SCIENCE TECHNOLOGY INTEGRATION

The DNA Investigator™

Cybergenetics Newsletter Fall 2009

Same Data, More Information — Murder, Match and DNA

Feature Article

DNA Information and Uncertainty

The crime scene is a messy place. A world away from pristine labs, biological evidence is scraped onto swabs under real-world conditions.

A specimen may be a mixture containing DNA from several people, with damaged molecules present in only small amounts. When the crime lab amplifies the DNA evidence into viewable signals, the peaks and valleys of the signal trace may have a complex pattern without any obvious solution. A DNA analyst must discern from these signals the uncertain genetic types (or "genotypes") of the individuals who contributed to the specimen. And, afterwards, evidence and suspect genotypes can be compared to see if they match.

But how does one interpret complex casework evidence, and produce a scientific uncertain genotype? What is an "uncertain" genotype?

A genotype of a person at some chromosome location (or "locus") is a pair of alleles – inherited genetic traits – one from their mother and one from their father. When the DNA data has ambiguity, more than one allele pair might explain the observed signals.

Explanations that better fit the data have a higher probability. So an uncertain genotype is a probability distribution of allele pairs – out of the hundreds of possible values, typically only a few will have appreciable probability, while most will be unlikely. The genetic loci used to identify people look at length variation in inherited junk DNA genes that do not code for life-sustaining proteins. Instead, these short tandem repeats (STR) loci evolve relatively unconstrained by biological function. Each STR locus allele corresponds to a DNA sequence length. A forensic locus has ten or more possible allele lengths, so there are hundreds of possible allele pairings. This genotype diversity gives DNA its identification power.

The most informative scientific methods make good use of all the data. An STR data signal shows a tall peak where there is a lot of some allele length, a short peak where there is less of an allele, and usually no peak at all when the allele is not there. The location and height of STR peaks form the quantitative data used to infer genotypes.

In principle, it is easy enough to infer genotypes from quantitative DNA data. Suppose that there are two individuals in the DNA mixture – the victim and the culprit – in some combination.

Continued on Page 4

Tutorial

DNA Match Information

Page 2

Case Example

Commonwealth of Pennsylvania v. Kevin Foley

Page 3

www.cybgen.com

Testifying Tip

Presenting Match Information in Court

Today's DNA analyst goes to court with significant advantages over her predecessors. The continual laboratory validation and audits reduce the threat of effective challenge to DNA data. Population databases are in general use. DNA mixture interpretation has been ruled always admissible, though subject to cross-examination. And likelihood ratio match statistics have general acceptance in the scientific literature, law and precedent.

The match statistic is at the heart of the analyst's report and testimony. Explicitly or implicitly, the report answers a single question: what is the scientific basis for believing that the suspect is contained in the DNA evidence?

The match likelihood ratio number fully answers this question. The statistic gives the ratio of the probability of match between evidence and suspect, relative to the match probability between evidence and another person. It tells us how much more likely it is than not the suspect contributed his DNA to the evidence. And that is all the trier of fact (judge or jury) needs to know to properly weigh the scientific evidence.

So the analyst's written report is designed to support the reported DNA match statistic. It describes the data, genotype inferences, match comparisons and population databases used to form the match statistic. This legal document is the foundation of the prosecutor's direct examination, and the defense cross-examination.

On direct examination, the analyst presents her match statistic, and any necessary support for her conclusions. With multiple specimens or suspects, several match statistics can be reported.

Continued on Page 4

Tutorial

DNA Match Information

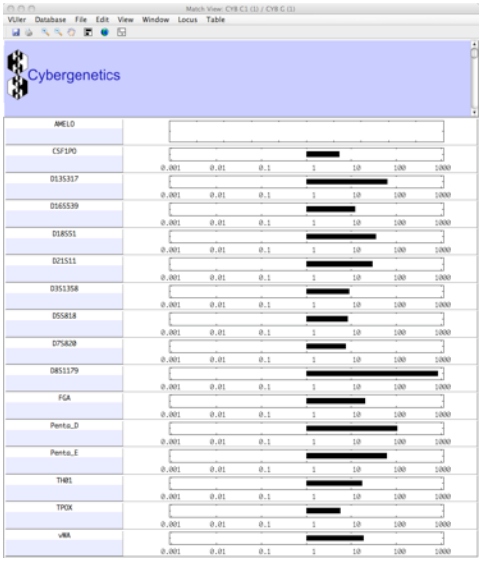
The goal of DNA identification is to elicit sufficient match information from the evidence to declare a confident match (or not) to a suspect. We express this "confidence" numerically in a match statistic. This statistic tells us how much more likely it is the suspect contributed DNA to the evidence than not.

By putting a number to this match hypothesis, an investigator or jury can decide its importance. Is that match a trillion to one? A billion, a thousand, ten? Does the match strength provide probable cause for police action? Does it prove to a jury a suspect's involvement beyond a reasonable doubt?

Only a quantitative number can properly convey DNA match information. Many courts require such a match statistic for DNA evidence to be admissible. Most crime labs routinely provide such match quantification when they report on bloodstains (RMP, CMP), sexual assault mixtures (CPI, CLR) and paternity (PI). Without a match statistic, a DNA information consumer hears only a qualitative judgment – this subjective opinion is not quantifiable, verifiable science.

A DNA identification scientist examines DNA evidence to determine the genotypes of the contributors. When compared with a suspect relative to a population, this evidence genotype completely determines the match statistic. So an informative genotype yields an informative DNA match result. This is why some DNA interpretation methods are more powerful than others – better methods produce more informative genotypes.

Science analyzes data to move us along the spectrum of uncertainty toward greater information. Statistical computers do a more in-depth analysis than people, and so generally produce a more informative match statistic.



MatchView. Presented with a mixture containing two unknown contributors, TrueAllele solved for the two genotypes. The MatchView window shows a match between the inferred 30% minor contributor genotype and a suspect, relative to a reference population. The match rarity (available from the Table menu) of this computer-inferred genotype is $10^{20.63}$, or 426 quintillion.

To solve a genetic identity problem in Cybergenetics TrueAllele® Casework system, a scientist takes a few minutes on their VUIer™ client computer to visually specify the data and method. A TrueAllele interpretation server solves for the unknown genotypes that best account for the data. The TrueAllele match server then compares these inferred evidence genotypes with suspects in the case.

The TrueAllele scientist can then visually review their match results (see the MatchView figure) on their VUIer computer, along with the original case data, inferred genotypes and mixture weights. The MatchView window shows the match rarity statistic in a bar chart at every locus (whether favorable or unfavorable). MatchView uses a logarithmic scale in order to display a wide range of match scores (e.g., from 1/1000 up to 1000), and to let the user visually add up the locus match information.

The DNA match rarity statistic (often called a "likelihood ratio") is easy to understand and present. Suppose that the match number is a trillion to one. The following English statements are equivalent:

- It is a trillion times more likely than not that the suspect's DNA is contained in the evidence.
- A match between the evidence and the suspect is a trillion times more likely than a match between the evidence and someone else.
- A match between a contributor genotype of the evidence and the suspect's genotype is a trillion times more probable than a match between the evidence genotype and a random person in a reference population.

The choice of match language is a matter of personal style, taste and laboratory protocol. Does the expert want to talk about a suspect contained in evidence, a match between samples, or a genotype match? The language for all of these approaches is given above. Every statement accomplishes the same result – numerically comparing the probability of a match between evidence and suspect with the probability of a random match.

The DNA match statistic is a single number that summarizes the import of DNA evidence. After DNA evidence has been objectively analyzed – without ever seeing the suspect (otherwise the statistic is invalid) – comparing evidence and suspect genotypes (relative to a population) determines the match information. Match information, expressed as a numerical rarity statistic, is a powerful and intuitive tool for DNA forensic scientists to communicate their findings to police, lawyers and everyone else.

More information is provided in our peer-reviewed article on matching uncertain genotypes, found at www.cybgen.com/information/publications.

Case Example

Commonwealth of Pennsylvania v. Kevin Foley

At the 2009 Pennsylvania trial of State Trooper Kevin Foley for the murder of his girlfriend's husband, three different match statistics were presented for the same DNA evidence. Each prosecution expert used a different interpretation method, with the match strengths ranging from 10 thousand to 100 billion.

How can the same DNA mixture evidence – fingernail scrapings containing 93% of the victim and 7% of the unknown second contributor – lead to such apparently disparate results? Why was this match statistic variation so readily admitted into evidence, yet so persuasive at the trial? We explore these questions in this case example.

The first match result for the pivotal DNA evidence was the FBI's 13 thousand inclusion statistic. The defense publicly derided this DNA match result, noting that it was not very specific (there are 13 million people in Pennsylvania). The prosecutor then asked two outside experts (Drs. Mark Perlin and Robin Cotton) to provide potentially more informative interpretations of the fingernail DNA data.

There are many ways to look at the same DNA mixture. The most informative approach is to consider all the available data – the victim's known genotype and the quantitative peak heights. The victim's DNA comprised 93% of the fingernail specimen, with the STR signals showing huge allele peaks towering over the minor 7% contributor peaks. The jury could see in the data that the victim's DNA was in his own fingernails. With a 15:1 ratio of victim to unknown DNA, the peak heights can help identify the genotype of the unknown contributor.

- Dr. Perlin used Cybergenetics TrueAllele® Casework system to statistically examine the evidence, and fully engage the quantitative peak height data. Separating out the uncertain genotype of the second contributor from the victim genotype, the computer produced a probability distribution of allele pairs that gave a 189 billion match statistic when compared with Trooper Foley's genotype.
- Dr. Cotton's mixture interpretation used the victim's genotype to determine "obligate alleles" of the unknown second contributor genotype. This is a qualitative method that does not use peak height data, and so the "obligate allele" match score was 23 million.
- The FBI had used the controversial "inclusion" method, which applies "thresholds" to discard peak height data, and ignores the victim's genotype. This information-poor analysis gave only a 13 thousand match rarity.

All three interpretation methods were admitted into evidence after a pretrial hearing. All of these methods have support in the scientific literature, and are generally accepted by the relevant scientific community. Courts have been ruling consistently that DNA mixture interpretations are always admissible, though subject to defense cross-examination and expert witness testimony.

Why is the FBI's inclusion method so uninformative in cases like these? By not considering all the evidence, inclusion discards considerable identification information. With data showing four alleles – two tall peaks from the victim, and two short peaks from the unknown – using the victim genotype would give exactly one correct allele pair solution. But inclusion pretends that the victim is not there, and so suggests all ten possible combinations of the four alleles, giving each a probability of only one tenth. By ignoring the victim, and quantitative peak heights, inclusion manages to reduce a jury-convincing match statistic of 189 billion down to a weak approximation of 13 thousand.

Of the many lessons learned from the DNA evidence used to help convict Kevin Foley of first-degree murder, two stand out for the prosecutor:

1. Never accept a DNA match result of under a million-to-one as definitive. More sophisticated interpretations can be done, whether by man or machine, that make much better use of the same DNA data. Although your crime lab may not be able to extract all the identification information in their data, you can call on other experts to conduct an independent review of the evidence.
2. It is perfectly acceptable for juries and judges to see multiple DNA match statistics. Reasonable people understand that using more of the data gives more information. Simple explanations of the underlying assumptions (e.g., "the victim's DNA was found in his own fingernails", or "data peaks have different heights") clarify the methodological differences.

A scientist should seek the truth in every case. The inclusion method of DNA mixture interpretation is certainly easy to do, requires less training, and is easy to explain. But the scientific literature disfavors "inclusion" because it can discard considerable match information – as we saw here in the Foley case.

When a simple DNA interpretation approach produces a match score of under a million-to-one, the scientist should ensure that a more informative method is applied to the same data. Otherwise, the information failure of weaker DNA mixture interpretation methods could lead to a serious miscarriage of justice.

Two narrated PowerPoint lectures on the Foley case can be viewed at www.cybgen.com/information/presentations.

DNA Information and Uncertainty (Feature Article, continued from page 1)

For any given victim genotype allele pair and culprit allele pair, in some mixture weight combination of these two contributors (say, 30% of one and 70% of the other), we can immediately visualize a quantitative pattern of allele peaks. The better a quantitative combination of genotypes and their weight fits the observed quantitative evidence, the higher its probability. We can have a computer try out all possible genotype value combinations and mixture weights, and compare their patterns with the DNA data to see how well they fit. This exhaustive comparison is the underlying principle of all DNA evidence interpretation.

With modern statistical computing, far more can be done. The computer can deliver a scientific probability distribution of allele pairs for every contributor genotype at every genetic locus. The mixture weight information can be inferred for a specimen, using the quantitative data at all STR loci, helping to separate out the genotypes at every locus. Computation can measure how degraded a DNA sample is, which helps solve for the genotypes. Statistical modeling can usefully explain many other STR experiment parameters, such as stutter and imbalance artifacts, as well as data peak uncertainty. And, being a statistical system, the computer determines its confidence in every reported parameter.

Not all DNA laboratories use scientific computing to examine their quantitative STR evidence. Instead, they may simplify the data to make it easier for a person to manually solve for genotypes. For example, some labs apply a threshold that reduces the highly informative quantitative data peaks to simple all-or-none allele events. This simplification comes at a price, however – the loss of identification information. By discarding quantitative data for a simplified qualitative review, the genetic identities become blurred. Rather than comparing genotype hypotheses to the original evidence, the thresholded on-off data can no longer separate out genotypes from DNA mixture data in a highly informative way. Thus, a rapist's genotype merges with his victim, and identification power is lost.

Just how bad is the information loss from qualitative human review? The most common DNA mixture interpretation method used in the United States is "inclusion", where all-or-none thresholded peaks are compared with a suspect genotype. The disparity between quantitative TrueAllele® computer interpretation and simple inclusion is a factor of one million. That is, on average, the inclusion match rarity statistic reported in court is a million times less than what is actually contained in the DNA evidence.

Dr. Mark Perlin (CEO of Cybergenetics) testified this year at a homicide trial where the true match information of the fingernail mixture evidence to the suspect was over 100 billion, while inclusion reported only about 10 thousand (a 10 million-fold reduction). He reported on another homicide case this year that also saw a 10 million-fold match information disparity. Working together with crime labs, Cybergenetics has done extensive comparisons that reveal comparable information loss across all the cases studied.

There is risk in a prosecutor going to court with uninformative DNA match results. Research has shown that juries require at least a million to one DNA match statistic to be convinced. (This statistic means that a match between the evidence and suspect is at least a million times more likely than a random match.) When DNA evidence has data ambiguities, simple qualitative analysis may not reach this level of persuasion. In these situations, a more informative analysis of the same evidence is essential to ensure that justice is properly served. Modern statistical computers, such as the TrueAllele Casework system, can provide this needed quantitative DNA interpretation, and preserve the identification information already present in the data.

More information is provided in our peer-reviewed article on matching uncertain genotypes found at www.cybgen.com/information/publications.

Presenting Match Information in Court (Testifying Tip, continued from page 1)

When presenting more than one DNA interpretation result on the same evidence, the likelihood ratio aspect of the match statistic becomes very helpful. It permits less informative match results (e.g., inclusion in thresholded peaks) to be presented alongside more informative ones (e.g., using victim information, or TrueAllele® computer examination of the quantitative data). This is because, in each case, the likelihood ratio says that under the interpretation assumptions used on the data, the match rarity statistic follows immediately from the inferred evidence genotype.

The following information analogy can be very helpful when testifying to juries about multiple match scores. Imagine that you are a physician trying to diagnose pneumonia from a sputum smear on a glass microscope slide. We need to identify the bacterial organism in order to treat it. Looking at the slide with the naked eye (one type of interpretation) is not very helpful in this case. A higher resolution instrument – say, a magnifying glass – looking at the same data might help, but even Sherlock Holmes won't see the bacteria this way. What is most appropriate in this case is a high-power microscope that lets us visualize the bacteria in the data that we have, make the diagnosis, and then treat the patient.

In the same way, all the different DNA interpretation methods are valid. But some (e.g., qualitative inclusion) are not always as well suited to the case as more informative approaches (e.g., quantitative computer review) might be. Systems like TrueAllele Casework simply offer a more appropriate resolving power in such cases.