Transcript of Dr. Mark Perlin’s lecture on “DNA Identification: Biology and Information” in the TrueAllele Basic Science Course.

Dr. Perlin: These lectures are about DNA identification. The underlying principle is that uncertain genotypes are probability distributions that are easily inserted into standard likelihood ratio (LR) formulas and are readily explained in court. We start by looking at DNA. What is its biology? And what do we mean by DNA identification information?

(Next Slide)

Virtually all of a person's cells have a nucleus that contains our DNA.

(Next Slide)

Inside the nucleus, the DNA is packaged into 23 pairs of chromosomes.

(Next Slide)

We can zoom into one location on a chromosome, which is called a locus. In forensic identification, we are interested in genetic loci whose DNA sequences exhibit very many different distinguishing features.

(Next Slide)

These different locus DNA values are called alleles. The short tandem repeat, or STR, locus contains alleles made up of repeating units. Each tandemly repeating unit is usually a sequence of 4 DNA letters, such as the "AATG" DNA word shown here. An allele value of 10 indicates a DNA sequence length of 10 units times 4 letters per unit, or 40 DNA letters. An individual has a pair of alleles at a genetic locus: one inherited from their mother and another one from their father.

(Next Slide)

An individual's genotype at a chromosomal locus is an allele pair, like the STR lengths shown here. The '7' designates seven repeat units inherited from one parent, while the '8' is for the eight repeating units from the other parent. At this locus, the individual's genotype is a [7,8]. With about 10 different allele lengths at a locus, there are roughly 10 times 10, or 100, possible pairs of alleles. Those 100 possible genotype values at one locus provide the very many different distinguishing features that are so useful in forensic identification. By combining 10 to 15 independent loci across different chromosomes, it becomes very unlikely that two unrelated people will share the same multi-locus DNA genotype.

(Next Slide)

Here is biological evidence – a bloodstain on a pavement. We want to make a biological identification using such biological evidence.

(Next Slide)

The DNA laboratory transforms biological evidence into DNA data in order to preserve identification information. First, the lab extracts the DNA material contained in the evidence. Then, it applies the polymerase chain reaction, or PCR, to amplify the small amount of evidence DNA into millions of fluorescently labeled fragments. A DNA sequencer measures each length of the DNA fragment to determine how many labeled copies were produced. The sequencer detects the amount of fluorescent light to form data peaks whose heights reflect the number of amplified DNA fragments. Here we see two such DNA peaks corresponding to STR alleles having 10 and 12 repeat units.

(Next Slide)

In the same way, the goal is to preserve identification information when moving to infer a genotype from the questioned data. The whole purpose of forensic science is to preserve all of the identification information that is present in the data. Of all the variables in the STR process, only the genotype will enter into the likelihood ratio match statistic, not PCR stutter, extent of DNA degradation, or the amount of mixture weight – just the genotype. For accurate statistical computing, every uncertain factor affecting the experiment data must be modeled as a random variable, but the most crucial part of DNA identification information is getting the most accurate genotype. Those are the two main steps. The lab needs to do its part in preserving DNA evidence into analyzable STR data, and the DNA interpretation method must do its part when inferring the most accurate questioned genotype.

(Next Slide)

Once we have inferred a genotype from the evidence, then a comparison can be made. This genotype comparison is automatically done for us in the likelihood ratio. We just put the results of a genotype inference into a likelihood ratio formula and out will come a match statistic. All match statistics are likelihood ratios. We see on the bottom right a comparison with a suspect. Notice that when we infer an evidence genotype the suspect is not known or used since it is completely irrelevant and meaningless at that point. We can never use the suspect genotype when objectively inferring a genotype from evidence since this would introduce bias against a defendant. In the statistical TrueAllele system, the question itself makes no sense. When solving for an unknown variable, we cannot assume the answer we are looking for. Indeed, whatever genotype assumptions we make are sort of subtracted away from the problem and leave us searching for other explanatory genotypes.

(Next Slide)

We all understand the meaning of probability. The probability of an event (or hypothesis or proposition) is our degree of belief in that event. The identification event happens when a suspect matches the evidence. We are interested in the probability of this event. We can generally assume that a suspect's genotype has a unique allele pair. This may not be true when doing kinship or inferring missing people, but in criminal mixture identification, we can assume that the suspects we care about have a single allele pair at each locus. If the evidence data are perfect and there is only one allele pair present at a locus in the questioned genotype, then the probability is one that the evidence allele pair is [10,12] since we have certainty about this event. Additionally, the suspect's genotype is [10,12] with certainty. Since each genotype is a certain event and their allele pairs agree, the probability of a match is one. We are certain that these two independent genotypes, the suspect and the evidence, match. So, we have made an identification. However, what if it is based on a highly common feature that everybody shares? The suspect may have a nose, but everyone has a nose, so this is not terribly distinguishing or persuasive evidence. To address this issue, we consider the alternative hypothesis: coincidence.

(Next Slide)

There is an entire biological population of people who have genotypes. How likely is it that someone else has the matching allele pair by coincidence?

(Next Slide)

To characterize the population distribution of alleles at an STR locus, laboratories generate allele frequency data. Some groups use tens of thousands of individuals in order to reduce statistical sampling effects from the population. However, American labs only sample a few hundred people, which increases the LR uncertainty ten-fold. A small population sample can further reduce an already weak DNA match statistic. Consider an inclusion mixture interpretation that gives its usual likelihood ratio match score of one million. Population under sampling can lose a log unit of likelihood ratio. A co-ancestry coefficient theta value of 3% can lose another log unit or two. After making these corrections, the originally reported million-to-one inclusion stat is really only a thousand-to-one. This no longer sounds like a very compelling match statistic.

(Next Slide)

From the allele frequency data, we can infer a population genotype. The definition of a genotype is a probability distribution over allele pairs. The use of probability lets a genotype capture and characterize the genetic data uncertainty. With uncertainty, the probability at any allele pair may not reach one. Known suspect reference genotypes usually have a probability of 1 for one allele pair at each locus, but population and mixture evidence genotypes generally do not. It is best to recognize that there could be uncertainty in any genotype we see and recognize that a genotype is a probability distribution over allele pair values. Probability provides a clean and entirely general genotype description. How do we construct a population genotype? We use the product rule for independent events that lets us multiply their probabilities together, and we assume Hardy-Weinberg genetic equilibrium. We are describing diploid human populations where individuals have two chromosome copies. A homozygotic allele pair, [10,10], has a p_2 genotype probability with p being the allele's population frequency. A heterozygote, [10,12], has 2pq genotype probability since the distinguishable alleles could have come from the parents in two ways. In our running example, let us use round numbers and say that the probability of population genotype [10,12] is 5%.

(Next Slide)

Now, we can look at coincidental matches. Let us compare that population genotype of a [10,12] allele pair having 5% probability with the suspect genotype of [10,12]. Why are we focusing on the [10,12] allele pair? This is because that is the genotype that the suspect has. Any other genotype allele pair value is not relevant to a likelihood ratio match score because the suspect does not have it. Should a different suspect have some other allele pair value, then that would be relevant for that other suspect. This distinction becomes important when testifying in court. Defense attorneys often ask, "What if it were a different allele?" Well, who cares? That is not relevant for the match statistic. Produce a person having that genotype, and we will happily generate another match statistic that is appropriate for that different suspect hypothesis. These "what-ifs" are entirely irrelevant to the question about this defendant, but, as we shall see, they do illustrate how probability can help quantify uncertainty for the court.

(Next Slide)

Now we make a comparison between the suspect and the population. The bottom left describes a coincidence event of the suspect matching the population. The probability of this coincidental match is 5%. Here is why. The suspect has a definite [10,12] with probability of one. Independently without accounting for co-ancestry, the probability of a random person having this [10,12] genotype is 5%. The product rule of probability for combining independent events lets us multiply the suspect genotype probability of 1 by the population genotype probability of 5%, which gives us a coincidental match probability of 5%.

(Next Slide)

We want match information. At the suspect's genotype, is the match an identification or a coincidence? This probability ratio turns out to be one of the many forms of writing down the likelihood ratio. We can start with the probability beforehand, which is in the denominator, of what could occur in the population by coincidence. After seeing the evidence data, we assess the probability of the evidence matching the suspect, making an identification along with our uncertainty. The probability ratio of an evidence identification numerator divided by a match coincidence denominator is one way of writing down the likelihood ratio. The logarithm of the likelihood ratio is a standard measure of information in fields that use probability. Such fields include computer science, astrophysics, forensic science, sociology, electrical engineering, economics, and genetics. If we can express a hypothesis and its alternative in probability terms, then we can start writing down a likelihood ratio. In this case, the hypothesis in the numerator is that the suspect contributed to the evidence. The alternative hypothesis is that his genotype is coincidently matching someone else in the population.

(Next Slide)

It turns out the word “matches” can be used just as well in the event statement. This fact is important because the match concept really has some tangible meaning to most people. It is something that a jury can readily understand. If we say that a match to one thing is more probable than a match to another thing, then we are now speaking plain English that has some intuitive meaning.

(Next Slide)

Let us now compute the information number. The numerator is 100%, while the denominator is 5%. So the likelihood ratio is 20. This tells us how much weight of evidence is being provided by this particular locus experiment. Like mass or energy, the amount of information is an additive quantity. Therefore, we use logarithms to add information instead of multiplying the likelihood ratios. The logarithm of likelihood ratio 20 is around 1.3 in base 10, which is our information number. We have looked at just one out of 13 or so STR loci tested in a typical DNA multiplex experiment. Adding up the likelihood ratio logarithm values at all 13 loci gives the total weight of evidence. With DNA identification, each independent STR locus typically contributes about 1 log unit. Adding together those separate values gives a total of 13 information units for a joint likelihood ratio of 10^{13}, or 10 trillion.

(Next Slide)

Suppose our evidence is a mixture of biological material from two people. Here, we see a happy perpetrator and an unhappy victim. With mixtures, there is often less identification certainty in the DNA evidence than there is with an unambiguous reference sample. A loss of genotype certainty usually translates into a loss of identification information.

(Next Slide)

When a laboratory generates DNA mixture data, we usually no longer see just one or two allele peaks. Instead, as shown, we see peaks for whatever genotype alleles may be present in the underlying mixture. We can also see artificial peaks that do not correspond to any contributing genotype.

(Next Slide)

We are no longer certain about the exact genotype allele pair composition of each mixture contributor. The evidence data simply do not contain enough information to make a unique determination. When we infer a questioned genotype at a contributor's locus, we may now have more than one allele pair possibility. To express our belief in each possibility, we assign a probability to each one. We will discuss how to determine such probabilities later on. Suppose we believe that the contributor genotype we are inferring at this locus must contain allele 12. Here are three allele pair candidates that are each assigned a probability value. Since we are working with probability, these values must add up to 1 (or 100%). We have implicitly assigned the other hundred possible allele pairs a probability of zero. This probability distribution over allele pair values constitutes the evidence genotype. Once we have inferred this genotype never having looked at a suspect, we can now objectively make a comparison to some suspect.

(Next Slide)

Here is that comparison. We see a match at allele pair [10,12] and can now compute a likelihood ratio match statistic.

(Next Slide)

What identification information did we obtain by examining the evidence? At the suspect's genotype, did we make a real identification or are we seeing a coincidence? We can look at the probability of identification after we have seen the data divided by the probability of coincidence before seeing the data. That ratio of probabilities is our match statistic, or the DNA likelihood ratio. The arrow here shows the direction of information gain through the data. It reminds us that learning from the data is an expression of Bayes' theorem, which underlies the likelihood ratio. The likelihood ratio quantifies how much information the data gave us when changing the likelihood of support for (or against) a pair of mutually exclusive and exhaustive match hypotheses.

(Next Slide)

When computing the likelihood ratio, we can use the match concept together with a genotype probability. The evidence genotype probability was only 50%. The probability of an evidence match in the numerator is now halved from 100%

down to 50%. We still have the same denominator, which is a coincidental random match of 5% probability. So, the likelihood ratio dropped from 20 down to 10. By dissipating genotype probability, whether inherently in the evidence DNA, in laboratory data generation, or in genotype interpretation methods, the likelihood ratio can fall from quadrillions down to thousands. Our primary goal as forensic scientists is to preserve identification information, and report an accurate likelihood ratio for the strength of match. Thank you.