SCIENCE TECHNOLOGY INTEGRATION

The DNA Investigator™

Cybergenetics Newsletter Winter 2011

DNA Intelligence and Forensic Failure: What you don’t know can kill you

Finding the genotype

Over nine months, a human being develops from a single cell into a whole organism. These cells share the identical DNA, arranged in 23 chromosome pairs. At a location (or "locus") on a non-sex chromosome, a person inherits one DNA sequence (or "allele") from each parent. So a person's genetic type (or "genotype") at a locus is a pair of alleles, called an "allele pair."

Most of our DNA does not code for anything useful. Possibly a vestige of our evolutionary past, this "junk" DNA can contain long stretches of repeating DNA units, like the four DNA letters "ACGT." Unlike a coding gene, evolution does not greatly constrain mutations in these short tandem repeat (or "STR") junk DNA regions.

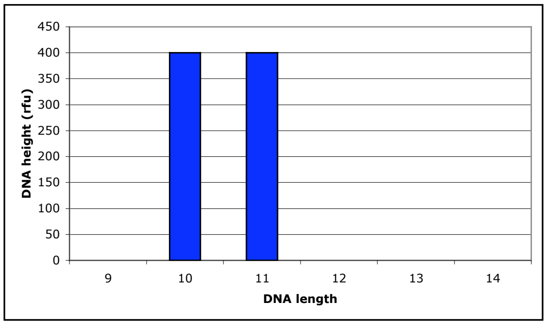

Figure 1. STR data from one person's genotype at a locus.

Thus a locus may exhibit considerable diversity in its allele lengths, say from 10 to 20 tandemly repeated units. With ten different alleles, a STR locus genotype has a hundred possible allele pairings (100 = 10 x 10). These very many different allele pair values can help distinguish between individuals, and so are useful for forensic identification.

It is easy for a DNA laboratory to find the genotype of an individual at a locus. They simply extract the DNA, amplify the two alleles at a locus, and then separate the amplified molecules by their size. The result is a DNA signal (Figure 1), where every identifiable peak in the data tells us the DNA length as the number of repeat units (x-axis) and the amount of amplified DNA (y-axis).

Suppose that a person has a locus genotype with one allele having 10 repeat units, and another allele with 11 repeat units. Figure 1 shows a typical data signal from this genotype allele pair. With clean DNA taken for reference from an individual, we can easily read off the genotype directly from the data as a [10, 11] allele pair.

But crime scene evidence is usually not so simple. For example, in a sexual assault, two people (the victim and assailant) are both present in the DNA. Suppose that the victim's genotype at a locus is a [10, 11] allele pair, and the assailant’s genotype is [12, 13]. Further suppose that there is four times as much victim DNA than assailant DNA in the evidence sample. Then the DNA signal of this mixed sample at the genetic locus would look like Figure 2. With a mixture ratio of 80% to 20%, it is visually evident that in addition to the victim’s known genotype of [10, 11] (blue), we find another person's genotype having allele pair [12, 13] (green).

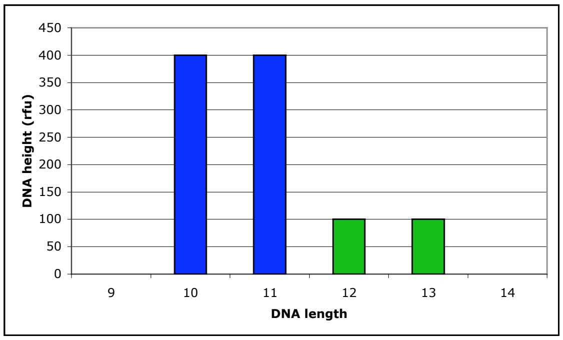

Figure 2. STR mixture data from two people's genotypes at a locus.

www.cybgen.com

Losing the genotype

Suppose you are looking at a hilly terrain, with peaks ranging from dozens to thousands of feet high. Imagine someone told you that every peak over 100 feet tall was a real hill. And all real hills share the same height of 100 feet. But any peak under 100 feet wasn't a hill, and wasn't really there at all. Would you believe them?

Or if your doctor said that all electrocardiogram peaks over 1 inch high meant exactly the same thing to him, and he just ignored all ECG peaks under 1 inch. Would you find another doctor? How many missed diagnoses and dead patients would it take before he was barred from medical practice, or jailed for his diagnostic misdeeds?

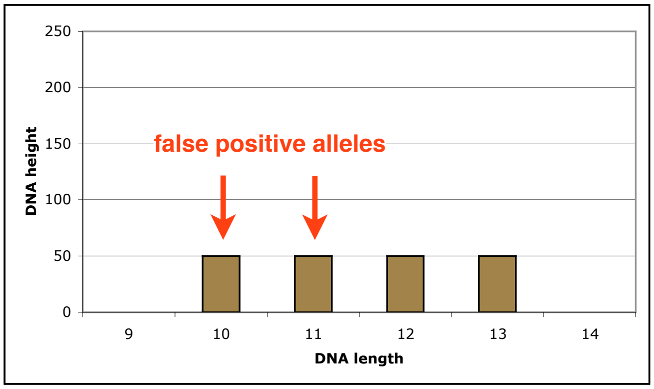

Figure 3. A low threshold applied to mixture data, introducing false positive error.

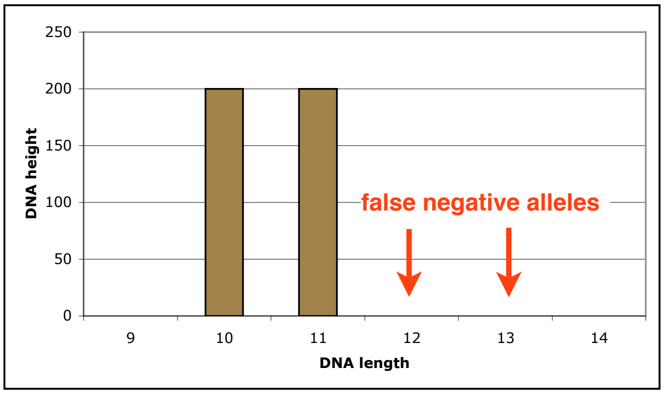

Figure 4. A high threshold applied to mixture data, introducing false negative error.

Yet this is the all-or-none data approach of forensic DNA analysis. A DNA lab sets a "threshold", and then considers all data peaks above that threshold to have equal stature. Peaks below that threshold are largely ignored. The intent is worthy (controlling uncertain data), but these thresholds have no scientific justification based on any rigorous math, logic or chemistry. Practitioners are taught early on how to discard evidence in this way, and they faithfully follow tradition.

Let us see what this data decimation does to our mixture example (Figure 2). Suppose we set a threshold of 50 relative fluorescence units (rfu). Then all four peaks would be considered to be "alleles," without regard to their height (Figure 3). Of course, a person's genotype is an *allele pair*, not an allele. The widely practiced "inclusion" approach would form all ten imaginable allele pairs from these four alleles (Table 1).

Table 1. How thresholds lose information by introducing spurious possibilities.

[10 10]
[10 11]
[10 12]
[10 13]
[11 11]
[11 12]
[11 13]
[12 12]
[12 13]
[13 13]

All ten possible allele pairs are formed from the "allele" set {10, 11, 12, 13}. In our example, only the allele pair [12 13] has support in the quantitative evidence. The nine other conjectures for the second genotype are wrong.

The computer can tame uncertainty through statistical modeling of all the quantitative data (i.e., no thresholds), and here finds the one true allele pair. But applying thresholds here introduced ten answers, nine of which must be wrong. This artificially induced ambiguity at a locus reduces DNA match information about ten-fold.

A different lab might set their threshold at 200 rfu. Applying this higher peak cutoff to our data (Figure 2) slices off the tops of peaks 10 and 11, and renders invisible peaks 12 and 13. In the minds of many forensically trained analysts, our mixture data has been reduced to just the victim's allele pair, while the perpetrator peaks completely escape attention (Figure 4). The criminal-identifying data, genotype and information have entirely vanished. A higher threshold loses the second genotype, and thus fails to find the assailant (Table 2).

Table 2. Why thresholds are scientifically invalid for DNA identification.

a. Genetic identity involves comparing genotypes (allele pairs). Thresholds produce sets of (often incorrect) alleles. But alleles are not genotypes. Solving the wrong problem will lead to the wrong answer.

b. Valid scientific inference mandates that the observed data must not be altered. But thresholds radically change the data from information-rich peak heights to error-prone all-or-none events.

c. Valid scientific inference requires complete consideration of all possible values of all relevant variables. But threshold methods examine only a few values of a few variables, and fail to adequately explain the data.

d. STR data have a physical uncertainty that varies in direct proportion to the DNA quantity (measured as peak height). But threshold methods incorrectly assume a constant data uncertainty.

What you don't know

There are two types of scientific error. Forensic DNA strives to avoid "false positive" mismatches that might wrongly implicate someone. But equally troublesome is the "false negative" error that fails to identify. A small error rate in science is 1%; a large one is 10%. How large is the forensic DNA false negative error rate?

Published studies show that mathematical computing can preserve DNA identification information. Proper statistical reasoning requires that (a) the quantitative peak data are not altered, and (b) all genotype allele pair possibilities are considered. Computers can be programmed to reason in this way. Compared with real computation, human discarding of DNA evidence data with simplistic "thresholds" makes match scores a million times weaker. Some recent studies shed light on the false negative errors introduced by thresholds.

Threshold methods make real data peaks disappear, falsely concluding that a true allele is not present. Let us count how many alleles disappear from each genetic locus. This will be our false negative statistic – the average number of missing alleles per locus.

We examined forty two-person DNA mixtures of known composition. The mixture proportions were 50:50, 30:70 and 10:90, and the DNA quantities were 1 ng, 1/2 ng, 1/4 ng and 1/8 ng. (A nanogram, or "ng", of DNA contains about 300 chromosome copies.) We had forensic scientists apply different thresholds to the fifteen loci of each mixture's STR data, and tallied up the missing alleles.

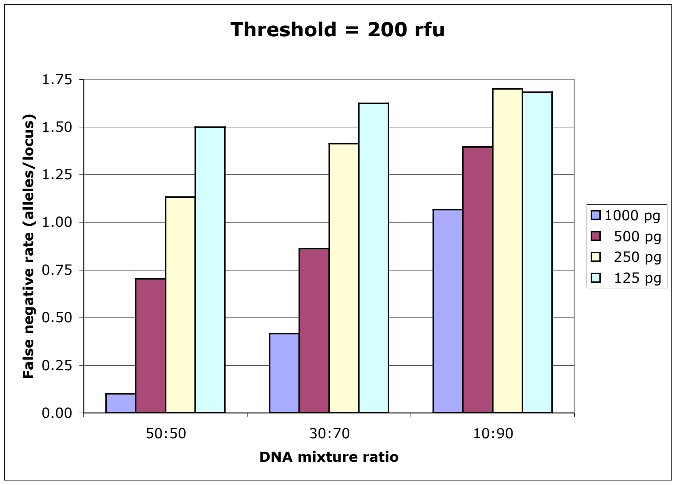

Figure 5. False negative error rates caused by thresholds under different conditions.

We examined the effect of thresholds set at 50, 100 and 200 rfu. Until recently, many labs used more sensitive thresholds between 50 and 100. However, in 2010, new DNA interpretation guidelines raised American laboratory thresholds closer to 200. In Figure 5 we see the resulting false error rates.

With 50:50 mixtures, the error rate ranges from 10% at 1 ng of DNA, up to 150% with a smaller 1/8 ng DNA amount. The missing alleles per locus increase to a 40% - 160% range with more typical 30:70 mixtures. When the perpetrator's DNA comprises only a small 10% proportion of the mixture, the error rate reaches 100% to 170%.

Error rates exceeding 100% are unprecedented in science. So perhaps applying thresholds to quantitative DNA evidence data does not constitute the best science. The principles of probability really do prohibit deducing directly from the data in this simple way. Since thresholds skirt the laws of science, their exaggerated error rates are not surprising.

How it can kill you

Is it really so bad that these FBI sanctioned threshold methods fail to identify? What harm could come from being "conservative" and discarding most of the DNA information? Let us consider three situations.

First, suppose you are falsely accused of a terrible crime and face decades in prison. Looking at the mixture, there is someone else's DNA in the data that clearly doesn't belong to you, but this stranger's peaks fall below threshold. Depending on how far down the peaks have fallen, some labs might report this evidence as "inconclusive". Even though a mathematical computer can correctly identify the true culprit, an uninformative human interpretation might not. Because thresholds hide data, they can suppress DNA evidence favorable to a defendant.

In another scenario, your friendly neighborhood thug oversteps again; he is now charged with his violent crimes. But his DNA data peaks fall below threshold. A computer would have accurately and objectively identified him, but thresholds were used instead. The guilty man is not convicted, and so continues his local reign of terror. You may well know his next victim.

Here is one last situation that is not so hypothetical. A national DNA database of convicted offenders provides police with investigative leads to unsolved crimes. DNA genotypes from a crime scene are entered, and the database searches for a match to some registered criminal. Ideally, highly informative evidence genotypes would be stored in this database, so that the one true perpetrator could be DNA identified, apprehended, convicted and incarcerated. And society would enjoy the protective safety of accurate scientific DNA investigation.

Instead, the FBI's CODIS (Combined DNA Index System) stores DNA evidence as "alleles", an improper genotype representation that discards most of the identification information. When the police run a "low stringency" database search of uncertain evidence against convicted offender profiles, CODIS can nonspecifically return hundreds of candidate matches. The true perpetrator might be on the list, or not, but all the rest of these false positive "matches" are made to the wrong people.

When running a "high stringency" search, as with high thresholds, the lost sensitivity can yield no match at all. The resulting false negative non-matches fail to detect the true perpetrator. And so, instead of finding and stopping criminals, CODIS match failure lets felons circulate in society, continuously victimizing innocent citizens.

An enlightened solution would be to require crime labs to preserve all DNA information, instead of routinely throwing it away. Competent computers should solve mixtures, and other complex DNA evidence, by considering all possibilities to accurately infer genotypes. A more intelligent DNA database should store these information-rich genotype probabilities, quantitatively matching them to preserve identification power. And the less effective utilitarian "threshold" methods should be abandoned. Dumbing down DNA does not secure a safer society.

Where to learn more

There is much information about forensic DNA evidence freely available on the Internet. The laboratory procedures that transform biological samples into DNA data signals are often described quite well. However, articles on overly simplistic threshold-based STR interpretation methods should be traversed with caution.

Cybergenetics provides free educational material on its website about accurate quantitative interpretation of DNA evidence. If you visit www.cybgen.com/information, you will find many courses, presentations and publications. A good starting point is our newsletters (such as this one), which are written for the general public.

On January 11, 2011, Cybergenetics CEO Dr. Mark Perlin addressed the FBI's Scientific Working Group on DNA Analysis Methods (SWGDAM) on "The science of quantitative DNA mixture interpretation." He explained the underlying principles of preserving DNA identification information, showing comparison studies and criminal cases. He also clarified why "threshold" methods do not work. The narrated slide movie can be found at www.cybgen.com/information/presentations.shtml.

If you know of an information-poor DNA interpretation result that could lead to a miscarriage of justice, please contact Cybergenetics. Our mission is to protect the public through better science.

Cybergenetics © 2011, All rights reserved