Transcript of Dr. Mark Perlin's talk on "Reliable interpretation of stochastic DNA evidence" delivered on 2 December 2010 in Toronto, Canada at the Canadian Society of Forensic Sciences 57th Annual Meeting.

Dr. Perlin: Today I will be talking about reliable interpretation of what we might think of as unreliable or stochastic DNA evidence.

(Next Slide)

In nature, we have uncertainty. All real data that we observe in science is not completely reproducible. If we were to re-amplify data and see the exact same peaks, the we would know that there was some mistake, such as a duplicate file. Reproducible DNA data exhibits stochastic variation. The question is, what do we mean by "reproducible"? I will discuss that in the next few minutes.

Probability models can capture the stochastic nature of the data. In all methods of interpretation including mixtures and the protocols that we follow, genotypes are inferred from these uncertain data either implicitly or explicitly. These uncertain genotypes are all described using probability distributions. Even inclusion is a probability distribution. Science expects us to account for the stochastic effects somehow, and the law expects us to testify within our certainty. All match statistics, including CPI, are likelihood ratios, and this meets the demands of both science and law. That is the measure that will be using in these

studies.

(Next Slide)

At the heart of the data randomness that we see in nuclear DNA is the polymerase chain reaction (PCR). Inherently, every cycle there are a number of copies of DNA exist, and in that random PCR branching process at each cycle, either there is a copy made or there is not a copy made. The efficiency is not 100%. Therefore, this copying probability, p, is not equal to one. The result is a random distribution of what we might observe as a peak.

(Next Slide)

For example, under the same conditions in the laboratory or in a computer simulation, an STR peak is a random variable. With one amplification, if we do a certain number of cycles, then we may end up at 1300 copies. With another amplification, we may end up with more than twice that number of copies.

(Next Slide)

What we are really getting is a probability distribution of the peak heights that reflects the underlying quantity of DNA. If we were to do 1000 or 2000 amplifications of the identical template or simulate this on a computer, then we

would get a probability distribution. Here, we see a peak distribution that covers maybe half of the distance from baseline to the maximum observable peak. The x-axis represents the amount of DNA, which is measured on a fluorescent DNA sequencer as a peak height (horizontally for right now), and the probability distribution (y-axis) characterizes our uncertainty. The distribution has a mean and a standard deviation, which quantitatively describes the extent of our uncertainty.

(Next Slide)

One way of understanding the extent of our certainty through a single number is the coefficient of variation (CV). The CV divides the standard deviation by the mean value. What is the expected value of an amount of DNA or a peak height? What is the variation that we see? The long bar (in green) is the extent of the average peak height. One standard deviation unit is shown in red. If we take the CV ratio here at 12/85, we see a coefficient of variation of 14%. This bell curve of peak distributions is taking up about half of the distance from the baseline up to the maximal peak that we would see.

(Next Slide)

On the other hand, if we quadruple the amount of DNA or the peak height, the we get twice the peak certainty. That is from basic statistics or from mass action in

chemistry. Four times a quantity will yield twice the precision. We see that result whether we do these experiments in the lab or if we do them by computer simulation.

Here, we have quadrupled the DNA amount up to 350 rfu. We have a much greater quantity of DNA. The standard deviation has doubled. Dividing a doubled standard deviation by a quadrupled mean, we end up halving the coefficient of variation. That is reflected visually – the peak probability distribution is now taking up a quarter of the distance from baseline up to the maximum peak. However, before with a quarter of the DNA amount, the distribution had taken up half of the total space. So, there is more concentration of probability of where the peak height will be around the mean value, as there is more and more DNA.

(Next Slide)

We see STR data of different fragment lengths (x-axis). Now, we are drawing the distribution in the usual way, with *rfu* on the y-axis. These are probability distributions. So, when we observe a particular peak height, that value is just a sampling from the peak probability distribution. With taller peaks, more of a probability mass is concentrated in one area very far from baseline. With smaller amounts of DNA, that probability mass is more diffuse on a relative scale to the baseline. We can see how a lower peak is covering as much as half of the space to the baseline while a taller peak the DNA is covering maybe only a quarter of

that relative distance.

(Next Slide)

How do we interpret quantitative data that has uncertainty? Some forensic groups like to use qualitative thresholds. In that method, one threshold is applied to the data. All peaks that are over that threshold are treated as bona fide allele events. Those peaks that are under threshold are discarded, and the result is that the quantitative data is reduced to a list of alleles that are "included."

(Next Slide)

However, nature is providing us with peak events that are samples of a probability distribution from the PCR of the underlying DNA. So, the instant we do something that is not working with a probability distribution itself but instead start drawing lines, we introduce error inherently. The reason is that when we draw a threshold, there is now a notion of the "wrong" side of the probability distribution. Starting with this peak probability distribution, if the PCR samples a true peak on the lower side of the threshold, then we produce a false negative error. If we sample a nonevent above the threshold, then we produce a false positive error. We can see this clearly at the 200 *rfu* threshold line for a real peak event. The peak event really did occur, but with the threshold, we are choosing not to see them. The peaks have now become invisible. The instant we draw a line, we

have created a statistical test that casts everything on one side of the distribution as an error or everything on the other side as an error, depending on whether we observe a peak event or not. Before we drew the threshold, we could not be wrong – it was just the original probability distribution. Once we institute a threshold, there is now a mathematical way to quantify our probability of error.

(Next Slide)

We did some experiments on this. This involved 40 mixture samples with one nanogram, ½ nanogram, ¼ nanogram, and 1/8 of a nanogram at five different mixture proportions of 10:90, 30:70, 50:50, 70:30, and 90:10 for two different pairs of individuals. Overall, this was 600 loci, and what we measured was what the false allele rate was. What is the false negative rate measured in terms of the number of alleles that were missed per locus? So, 1 here means 100% missing of how many alleles there would be per locus on average. We see with 50:50 mixtures with a lot of DNA, the error is low, but even with a reasonable amount of DNA (500 ng, 250 ng) with an imbalanced mixture, the error rate is getting up to 100% of the number of incorrectly missed alleles per locus.

(Next Slide)

That was at a threshold of 50 *rfu*. When we use stochastic thresholds and thus raise the cut off level to say 150 or 200 *rfu*, then the peaks that are under

threshold begin to disappear from our review, as do the criminals that we are trying to detect and identify. The result is that there is no genotype. There is no match score. There is really nothing left to say to the police, prosecutors, or society because the data has now become invisible to our interpretation method. The peak data are there, but we are not seeing them because of the artificial thresholds.

(Next Slide)

The interpretation result is a higher false exclusion rate. If the threshold goes up from 50 rfu to 200 rfu, then on the same 600 loci that we looked at, we are averaging a false negative rate of one allele per locus. Basically, for most the locus data, the threshold is falsely discarding alleles a rate of about 100%. For example, for a minor 10%:90% highly imbalanced two-person mixture, the allele drop out error rate is at 100% or higher, regardless of DNA concentration.

(Next Slide)

However, the new SWGDAM guidelines in paragraph 3.2.2 provide for this. In essence, the say that if we are using quantitative data with probabilistic genotype methods and we empirically support it, then we can essentially ignore the rest of the document. However, this is only if we use a validated probabilistic genotype method. If we do not and we are just guessing at peak certainty with a qualitative

approach, then we must use a stochastic threshold. However, high peak thresholds will discard identification information, whereas probability modeling, as done and published by many groups around the world, can preserve a lot more match strength.

(Next Slide)

What is probability modeling? All DNA interpretation methods use probability modeling. The threshold mixture methods also use probability modeling but employ a fairly uninformative likelihood function. When using quantitative STR peak data, the likelihood functions become more informative, and so it pays to discuss how probability modeling is done.

At the heart of the forensic sciences is Bayes' theorem, which provides an accepted way of addressing data uncertainty in most fields of natural and social science. The idea is to use a likelihood function to update our probabilities. (I will show a picture of one soon). We start off assuming that we do not know anything about an individual's genotype, and so the probably distribution is that of the population. When we see STR data, we then update our beliefs. With DNA mixtures, data ambiguity will often remain in the genotype. With single source data, our belief can be focused on one allele pair to give a definite solution.

A joint likelihood function combines independent evidence, whether it is from

independent peaks within a locus or across loci. The purpose of a likelihood function is to tell us how well model parameters explain the data. I will give an example of this in a second. With STR data, what we want to explain are the peaks. Formally, a likelihood function gives the probability at one peak, and when we combine likelihoods, we can explain the whole pattern.

(Next Slide)

Here is how one could propose a genotype. Here is a mixture with two alleles (shown in blue) from one contributor and one allele (shown in orange) from another contributor. We could propose a certain amount of DNA or a mixture weight. We could also propose, as we do with our systems, stutter amounts, relative amplification, degraded DNA extent, and so on. We then ask: how well does the proposed pattern compare with the peaks?

The peaks are probability distributions. We can try out every possible pattern by considering every possible genotype value, including the ones that we do not see because maybe there is drop out, across all possible mixture weights, all possible stutters, degraded DNA, and all other possible model parameters. This activity is not for the faint of heart. If we do not have a computer, then forget it because a computer is needed to try out billions of possibilities. Those patterns that better explain the data have higher probability. Therefore, the genotypes that produce those more explanatory patterns have higher probability. That is the

heart of all modern statistical computation. Everything gets tried out, and whatever better explains has a higher likelihood. This comparison picture shows the data and a proposed pattern. The picture shows a likelihood function that can account for stochastic data of any peak heights, including baseline.

(Next Slide)

What is interesting is that all variables, not just genotypes and allele pairs, but also stutter, pref amp, and (importantly) peak variation can be solved in the same way using a probability distribution. The distribution of the data – how confident we are in one peak as a representative of every observable peak, as well as around baseline – can be modeled as just one more parameter in the data. The formula for a bell curve has a mean and a variance. The mean is one parameter. The variance is another parameter. Computers can solve for both parameters as probability distributions. So, with modern computation, the certainty of every peak in the data can be computed. That is how we can use quantitative likelihood functions in a reliable way with quantitative data for DNA mixtures and other problems.

(Next Slide)

TrueAllele is a computer system that implements this probability approach. There are other groups in the world that do this as well. The approach is quantitative

computer interpretation. TrueAllele does a statistical search across all of the parameters in the probability model by examining all of the possible genotypes at all of the contributors, mixture weights, degraded DNA, stutter values, and so on. The goal is to preserve all of the identification information present in the data.

TrueAllele objectively infers a genotype having never seen the suspect. It would not know what to do with a suspect genotype since that is an unknown variable that it is solving for. Then, after the system has inferred its genotypes from the data, the genotype probability distribution is put into a formula, and a likelihood ratio is computed. TrueAllele can handle any number of mixture contributors, though I do not think that we have gone beyond five unknown contributors in our studies.

System random variables include stutter, peak imbalance, and degraded DNA. TrueAllele also calculates the uncertainty of every peak. If a computer does not compute the uncertainty of every peak, do not bother using it. Just stick with thresholds. This is because the results would be unreliable and easily destroyed in court. The TrueAllele system is over 10 years old, now in version 25. We have used it on over 100,000 evidence samples, including the World Trade Center, reprocessing all of the DNA data. Forensic groups often test out TrueAllele by sending us one or two interesting cases. They select ones having a very low match score or cases where they believe some information is there but their guidelines will not let them do anything with it. We do not charge for this

examination. We then analyze the data in TrueAllele and do a webinar to walk through the user interface and show the genotypes and match scores that the system found. There have actually been some convictions that resulted from these free trial test cases that people sent us. The police would present the match results that we found in TrueAllele to a suspect, who would then say, “Ok, I confess.”

(Next Slide)

This was the first case that I know of in the world where statistical computing was used for DNA. I testified in this a year and a half ago in Pennsylvania. It was a 7% two-person contributor. The unknown was under someone's fingernails from a homicide case. The inclusion score by a national laboratory in the US was 13,000, and TrueAllele produced a score about 1 million times more. The whole theory was presented in the admissibility hearing, explaining why peak thresholds discard information whereas probability modeling preserves it.

(Next Slide)

In this case, here is the vWA locus. Suppose we have three alleles, and in this case with a 7% mixture, we would have some very tall peaks from the victim and some very small peaks also visible. With CPI or RMNE, all of the alleles are considered to have equal standing as in or out, and the victim genotype is not

used. Of the three possible alleles, there are six possible allele pairs. The result is that the inclusion method disperses its probability over all of the allele pairs, including ones that are obviously incorrect by visually looking at the data. But, a quantitative method, like TrueAllele, forms a likelihood function that, after its calculations, focuses the probability on the true allele pair value for the true genotype. When we focus probability on the right allele pair, then we then get a higher match statistic.

(Next Slide)

This improvement was not just seen in this case. I will give you the URL at the end, and also the slides are on our website for the talk as a hand-out. This paper with the New York State lab is coming out in JFS next fall. In it, we showed on the same eight mixture cases that the average match score using a quantitative TrueAllele probabilistic genotype interpretation was about 10^{13} (note the logarithmic scale) relative to their reported CPI value of 10^{7}. That is, the data tell us to expect to get about a 1 million to one improvement in match score without a victim reference on those cases where people can produce a result.

(Next Slide)

SWGDAM also provides for combining evidence. Here a joint likelihood function is useful.

(Next Slide)

This is a case that I testified in over the summer in England: The Queen vs. an arsonist. This was fascinating data. If it can be believed, these are all amplifications of the same locus of the same template for three contributors. They look a little bit different. It is enhanced, which is why these peak heights are actually really all below 50 *rfu* in the real world. When we looked at it, the human result was inconclusive, and with the TrueAllele interpretation

(Next Slide)

this is what we found. Let us focus on locus vWA. In brown, we see the prior genotype probability before seeing the data. After the computer spent about a day processing since this is a very hard case using a lot of different evidence, the joint likelihood function was able to combine that evidence and we see the posterior genotype distribution (in blue). The system does not know who the suspect is, but at some allele pairs there is a gain in probability while at some other values there is a loss. It turns out that at this locus, the suspect's genotype is the allele pair [14,18], and there is a 6-fold probability gain from prior to posterior. This can be seen by looking at the ratio of the blue bar to the brown bar. That probability gain is the likelihood ratio match statistic at the vWA locus. Multiplying across all of the SMGplus loci in this case, the joint likelihood ratio

was over 3 million, even when accounting for population substructure at 1%.

(Next Slide)

In a larger study with New York State, we looked at about 85 items of evidence. What we found, as seen on this descending blue curve, is that the computer always produced a match score. The y-axis shows the likelihood ratio of 10^{15}, 10^{10}, 10^{20} and so on. People also put a match score to the data and recorded it in the case record. New York is a top lab. They use a peak height threshold of 50 until the new SWGDAM guidelines came out. They aggressively make use of the available data, never overstating a result, and they are very clean in their analysis. The New York State lab was observed to put a match score to the evidence about 30% of the time. When they did find a score (unless the mixture was treated like a single source RMP) using peak threshold methods, there was the expected loss of information. However, with their qualitative threshold methods, they were unable to make a match statement or put any match score to the data over 70% of the time. However, the quantitative TrueAllele interpretation method was able to preserve the data's identification information all of the time.

(Next Slide)

In conclusion, real quantitative data has stochastic effects. If it did not exhibit random variation, it would just be fake data drawn on Photoshop – it would not

be real. In nature, all data has stochastic effects. That is good because with modern probability analysis, we can model quantitative data using joint likelihood functions, as is done in hundreds of other fields. Probability modeling exploits the stochastic effects. It tames and captures them by working out the peak variation's probability distributions in order to preserve identification information. These interpretation methods are based on established probability methods that are hundreds of years old. A joint likelihood function can rigorously combine the DNA evidence's quantitative data, a TrueAllele computer can do the necessary statistical calculations, and an exact modeling of peak variation can replace inexact thresholds and scientifically overcome stochastic effects.

We have written a number of papers and have presented our findings at scientific meetings. We gave about ten scientific presentations in 2010. We began at the AAFS meeting in February, speaking about the Foley case and the New York State validation study. We prepare movies of the slides along with the talk's audio recording. If you do not want to listen to my voice, we also provide transcripts and handouts. This scientific and educational material is available on our website under "Information". We also provide preprints and reprints of our papers, should you want to read about the interpretation methods or validation studies. If you are interested in seeing how TrueAllele works on cases that intrigue you, please send me an e-mail. Cybergenetics would be happy to take a look, and show you on your own data at no cost what highly informative quantitative interpretation is all about. Thank you.

