Transcript of Dr. Mark Perlin's talk on "Profiles in productivity: Greater yield at lower cost with computer DNA interpretation" delivered on 8 September 2010 in Sydney, Australia at the Twentieth International Symposium on the Forensic Sciences of the Australian and New Zealand Forensic Science Society.

Dr. Perlin: Thank you. I would like to thank the organizers for inviting us. This is joint work with Dr. Barry Duceman of the New York State Police. We have been collaborating for almost ten years on this project.

(Next Slide)

To use another meaning of the phrase, there is currently a crushing “weight of DNA evidence.” The amount of DNA quantity just keeps going up and up with property crime, touch DNA, and demands on laboratories. What some people perceive as DNA quality with mixed samples, degraded samples, low-level, and so on seems to be going down. This leads to some DNA problems, such as backlogs and low information yield. The crushing weight also leads to a possible long wait for evidence. Certainly in the US, we are seeing more and more criteria for when to not analyze the data at all and move onto something else — and that is a very long wait indeed.

(Next Slide)

TrueAllele Casework provides computer interpretation of DNA evidence, as some of us may have heard yesterday. It mathematically models the STR process. The primary goals of the project with New York and many other groups are to achieve objectivity in DNA interpretation, ease-of-use, and being able to get interpretation happening for us by machine. Information is a significant goal. We want as much information yield as possible from the data that is being brought into the lab. We do not want to throw anything out if it contains information. The focus of this talk is productivity, and these are the same four terms that Barry uses in his talks all of the time about TrueAllele. In this talk, productivity will measure how much work is done redundantly that could be avoided.

(Next Slide)

Let us go through each of these four points. Let us begin with *objectivity*. Objectivity is inherent in the interpretation of evidence by computer using Bayesian inference of genotypes because we begin with the DNA evidence, interpretation occurs, and then we end up with the DNA evidence genotype. We have never seen the suspect. We have only seen crime scene DNA evidence. It might be one sample or it might be 10. It might involve elimination samples or victims, but whatever it is, it does not involve suspects.

(Next Slide)

Then, a match is made to suspects: one, two, ten, a hundred thousand. It does not make any difference. We have already locked that genotype into the database of what the answer was. As we see on the right, when that match is made and when we combine an evidence genotype together with the suspect genotype and a population genotype, we have a DNA match score (or a likelihood ratio), and the result is an objective computer interpretation of the quantitative data.

(Next Slide)

Let me move to the next point, which is *ease-of-use*. In TrueAllele, what a user does is ask questions. In a routine production process, like volume crime, we can upload 200 questions simultaneously from one place, ask the computer to solve everything for one and two unknowns from one amplification, and the system happily goes off and processes. For casework, workers typically prefer to instead maybe take two or three minutes and ask a question about the data in a case. We certainly do this when we get into the data. If it is an interesting case, such as 40 pg of DNA, then we often ask several questions to see how much information we can get. After the DNA comes off of the sequencer, it is uploaded by plate into a database.

We open the Request interface of the VUIer program and say “please give me all of the data that is associated with our case.” We enter the case name, and the

computer finds our data (first panel on the left). The top two rows, coded by the by the user and the computer interface with the darker blue, show ProfilerPlus and Cofiler as evidence items. Underneath is the suspect, which is irrelevant for this purpose and who is not involved in the evidence request. The user then says “I would like to analyze an item of evidence.” We see in the middle panel on the top that there is an icon for the DNA for the evidence. From there, the user says “in this case let us analyze it as having two unknown contributors.” There are all sorts of questions we could ask, and this is what we see. We might also want to analyze a suspect so that the matching within the case after genotype inference happens automatically. A trained user would do this in about two minutes, and if it is an interesting case, then we have to stop people from going back and asking more questions.

TrueAllele is a fun calculator. The computer goes off and processes. How does it do that? We run parallel systems in our office. Our machine has 24 processors, and we are adding another 12 next week because we are doing more validation projects. How long does it take? Every day we crank out about 300 cases from the 24 processors, which are typically mixtures, low-level, and interesting things. That tells us how to schedule our work. We solve more cases with parallelism because computers are cheap.

(Next Slide)

When we are done, the VUIer™ interface can show us the data and answers in different visual ways. This is what the mixture distributions look like. We see a 0 to 100% scale on the x-axis. Here are two contributor distributions that are about 40% and 60%. The minor contributor is in blue, and we will continue looking at that.

(Next Slide)

The genotype was also inferred. What we are seeing in every row is another locus, and its probability distribution is to the right. In all mixture interpretation, anything that is uncertain is always interpreted to produce a probability distribution. What else could it be? It would not be science if it were anything else. Certainly, RMNE, CLR… — everything produces a probability distribution over the allele pairs.

Here we see a genotype. When we get a lot of bars to the right, there is little certainty in one solution. However, when we are getting closer on the 0 to 100% probability scale to one bar, a definite allele pair is trying to come out. It typically means we are going to get more information, and it was an easier sort of case to resolve with a computer.

(Next Slide)

On this minor contributor, I am going to quickly describe this again. This is a picture where we see the base pair on the x-axis and the *rfu* on the y-axis. The data are in black for D13, and there is a cartoon drawn over it by the computer. All of the other decorations, the lettering, and so on is just PowerPoint that illustrates what the genotypes, the allele pairs in D13 are. What the computer is showing us is what happens with roughly 60:40 of the orange contributor to the blue contributor with those allele pairs. We can do a “what-if” analysis and choose different genotype or mixture values. The computer is considering everything that could possibly exist in its calculation.

(Next Slide)

Then it brings in the stutter variable. We can see the stutter peak all the way on the left. We can see the relative amplification where the peaks are sort of skewing. TrueAllele creates a pattern that we see in gray, and it compares with the quantitative pattern. The computer does tens or hundreds of thousands of these proposals against the data, and those patterns that quantitatively better fit the data tend to have a higher likelihood and are given a higher probability. Notice that this might be very visually evident to people and to a quantitative computer, but if we had a threshold much over 100 rfu, then we would start throwing out the information we were looking for.

(Next Slide)

We have another interface that does a lot of reporting. It gives us a snapshot. We use it for validations. We sometimes will be looking at a hundred comparisons at once. On the top, we can see and set what theta value we want for co-ancestry. We can look at this on a log scale or on a linear scale. We can look at every locus individually or look at them jointly. What we are seeing here is any number of evidence genotypes against any number of controls with any number of populations.

To simplify, let us look at the one minor contributor genotype that we just saw. For the total likelihood ratio (log likelihood ratio information), we will have two different comparisons: a suspect and an elimination profile. The elimination genotype does not match. It is moved over to the left. The scale over here is -20 to +20 log units. That, we might call an exclusion. On the right, we are seeing a log(LR) in the 15 to 20 range, shown for two different ethnic groups. Report is the interface where people have the most fun because they get to try out “what does the answer mean,” whether it is for a case or for a validation study.

(Next Slide)

In the Report interface, we can also generate detailed calculations and reports that are useful for the case folder. Here is a more detailed output. It gives a likelihood ratio statement for every locus. The report shows a lot of information,

including the genotype, which (of course) is a probability distribution over allele pairs. It gives the likelihood ratio both in linear and logarithmic form. The long form can be useful if we want to establish how a calculation was done.

(Next Slide)

I would like to move on to information and productivity, now that we have a sense of what the system looks like. In this study, we collaborated with New York State. We will see all of the different people who participated out at the end on the Acknowledgments slide. They gathered together a lot of data. It took about a year to gather the data. It took about another year to get administrative permission. It took another six months to get permission to submit the paper. This has been a very long study. There are 368 evidence items that range the full spectrum of what the New York State lab sees, as can be read here.

(Next Slide)

I want to quickly show some pictures of what the results are. What we are seeing here are eight cases. Each point on the x-axis is another case. These were cases that New York State looked at as a two-person mixture, and they had a victim profile. We will see in a second what they did with it, but they did a CLR analysis assuming the victim. The vertical scale is information, or log likelihood ratio, going up from 0 to 25. Each of these cases was done in duplicate. Blue is

the first computer run, and green is the second computer run. We measured reproducibility. John Butler has shown that, with US labs, reproducibility on two-person mixtures ranges over ten orders of magnitude. When we consider population variation, it typically varies by plus or minus one log unit in the US (as opposed to ten). The variation that we see with TrueAllele and simple two-person mixtures like this is on the order of *hundredths* or *tenths* of a log unit. So, it is a very reproducible process.

(Next Slide)

Now let us see how analysts can lose information. In blue again, on the same scale for the same eight cases, is a one unknown two-person mixture. This is what TrueAllele did for each case with an average of about 10^{17} likelihood ratio. Human review, using the usual sorts of thresholds and equal likelihood assignment to allele pairs instead of basing it on quantitative data, lost about four and a half log units. We see case-by-case the information drop to where the orange bars are.

(Next Slide)

When we go to two unknowns (this is RMNE, what people were doing), again these are all of the cases for which New York State reported a statistic. I will change that assumption in just a second. So, these are eight different cases

where they did *not* have a victim reference, and there is less information. We can see that the computer dropped from maybe 17 log units to 13. However, the human review lost over six log units relative to the gold standard of a quantitative Interpretation. We can see this by looking at the average drop from the blue bars of the computer case-by-case to the human review. Typically, 10^6, 10^7, which we see on the right in orange, is what a normal RMNE match score would be like in the US on 13 loci. The information loss here is due to the fact that we are not using a victim genotype and that we are using equal likelihood assignment to all possible allele pairs, which diffuses the probability. Now we have seen that when New York State got a match score, human review lost more information as the cases became more interesting.

(Next Slide)

What happens when we look at all 85 or so mixtures in this case? This is probably the most interesting slide from the whole study, and so I am going to spend a minute or two on it. What we are seeing on the x-axis is a listing of the 85 evidence items, ranked by decreasing information (that blue edge) as determined by TrueAllele interpretation. Sorted from the left with the most information, 10^{20}, going down on the right to maybe 10^3. The median information value in the center is about 10^{15}, or 10 quadrillion to one, as a likelihood ratio. The height of the curve is telling us what the log likelihood ratio information is for each interpreted case. The area of that blue region is telling us

what the total information extracted across all of the cases for the study was.

Now, what people achieved is shown in these bars. The best results are the four gray RMNPs where we apply thresholds, pretend the mixture is single source, and get the full answer. In green, we then had eight cases where the victim was known. The CLR was done and yielded a reasonable amount of information. Then, there were 12 in orange that were RMNEs or CPIs where they lose more information.

So, here are the two most interesting facts about this particular picture. The first is that if we look along the x-axis and ask for all of the cases for which the computer got an answer, what fraction of the mixture items that were done was there any reported likelihood ratio result at all? The answer is less than 30%. We are going to talk about that in second, but that implies that there is a whole host of information up there not being used. The second question is, what was the total amount of information found by human review? That is the area of those colored human review curves. That is less than 20% of the total area of what the computer found. If information is one of our main goals, then it is not being completely captured by current review processes.

(Next Slide)

We can start asking what happened. The computer, shown in blue again, always

gets an answer. We look at cases where they look at one mixture item, two, three, four, five, and so on. When they only looked at one item, about 75% of the time they would get an answer. Then, as it goes out, they get less and less information.

(Next Slide)

Let us take a look at why this is happening. New York classified the cases by degree of difficulty into simple, medium, and complex. Simple being a two-person mixture with a known victim, medium was two unknowns without any reference, and complex were things that went to three or more contributors, partial profiles, and so on.

(Next Slide)

Here are the productivity results. What is the extra effort that people are going through on data that would have been solved if they were using a genetic calculator? Let us look at the simple cases first (in green). There were 35 of these. The computer found about 10^{16} on average, which is the log likelihood ratio information. It had an answer for all of them, and the human yield on these is about one in two, or about 50%. This means that the expected extra effort that we can compute is about two to one, as it turns out. How many items do we have to analyze until we get an answer? If there is a 50% yield, then we are doing

twice as much work as we have to.

When we move to the 20 medium cases (in the middle column in purple), then the computer is averaging about ten trillion, or 10^{13}. It is getting answers for everything, and people are finding an answer for one in four on these two-person mixtures without a reference and done by RMNE. They have to do four times the work that would be needed based on the data that they already have. When we go to the 33 complex items in the third column, the computer is now down to only a trillion to one, or 10^{12}. The human yield drops to one in five, meaning that we expect to do about five items before we get an answer. This results in spending five times the amount of time and money on reagents, the delay, and all of the usual costs that go with processing samples.

If we look at the 88 total, the computer averaged about a hundred trillion to one, or 10^{14}. The human yield was just under 30%, and about three and a half times the effort was needed on average. That is the take home number from the talk — on this very diverse data set on mixtures, people are expending three and a half times the amount of effort to get some match statistic on an item than they would need to if they were instead using a quantitative genetic calculator. Notice, of course, that as the spectrum of cases shifts to more interesting samples, we are going to be closer to five times the effort than we are to three and a half.

(Next Slide)

Who benefits from all this? Probability is always done in service to utility, starting with gambling 300 years ago and moving to economics and such today. From the laboratory's perspective, it is nice to reduce costs and increase efficiency and so on. From society's perspective, it is good to get the maximal expected utility from forensic DNA to reduce crime and increase public safety. They are different objective functions, but finding information is good for both.

(Next Slide)

In conclusion, there is a rapidly growing increase in DNA demand from society, which taxes the lab with workload complexity of DNA, but there is a far smaller increase in the DNA capacity. As we know, the resources are not growing commensurate with that demand, nor are the number of analysts. Automated computer interpretation of DNA evidence is one way forward. The metrics that we use that I discussed today were: a) objectivity, which computers have if we program them right; b) ease-of-use; c) information extraction from the available evidence; and d) productivity that is directly related to the extracted information

(Next Slide)

I would like to thank Jamie Belrose, who reviewed of all this data and collected it. Some of our group at Cybergenetics was involved in the project. More staff at

New York State Police gathered quite a bit of the data. We have a submitted manuscript that we submitted to the Journal of Forensic Sciences, and in the interim, you can download that if you are interested. (http://www.cybgen.com/information/pub_a19.shtml). It goes into much of what I discussed today, and probably most interesting, it goes through one detailed example showing locus by locus how RMNE loses information relative to quantitative DNA interpretation. Thank you very much.