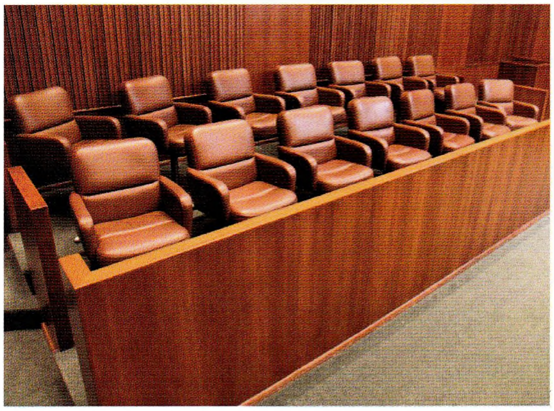
Easy Reporting of Hard DNA: Computer Comfort in the Courtroom



How can an analyst translate uncertain DNA data into understandable testimony for a non-expert jury?

Dr. Mark W. Perlin

Analyst Unplugged

A DNA analyst's life alternates between laboratory routine and courtroom unpredictability. In the lab, biological evidence is chemically separated into DNA data. Simple evidence (such as one man's blood on a knife) produces simple data, easily reported and explained in court. The greater complexity of today's low-level and mixed DNA introduces interpretation and courtroom complications, however. How can the analyst translate uncertain DNA data into understandable testimony for a non-expert jury? How well will evidence and analyst hold up under vigorous cross-examination?

The DNA match statistic is a powerful tool for taming DNA uncertainty. The statistic, or "likelihood ratio," summarizes in a single number the evidential support for a person having contributed DNA. With simple evidence from a single person, this ratio is easily explained:

- the numerator (upper half) is "1," the chance of a match if the prosecutor is correct, while
- the denominator (lower half) is the prevalence of the person's genotype in the population, i.e., the chance of a coincidental match when the prosecutor is mistaken.

Multiplication combines the statistics from all independent genetic loci. We can then read a one in quadrillion (a one followed by 15 zeros) genotype rarity as the understandable statement: "a match between evidence and suspect is a quadrillion times more probable than coincidence."

What about DNA mixtures? Alas, those are not so simple. The analyst must explain to a jury how a "threshold" classified data peaks into "alleles," but also that this threshold varies between laboratories and may give conflicting or inaccurate results. Scientific certainty devolves into subjective opinion, susceptible to cross-examination scrutiny. How can the analyst reestablish the comfort of single-source samples when testifying about mixtures in court? Without solid scientific assurance, forensic laboratory policy may discard informative crime-fighting DNA evidence as "inconclusive."

Fortunately, modern computing can help. An objective probabilistic computer can thoroughly examine the DNA mixture evidence, separating out the genotypes of each person who contributed. A separated contributor genotype can then be matched and explained with the ease and simplicity of single-source DNA. With this computer assistance, an analyst can confidently testify in court about objective scientific mixture results, without branding vital evidence "inconclusive."

Easy Reporting

Why is single-source short tandem repeat (STR) DNA so easy to interpret and explain? Because the genotype is immediately evident from the data (Figure 1). The biological specimen (e.g., a reference sample) has one true genotype at each STR locus. After the laboratory extracts, amplifies, and separates the DNA, a fluorescent data signal shows peaks that exactly correspond to this genotype. A single peak shows a homozygote allele pair (one allele repeated twice), while two peaks give a heterozygote pair of two different alleles. Data artifacts (PCR stutter, relative amplification, etc.) are minimal, so people or computers can easily infer the genotype's unique allele pair with high confidence. Sample, data, and inference may be conceptually different, but they all share the same unambiguous genotype.

The single-source genotype has an obvious match statistic, the random match probability (RMP). As mentioned, the RMP ratio is one over the population probability of a randomly selected person. Inverting a small coincidental probability gives a large match statistic, expressing high evidential force that the match observed between evidence and suspect is not coincidence. In court, this easy DNA evidence encounters little resistance, and the analyst is entirely at ease describing data, genotype, and match.

Hard DNA

With low-level DNA or mixtures, the unambiguous correspondence between data and genotype breaks down. The data may now support more than one genotype possibility. With multiple allele pair possibilities for each contributor, the data is no longer self-evident or easy to explain in court. What is an analyst to do?

Often, an analyst will examine the data and declare the mixture to be "inconclusive." There may be informative STR peaks, but some fall below their lab's interpretation threshold, so those data are not used or reported on in a match statistic. Sometimes an analyst will valiantly struggle with a complex mixture, knowing in their heart that the data are informative. But after a weeklong engagement, they reluctantly call it off as "inconclusive," losing the DNA evidence.

A common mixture interpretation method is the combined probability of inclusion (CPI). CPI treats the set of peaks above a threshold as alleles donated by contributors. The validity of CPI has been questioned, since changing the threshold gives different answers, some of which are wrong. Moreover, CPI ignores important data (e.g., a known victim genotype, or the different heights of data peaks), and so loses considerable information, often greatly understating the true match statistic. In some cases, through non-optimal data use or biased knowledge of the suspect's genotype, CPI can overstate the evidence against a defendant.

Analysts sometimes undertake more informative mixture interpretation, such as combined likelihood ratio (CLR) on a two person mixture having a known contributor reference. These methods entail writing down allele pairs that explain the data. The assumed thresholds and human decisions are harder to explain to a jury and let an opposing attorney raise reasonable questions about the analyst's testimony.

How can the analyst objectively and thoroughly interpret DNA mixtures, preserving match information? What reliable mixture computer tools offer protection from a courtroom barrage of undermining cross-examination? As we next discuss, a computer can restore the ease and comfort of single-source evidence when analysts report on DNA mixtures in court.

Computer Comfort

With DNA mixtures (Figure 2), the genotype may not be evident from the STR data.

- A biological mixture specimen is a combination of two or more contributor genotypes. Each contributor's allele pair at a locus may be definite in reality, but is unknown to us.
- We generate laboratory data that reflects these constituent genotypes. The STR signals and their peaks are not genotypes—rather, they are data derived from the underlying genotypes.
- From the observed data, we (or our computers) can infer a genotype. We list the multiple allele pair possibilities, describing our relative belief in each one by a probability.

Human review tries to extend the simplicity of a single-source situation (Figure 1) to complex mixtures (Figure 2). The hope is to ascribe to DNA data the stature of genotype by applying thresholds to envision alleles. But data are not genotypes, and so the analogy breaks down. Instead, we must continue forward from the data, inferring genotypes, and capturing our uncer-

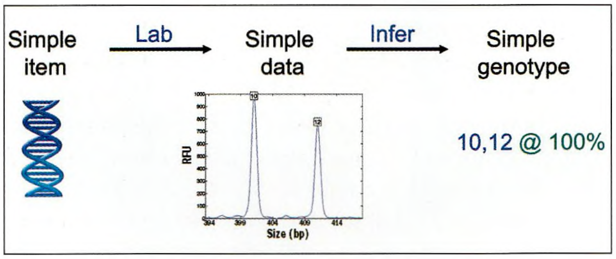


Figure 1: Single source genotype. A) In nature, the true genotype of an individual at a locus is some allele pair. B) Laboratory data casts the alleles of this genotype as peaks, where the peak's x-axis location designates the allele and the y-axis peak height indicates the DNA quantity. C) The individual's genotype is readily inferred from the data.

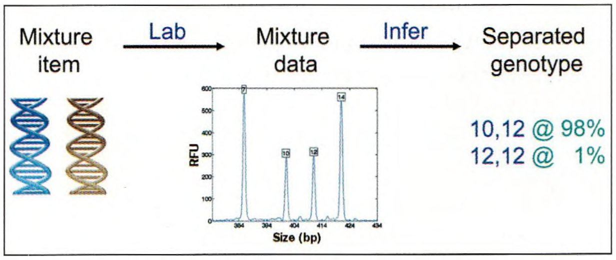


Figure 2: Mixture genotype. A) The unknown genotypes of the two (or more) contributors to the mixture are unknown to us. B) Laboratory data from this genotype generates peaks that suggest contributing alleles and their amounts. C) A contributor's genotype can be inferred from the mixture data.

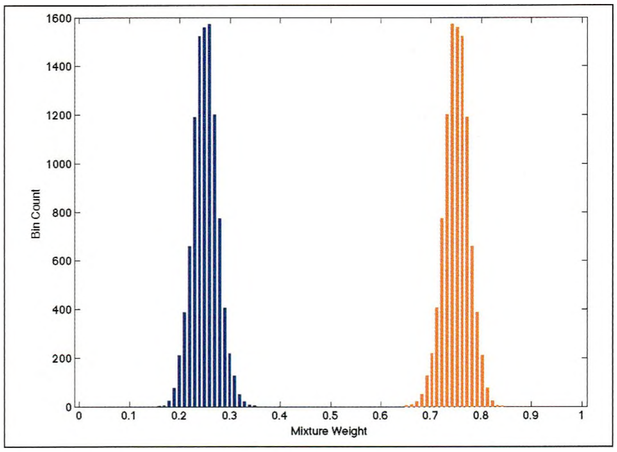


Figure 3: Mixture weight. The computer can separate a DNA mixture into its components, shown here as major (orange) and minor (blue) contributors. The center of each histogram bell curve is the average mixture weight of the contributor's DNA template, while the histogram's spread shows the mixture variation across locus experiments.

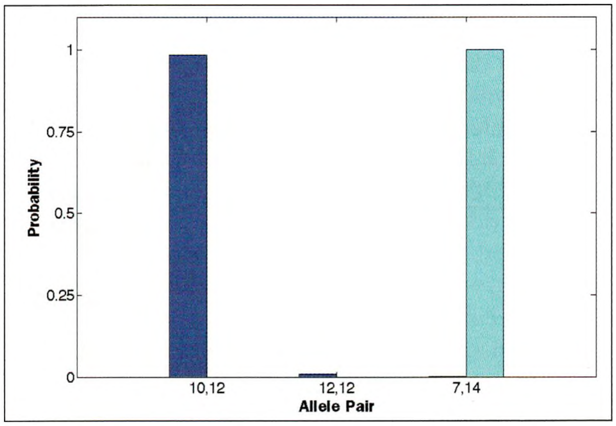


Figure 4: Separated genotypes. When the data permit, the computer can separate a mixture into essentially single source genotypes. The 3:1 ratio provides sufficient peak height differentiation between the two contributors to separate the mixture (shown at the Penta E locus) into major (light blue) and minor (dark blue) genotypes.

tainty through probability.

The computer does this mathematically by separating out mixture contributors, along with their genotypes. A two person mixture can be separated into major (Figure 3, orange) and minor (Figure 3, blue) contributors, each having a mixture weight (here 75% and 25% respectively, with a 5% uncertainty). A contributor's genotype at a locus is usually known only up to probability, but sometimes computer separation yields essentially single-source genotypes (Figure 4).

A comparison can be made between the inferred evidence (i.e., mixture contributor) genotype and a reference (e.g., suspect) genotype, relative to a population genotype. This match comparison is just like the easy RMP single-source framework, since it properly compares genotype with genotype. Importantly, all of the analyst's pictures, words, and intuition from a single contributor extend naturally to understanding and explaining DNA mixtures.

There is comfort in testifying about reliable computer mixture interpretation that has been:

1. Validated with both laboratory synthesized data and adjudicated cases
2. Published in peer-reviewed journals
3. Admitted as evidence after admissibility challenge
4. Accepted by trial and appellate courts and demonstrates proven sensitivity, specificity, and reproducibility.

The analyst can have confidence in an unbiased interpretation (no knowledge of the suspect's genotype) that thoroughly considered all feasible solutions.

Preparing for Court

A testifying analyst should be comfortable with the reliability of his or her conclusions. This comfort level can be achieved by examining replicate computer runs that were conducted on the same data. With concordant genotypes, the computer reproducibly concentrates probability onto the same allele pairs, as visualized in a series of similar looking probability bar charts (Figure 5).

The analyst prepares a DNA match report that describes the examined evidence, interpretation method, and match comparison results. A match between evidence item and reference is expressed numerically, with one statistic produced for each ethnic population considered. The match statement is written in plain language as: "a match between the evidence item and

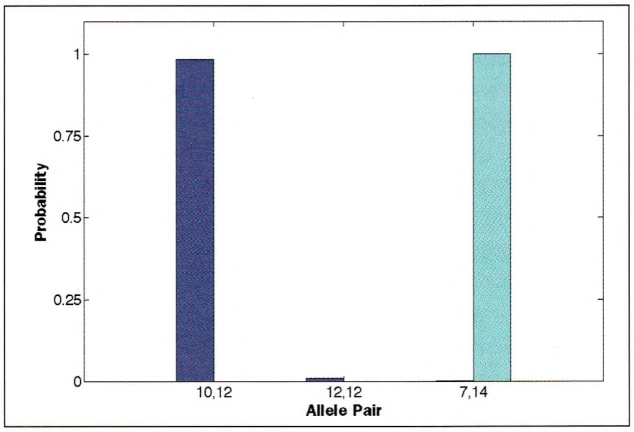


Figure 5: Concordant genotypes. Independent computer runs can establish the reproducibility of the computer's genotype determination, shown here in quadruplicate at the Penta E locus. Each independently inferred evidence genotype is shown in a different shade of blue, with probability concentrated on allele pair 10, 12.

the suspect is a quadrillion times more probable than coincidence." When there is no statistical support for a match (a statistic around 1 or smaller), the report should state this finding.

Before the trial, the testifying analyst reviews his or her report with the trial attorney, explaining the underlying science and results. When opposing counsel requests information, the analyst can meet with them and respond to their discovery request. For court, the analyst should prepare a short presentation (e.g., a PowerPoint with under ten slides) that explains the science and evidence. They should also bring a copy of the report and case folder, having on hand a summary of all computed match statistics.

Virginia v Michael Gardner

Michael Gardner, a respected lawyer in Arlington, VA, was accused in June 2011 by three young girls of having molested them when they stayed overnight at his Falls Church home for his daughter's tenth birthday celebrations. Consistent with the allegations of inappropriate touching, no semen or other intimate biological evidence was found. However, some of the girls' clothing was collected and analyzed for DNA by the state crime lab.

The inside crotch panel of one girl's underpants showed a two person DNA mixture. The analyst could not eliminate the girl or Mr. Gardner as contributors to this mixture but was unable to calculate a CPI match statistic. The lab therefore sent the STR data (.fsa files) to Cybergenetics in Pittsburgh, PA, to compute a match statistic using computer-based probabilistic genotyping.

The computer interpretation was entirely objective, having no knowledge of reference genotypes, and thoroughly considered a hundred thousand candidate solutions. The computer mathematically separated the mixture data into major and minor contributors. The separation was virtually complete, producing two definite genotypes. Each genotype had a single-source appearance, with probability heavily concentrated on just one allele pair (Figure 4). Multiple computer runs reliably inferred concordant genotypes (Figure 5).

Match comparisons were made between the inferred mixture genotypes and the available reference genotypes. As expected, a genotype (major contributor) from the underpants matched the girl who was wearing them. Because of the complete mixture separation, the 75% single-source appearing genotype gave an exact match, yielding a single-source level match statistic of 363 quadrillion. (All statistics here are computed relative to a Caucasian population, with a 1% co-ancestry theta adjustment.)

The minor contributor genotype from the underpants was compared with the references. The essentially complete mixture separation meant that comparison of this 25% genotype with any reference would either show a) a very high RMP-like single-source statistic establishing a definite match, or b) an equally strong match rejection. The computer found that a match between the underpants and Mr. Gardner was 20 quadrillion times more probable than coincidence.

Trial Testimony

The Gardner trial began on Monday, April 23, 2012. The three young girls each testified for about four hours that week, enduring long cross-examinations. The state DNA laboratory analyst presented the biological evidence. On Thursday morning, the author was sworn in to testify about the computer DNA match statistics.

In a short PowerPoint presentation, I first introduced the jury to STR genotypes and DNA evidence interpretation (Figure 2). I showed them a quantitative STR data signal (from the underpants at the Penta E locus), and explained how the computer separates DNA mixtures. Specifically, I described visually how the computer considers all possible genotype solutions, giving higher probability to proposed peak patterns that better explain the observed peak height data.

The computer had objectively inferred an evidence genotype (Figures 4 and 5), determined solely from the data, without any knowledge of the suspect's genotype. Using bar charts, I visually explained the DNA match statistic (Figure 6), comparing the probability of the evidence matching the suspect (numerator) with coincidence (denominator); this ratio was around 30 at the Penta E locus.

I showed a bar chart of the 15 independent locus match statistics (Figure 7). Multiplying these numbers

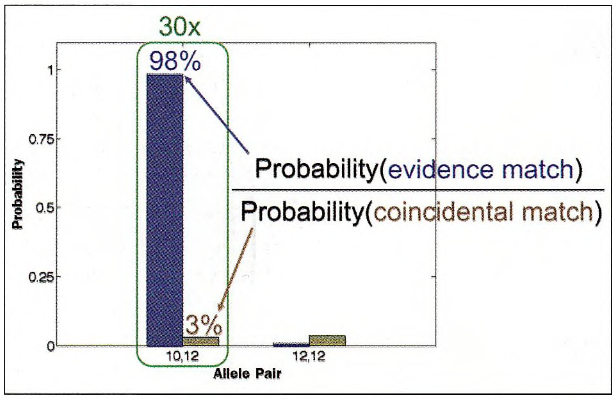


Figure 6: Locus match statistic. A DNA match statistic gives the probability of a match between the evidence and the suspect, relative to a coincidence. This ratio can be visualized at the suspect's genotype (green bar) as the relative heights of the evidence (blue bar) and population (brown bar) genotype probabilities. At the Penta E locus shown, this ratio of 98% (evidence) to 3% (population) at allele pair 10, 12 (suspect) gives a match statistic of around 30.

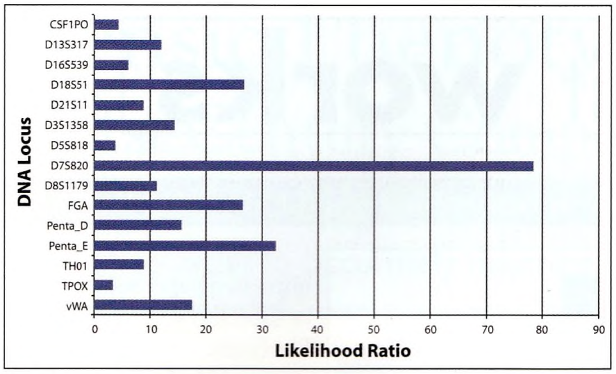


Figure 7: Match statistic. The bar chart shows the match statistics at all 15 loci. Multiplying these numbers together calculates the reported match statistic of 20 quadrillion.

together answered the question "Is the suspect in the evidence?" with the statement "A match between the underpants and Mr. Gardner is 20 quadrillion times more probable than coincidence."

The cross-examination took under an hour, never seriously challenging the reliability of the computer's DNA interpretation or its match findings. I explained why the 20 quadrillion DNA match statistic was scientifically expected.

The jury balanced Mr. Gardner's word against that of three fifth grade girls corroborated by DNA. On May 2, Michael Gardner was convicted of two counts of sexual battery and one count of object penetration. The jury sentenced him to 22 years in prison.

Analyst Reloaded

Forensic analysts currently testify about mixtures and other ambiguous DNA with justifiable trepidation. While their laboratory data are extremely reliable, the human interpretation of this data may lack rigor.

Much important DNA evidence, often crucial to a case or public safety, has been discarded by overly simplistic interpretation. Analysts often agonize over DNA mixtures, spending days wondering whether there is even a reportable match. Understating a statistic might free the guilty, while overstatement could wrongfully imprison an innocent man. Testifying can be stressful, with cross-examination questioning interpretation validity.

Interpreting DNA mixtures with a full statistical model, and hours of mathematical computing, can restore analyst confidence. Going into court with thorough and reliable match results and an understanding of computer interpretation establishes scientific comfort. The computer can separate out mixture data into component genotypes and represent uncertainty as probability, neither understating nor overstating the match statistic. A computationally empowered analyst assists the court through objectively derived fact, not subjective opinion.

***Dr. Mark Perlin** is Chief Scientific and Executive Officer for Cybergenetics. He has twenty years experience developing computer methods for information-rich interpretation of DNA evidence and providing TrueAllele products and services to the criminal justice community. Cybergenetics, 160 North Craig Street, Suite 210, Pittsburgh, PA 15213; (412) 683-3004; perlin@cybgen.com; www.cybgen.com.*