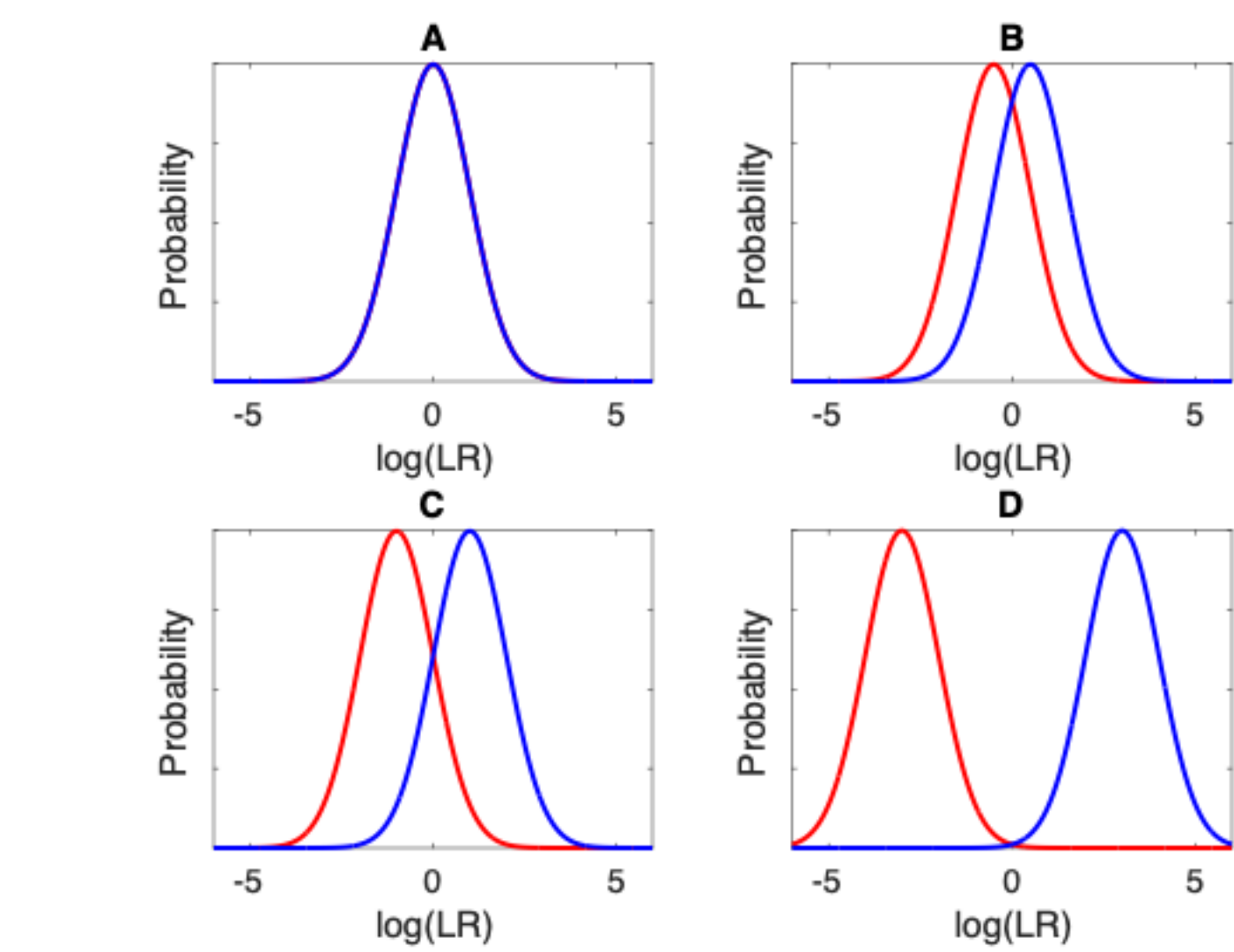


Fig. 1. An evidence genotype's *Contributor* and *Noncontributor* LR probability distributions. The paired distributions can overlap (A,B). Less overlap for better separation means more LR match discrimination (C, D).

Introduction

Forensic biology's reportable outcome is DNA match. The likelihood ratio (LR) match statistic quantifies the association (LR>1) – or non-association (LR<1) – between evidence and reference. The LR summarizes DNA evidence strength in one useful number.

But an LR statistic can sometimes point the wrong way, supporting a match even when a person's DNA isn't in the evidence. How often does such match misclassification occur? This error rate is governed by the item's *match discrimination accuracy*.

Match accuracy is an inherent property of DNA evidence data and its inferred genotypes. New science illuminates the error rates of incorrect LR scores for DNA match.

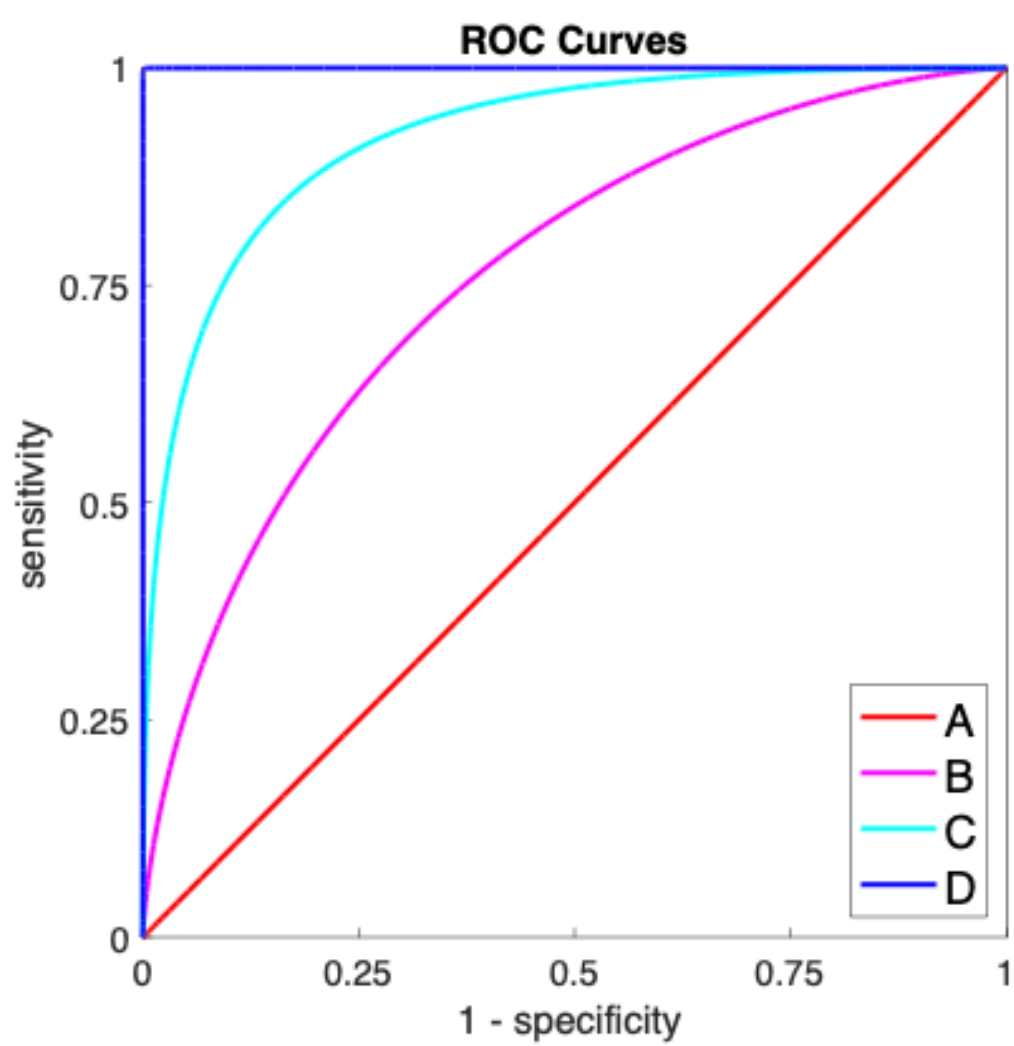


Fig. 2. ROC curves. The area under the curve (AUC) measures discrimination accuracy as the probability of correctly ordering LR pairs. A perfect LR classifier has AUC = 1 (D, blue).

Five DNA Separations

Three laboratory separations

Separation is central to forensic DNA analysis. Before alleles are developed or genotypes compared, biological evidence must first be transformed into data files of allele events. Genotypes are later derived from quantitative amounts of detected alleles.

A forensic biology lab performs three key DNA separations. **First**, they physically *extract* an item's DNA molecules from background materials. **Second**, they use polymerase chain reaction (PCR) to biochemically *amplify* DNA from just one millionth of the human genome, separating STR identification alleles.

Third, their electrophoretic DNA sequencer *separates* amplified alleles by fragment size. Fluorescent detection produces quantitative allele data, packaged as computer files for downstream genotyping analysis.

The fourth separation: contributor genotype

The **fourth** forensic DNA separation transforms allele data into genotypes, one for each evidence contributor. Simple data can produce a definite genetic profile. With complex data (degraded DNA or small amounts) genotypes have probability.

Genotype separation *splits complex DNA mixture data* into simpler genotype variables. Information-preserving methods like TrueAllele® information preserving genotyping (IPG) separate a mixture of people into corresponding single-contributor genotypes. Representing genotype possibilities with accurate probability preserves DNA identification information.

Comparing a (probabilistic) separated evidence genotype with a reference genotype, relative to all people, yields an LR match score. The score tells us how strongly the reference person contributed their DNA (or not) to the evidence. The LR logarithm measures identification information.

The fifth separation: match accuracy

An evidence genotype's Contributor LR probability distribution represents a state of the world where someone contributed their DNA to the biological evidence. The Noncontributor distribution is the alternative reality, expressed through LR, for people who didn't contribute their DNA.

If the two distribution curves were identical, the LR match statistic wouldn't distinguish contributors from noncontributors (**Fig. 1A**). With some curve separation, the LR becomes less ambiguous and more discriminating (**Fig. 1B**).

More informative DNA evidence leads to more informative genotypes that better separate the left and right LR curves, yielding greater match discrimination ability (**Fig. 1C**). With totally informative evidence there is complete curve separation, and the LR match statistic becomes a perfect classifier (**Fig. 1D**).

As introduced in our *Heliyon* paper this year [2], we can *combine the two LR distribution curves* to calculate a genotype's *classification accuracy*. Such “ROC” curve combination is widely used in fields like Artificial Intelligence and Medicine to assess the “AUC” accuracy of data analysis methods (**Fig. 2**). The inherent LR match accuracy can be derived directly from the evidence genotype, before ever comparing with a suspect.

The **fifth** forensic DNA separation is *match discrimination accuracy*, based on the separation of an evidence genotype's LR distribution curves. This match accuracy measures how well the genotype's LR match statistic discriminates between contributors and noncontributors. More curve overlap leads to more match misclassification error.

Results

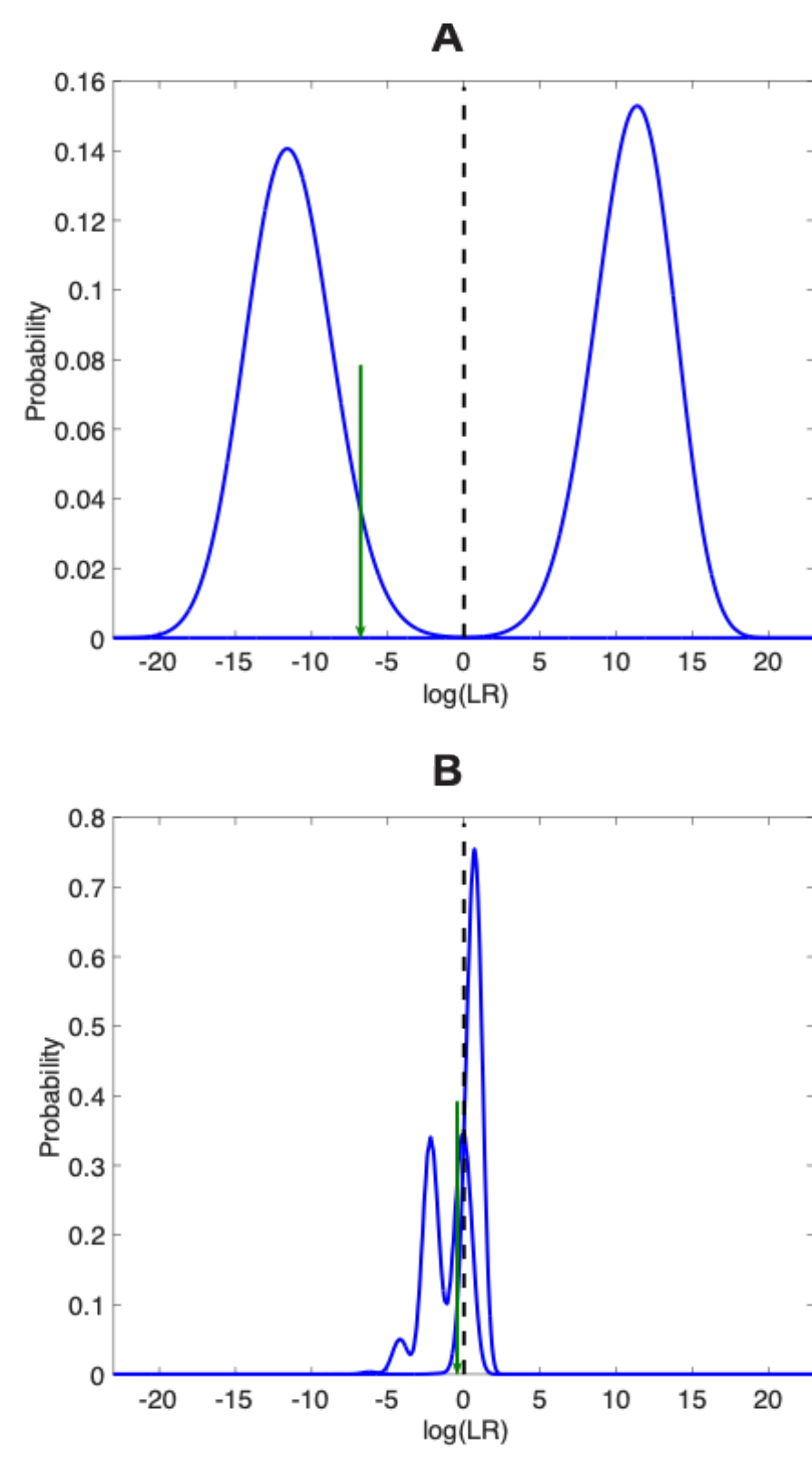


Fig. 3. US v. Sandoval LR distributions.

A. *TrueAllele*. Complete classification curve separation

B. *STRmix*. Overlapping LR distribution curves.

C. ROC curve analysis and AUC accuracy for *TrueAllele* & *STRmix*.

Heliyon 2026: \$4.9. Data thresholds discard DNA evidence

Most forensic DNA interpretation relies on an *analytic threshold* (AT) that filters out data peaks having heights below a laboratory's preset AT value. This data reduction attempts to distinguish allelic data from baseline noise or PCR artifact. The TrueAllele technology doesn't need or use thresholds. Its Bayesian model accounts for noise and artifacts, letting the software consider all the data.

United States v. Sandoval involved a two-person DNA mixture swabbed from a package containing methamphetamine. TrueAllele inferred contributor genotypes, using 210 data peaks across 21 STR loci. The log(LR) to the defendant was -6.08 ban, relative to an African-American population. The error rate logarithm was -8.35 for this exclusionary match statistic.

The DNA lab that generated the STR data applied their STRmix PG program to the same evidence item [3]. The lab's reported STRmix log(LR) match statistic to the defendant was a lower exclusionary -1.38 ban (AT = 40 RFU, 24 input peaks), relative to a Southwest Hispanic population. At different data thresholds, STRmix log(LR) values ranged 7 orders-of-magnitude from -0.53 ban (AT = 90, 11 peaks) to -7.48 ban (AT = 20, 38 peaks).

Discarding DNA data by applying peak height thresholds can lose considerable identification information, particularly with low-level mixtures. More data reduction leads to greater information loss. LR distributions [1] can help explain how data input impacts LR information. We standardize to the laboratory's choice of Southwest Hispanic population in the following LR comparisons.

Fig. 3A shows TrueAllele's threshold-free LR distributions for the exclusionary genotype. The informative Noncontributor distribution (left) has 99.99% specificity, with an average log(LR) of -11.41 . This $-KL$ is consistent with the defendant's exclusionary log(LR) = -6.78 match statistic (green arrow), distancing the defendant from the evidence. The informative Contributor distribution (right) has 99.99% sensitivity, with $KL = 11.03$. The FNR area to the left of the LR arrow gives a small error rate of log(ER) = -8.77 .

The lab applied a 40 RFU peak-height threshold to form STRmix input data. The software then separated the minor component genotype, at each locus giving a likelihood function over allele pairs. These genotype likelihood results were loaded into TrueAllele to derive genotype probability, and therefrom compute LR distributions, match statistics, and error rates.

RFU	nloci	npeaks	MW	-KL	+KL	ΔKL	AUC	log(LR)
TrueAllele Version 3.25.5840.1								
0	21	210	0.54	-11.4206	11.0281	22.4487	1.0000	-6.7828
STRmix Version 2.5.11								
10	19	54	0.09	-7.5981	4.7557	12.3538	1.0000	-3.1580
20	18	38	0.13	-4.1841	2.7046	6.8887	0.9982	-6.2917
30	16	29	0.13	-2.7232	1.3926	4.1158	0.9800	-3.2046
40	14	24	0.16	-1.1437	0.5802	1.7239	0.8971	-0.8440
50	12	21	0.15	-1.0415	0.5232	1.5647	0.8839	-0.8173
60	10	16	0.03	-0.2780	0.1460	0.4240	0.7189	-0.5886
70	9	14	0.01	-0.1510	0.0684	0.2194	0.6496	-0.5065
80	9	13	0.02	-0.1708	0.0904	0.2612	0.6736	-0.6204
90	8	11	0.03	-0.1437	0.0655	0.2092	0.6460	-0.6071

Table 1. Threshold sensitivity study showing how log(LR) information varies with RFU peak height threshold for STRmix PG computing in the case.

Fig. 3B shows the LR distributions for the STRmix genotype. The Noncontributor distribution (left) has 77.09% specificity and $-KL = -1.26$, near the recomputed log(LR) = -0.45 match statistic (green arrow). The STRmix Contributor distribution (right) has 88.17% sensitivity and $KL = 0.64$. The distribution's left tail reveals an appreciable false negative LR error rate of $ER_x = 2.53\%$. For the exclusionary LR of 0.36, only 1 in 40 people would be excluded as strongly. The 99.9% HPD credible intervals of the Contributor and Noncontributor distributions overlap, with the LR falling within their intersection, suggesting a potentially uninformative result.

Fig. 3C shows a perfect classifier TrueAllele ROC curve (AUC = 1, blue), with rectilinear edges overlying the x and y axes. A less discriminating STRmix curve (AUC = 0.9077, red) is shifted inward. The genotype-derived LR distributions, ROC curves, and associated statistics quantify the adverse impact of data thresholds on DNA identification information in this case.

Table 1 above shows how analytical thresholds can reduce DNA identification information. Threshold-free TrueAllele used 210 evidence peaks drawn from all 21 loci, producing well-separated Noncontributor ($-KL = -11.42$) and Contributor ($+KL = 11.03$) distributions, with $\Delta KL = 22.45$ and AUC = 1. As we raised the STRmix threshold from 10 to 90 RFU, the number of input loci and peaks shrank. A narrowing $[-KL, +KL]$ distribution interval gave diminishing AUCs. Typical threshold levels between 40 and 90 RFU produced small exclusionary log(LR) values between 0 and -1.

Conclusions

Fig. 3 explains the match statistic disparity. Less informative PG curves and LR values move inward, closer to zero information. The lab's distribution curves are close together. The curves overlap, reducing match classification accuracy to 90.77%. The small exclusionary log(LR) of -0.60 (green arrow) falls in the overlap region. This LR region cannot distinguish between Noncontributor and Contributor.

A 90.77% match classification accuracy means that 9.23% of the time (1 in 11) an LR comparison will point in the wrong direction. A positive log(LR) can erroneously misclassify someone whose DNA isn't in the evidence as a contributor. Or a negative match statistic can misclassify someone who really did leave their DNA as a noncontributor. A jury has 12 people. How reliable is genotype evidence whose LR comparisons misclassify a defendant 1 in 11 times?

References

- [1] Perlin MW. Efficient construction of match strength distributions for uncertain multi-locus genotypes. *Heliyon*. 2018;4(10):e00824.
- [2] Perlin MW, Allan WP, Bracamontes JM, Legler MM. The reliability and reporting of DNA match strength for uncertain genotype evidence. *Heliyon*. 2026;12:e44753.
- [3] Perlin, MW, Butt, N, and Wilson, MR. Commentary on: Thompson WC. Uncertainty in probabilistic genotyping of low template DNA: a case study comparing STRmix™ and TrueAllele®. *J Forensic Sci*. 2023;68(3):1049–63.