Transcript of Dr. Mark Perlin's talk on "Combining DNA Evidence for Greater Match Information" delivered on 24 February 2012 in Atlanta, GA at the American Academy of Forensic Sciences 64th annual meeting.

Dr. Perlin: Today I will be talking about how to combine DNA evidence to get more match information. Cybergenetics TrueAllele® system computed the various statistics shown in this presentation.

(Next slide)

We begin with a mixture sample. Sometimes there is more than one mixture sample; here we see two. The samples could come from the same item, through two different amplifications. The samples could be from two different items. Regardless, suppose the samples contain the same two contributors, but show very different short tandem repeat (STR) data patterns. We see this with Sample 1 and Sample 2 at the STR CSF locus.

(Next slide)

The computer can look at all 15 STR loci and mathematically separate out the contributors, determining the mixture weights. The mixture weight has both a mean and a standard deviation, because scientific data has uncertainty. For Sample 1, we get about 10% mixture (this orange component is the one we will be interested in), as well as a major 90% component (blue). The second sample is about a 50-50 mixture.

(Next slide)

The key scientific idea here is the joint likelihood function. The concept and its mathematics were described in our November JFS paper, and in other papers. The idea is how to explain the

evidence data based on some hypothesis.

Let us look at one hypothesis, that we have a pair of alleles from one contributor whose genotype is [10, 11] (blue), and we have another contributor whose genotype is allele pair [12, 12] (orange). If the contributor proportion is 90% blue to 10% orange in the first sample, and a 50-50 of blue to orange of those allele pairs (i.e., genotype values) in the second sample, then we see that nicely accounts for the data. That would be a good explanation, which would have a high likelihood.

The computer tries out all possible explanations – where the alleles are, where they are not. It may try out a [13, 14]. The computer does not care; that genotype hypothesis would just have low likelihood. After trying out ten thousand or a hundred thousand different possible explanations, it computes a genotype.

(Next slide)

When we begin, the number of possibilities at CSF was perhaps a hundred allele pairs. Looking at data constrains what the possible genotypes can be. When more than one possibility remains, that uncertainty is expressed in probability. Here are three different genotypes. The dark blue left bars correspond to Sample 1, placing probability at allele pair [10, 12], as well as [12, 12], which you can see are feasible from the data in the first sample.

Sample 2 also distributes its probability between two different allele pairs, [10, 11] and [11,12]. However, when we look at all the data together, there is really only one explanation that can account for all that data. That is, the STR data from Sample 1 and Sample 2 constrain what the possible genotype answers can be, and the probability (light blue-green bar on the right) is jointly just one explanation, the allele pair [12, 12].

Notice that these genotypes are inferred objectively. There is no concept yet of a suspect, or the

database we will be comparing against. It is all done just from the evidence data, in this case from two samples.

It turns out that one of the contributors is a [12, 12]. We will soon see the match strength. The match statistic is proportional to the genotype probability. There is a high likelihood ratio (LR) from a lot of genotype probability, and less LR from little genotype probability. The LR is almost nothing if interpretation has not placed genotype probability on the reference allele pair value.

(Next slide)

Here is the genotype match information at fifteen loci. Shown again is CSF, with the three LR bars, now horizontally. On top at CSF is the dark blue LR bbar from one sample, no match support from the second sample, and the joint LR bar that goes all the way out to 10. The LRs are shown on a logarithmic scale, because information is expressed in the exponents, the number of zeros (powers of 10). That way, we can add information to get the total. If you prefer, you can think of it as multiplying the likelihood ratios together, but most papers on information and DNA match statistics use the number of zeros (or the powers of ten).

At every locus, the first two bars represent Sample 1 and Sample 2, and the third bar (greenish blue) is larger, having more match strength than the match information from either of the separate samples. The total information from Sample 1 is about 10^{10}, more than a billion, which is nine zeros. Sample 2 by itself gives also about a billion, nine zeros, but when the computer has to explain both of the data items together in a joint way, we end up with twice that information, a billion billion, as the match statistic, or a quintillion. This example shows the basic principle, that using more data in a rigorous, mathematical way can provide more information in a match statistic.

(Next slide)

I would now like to turn to a case example. On March 7,009 in Antrim, Northern Ireland, there was an attack on the Massereene Barracks for which the Real Irish Republican Army claimed responsibility. There were four unarmed soldiers collecting pizza from two pizza deliverymen, when a car drove up this road. In less than 60 seconds, two hooded gunmen came out and fired over 60 rounds from automatic weapons into the people, repeatedly firing into soldiers on the ground. Two young soldiers died, Patrick Azimkar and Mark Quinsey.

(Next slide)

An investigation was launched that involved 60 police investigators. The getaway car was burned in order to hide the DNA evidence and destroy it. It was only partially burned, however. Teams of forensic scientists recovered DNA evidence.

(Next slide)

There were several items, including these three touch DNA items for which there was no DNA match statistic. They were a passenger-side safety belt buckle, a cell phone that was used to make a phone call describing the attack shortly after it happened, and a matchstick that was used to light the car that was found on the side of the road. The second two items were directly connected to the crime, while the first one was merely at the scene of the crime.

The biological samples were sent to DNA labs. These three evidence items were sent to Cellmark in the United Kingdom near Oxford, who did multiple amplifications on each one. In fact, on the cell phone they did several enhancements as well.

(Next slide)

Here is the first DNA profile. We see that the data from the matchstick has ambiguity. We are not going to get a unique genotype from this data. Notice that the peak height scale goes from 0 RFU to 120 RFU. Most of these peaks are around 50 RFU or below, so conventional human interpretation methods will not get much of a result.

We can see the sources of uncertainty. We have mixtures; at D3 and at D8 there were more than two allelic peaks. There is low DNA quantity at TH01, with peaks well under 50 RFU. There is no DNA visible at some loci like FGA, and there is probably allele dropout at others. This is highly ambiguous, uncertain DNA evidence.

(Next slide)

To address that uncertainty, multiple DNA amplifications were done. Here are three different amplifications *a*, *b*, and *c*, each showing data at 10 SGMPlus STR loci. A human review attempt was made to use a consensus method, but in fact there was little consensus between the three amplifications. At locus D21, we see that, in a comparison with suspect Brian Shivers (red dots), there are zero, two, or one numbers of alleles that have appeared. No match statistic was produced. Without a match statistic, there is a risk that the court will not admit the DNA into evidence, because there is no statement as to how rare this genotype is nor how specific it is to a defendant.

(Next slide)

At this point about a year ago, Cybergenetics was asked to look at the data. The slide shows a joint likelihood function. Here the computer is looking at three different amplifications (*a*, *b*, and *c*) at the TH01 locus. The computer tries out virtually all possible combinations of different allele pairs. The first contributor allele pair (blue bars) and second contributor allele pair (orange bars) move all over the place. The different bar heights indicate different DNA amounts. We see one

particular allele peak pattern that fits the data pretty well, and explains the data signal we are seeing. Not all signal peaks are explained, such as amplification. When the computer finished at TH01, in fact, most of the probability did land on this allele pair [6, 9.3].

(Next slide)

Combining the DNA amplification data increased identification information. Looking at the data of three samples, each of them in isolation (*a* alone, *b* alone, *c* alone), we see match statistics of 25, 27, and 6 over those ten STR tests. In gray above, I have shown the logarithm, the powers of 10. 25 is a number between 10 and 100, which has between one and two zeros, and so that number, 1.4, describes how much information is present. It is the exponent of the number 25.

Looking at pairs of amplification experiments (next column), the computer could look at *a* and *b* together, *a* and *c* together, or *b* and *c* together. Examining twenty locus tests simultaneously increased the information to about three or four zeros after the 1. Finally, when the computer looked at all three evidence items together over all 30 STR tests (last column), it arrived at a number that had an information of six zeros after the 1, a little bit over a million. These computer runs were done at least in duplicate for the single tests, and with more replicates for combined tests in order to establish reproducibility.

(Next slide)

The computer's match statistic for the matchstick on the side of the road was about a million to Mr. Shivers. Only after the genotypes were computed was a comparison made to the defendant. The match number was about six billion for the cell phone to Mr. Shivers, and about six trillion for another suspect, Colin Duffy, to the passenger-side belt buckle.

Since these two items (matchstick and cell phone) were directly linked to the crime, Mr. Shivers

was convicted and sentenced to 25 years. Mr. Duffy was believed by the judge to be in the car because of DNA evidence, but he was not linked to the crime.

(Next slide)

In December of this past year, after a seven week trial that I testified in for three days in November, Judge Hart ruled that the TrueAllele system was admissible in evidence, concluding that he was “satisfied that the stage has now been reached in the case of this system where it can be regarded as being reliable and accepted,” and he was “satisfied that Dr. Perlin has given his evidence in a credible and reliable fashion,” and “in light of these conclusions [he] saw no basis on which [he] could properly… exclude this evidence,” and therefore he admitted it in evidence.

It is also worth mentioning that last week the Pennsylvania Superior Court published a decision that establishes precedent for TrueAllele computer interpretation of DNA evidence in the Commonwealth of Pennsylvania.

(Next slide)

It was noted at the time, on the same day as the verdict in which Mr. Shivers was convicted, that these DNA techniques used in the Massereene conviction could pave the way for future trials, particularly in crimes that are terrorist attacks, where no witnesses will come forward and the main evidence is forensic. The TrueAllele computer is a new tool for investigators, prosecutors, and the defense to get more information out of the same DNA evidence.

It is also worth noting there are many crimes in the U.S. (e.g., drug homicides) where witnesses will not come forward, and, as in Massereene, such cases are made largely on the basis of forensic science.

(Next slide)

As the great seal of the U.S. says, “E pluribus unum.” Out of many, there comes one. In the Massereene case, we saw many police investigators working with many forensic scientists to gather evidence. We saw the crime lab with its data working together with the best that people could do with computers. Looking at multiple DNA amplifications and much data, combining them mathematically, we were able to infer more informative match scores in the range of a million that were persuasive in court. Out of many, we arrived at one answer. Thank you very much.