STRmix™ Training

Indiana State Police Laboratory
December 2017

Introduction

- On November 1, 2017, the Indiana State Police (ISP) Laboratory Biology Section began utilizing STRmix™ probabilistic genotyping software to aid in the interpretation and statistical evaluation of DNA profiles including some previously uninterpretable mixed DNA samples. This training will explain why we have chosen this new method, how the new statistical results will be reported and how they differ from previously reported statistics. Special attention will be given to how this will affect testimony.

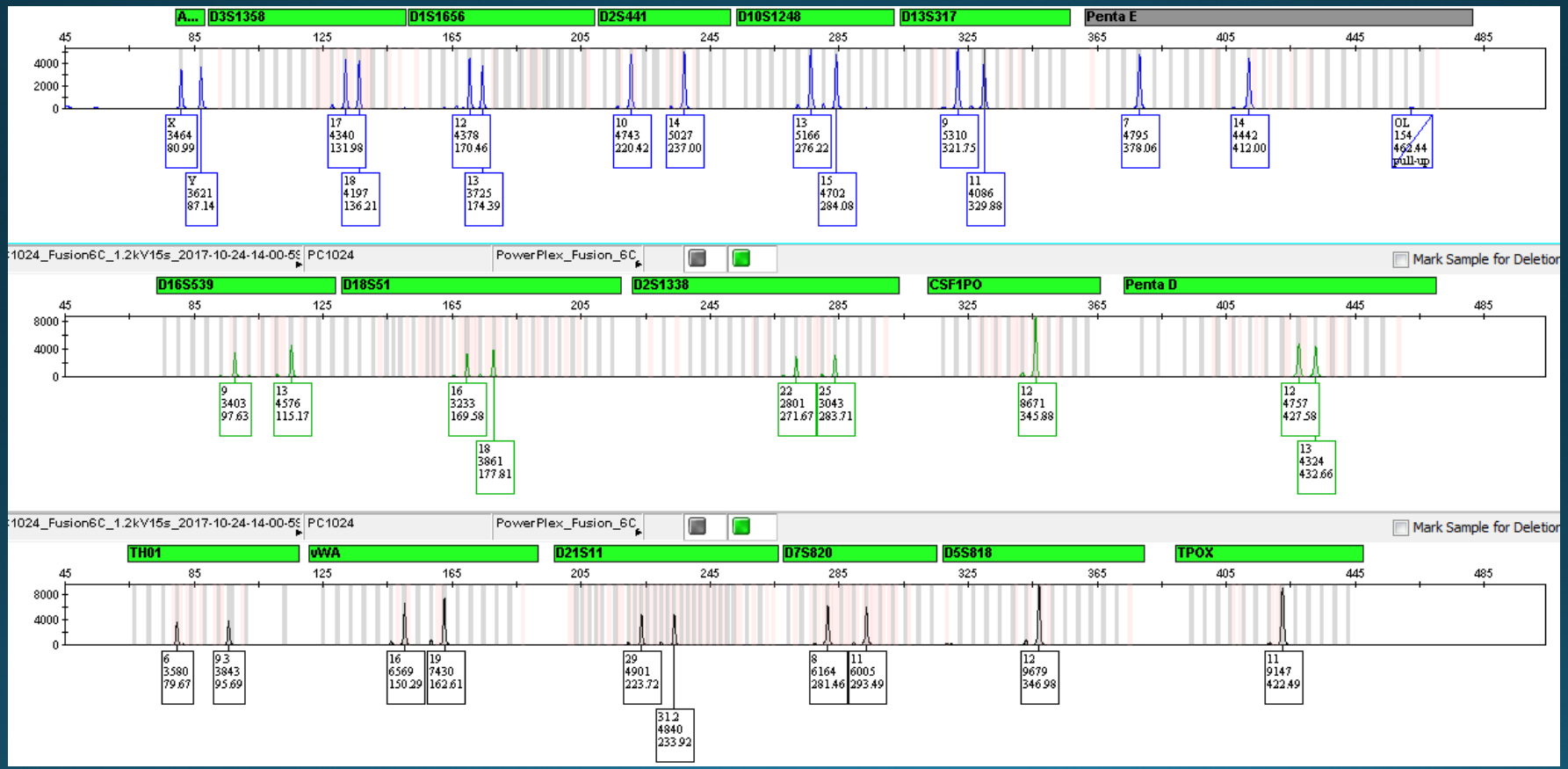
Overview

- Motivation for change
  - Basic terminology & core concepts of DNA profile interpretation
  - Old method: manual interpretation
  - New method: probabilistic genotyping & STRmix™
- Statistical Evaluation
  - Likelihood ratios
  - Case information and relevance for each item submitted
- Testimony
  - Communicating likelihood ratios
  - Laying a foundation – demonstrating scientific validity & reliability
  - References

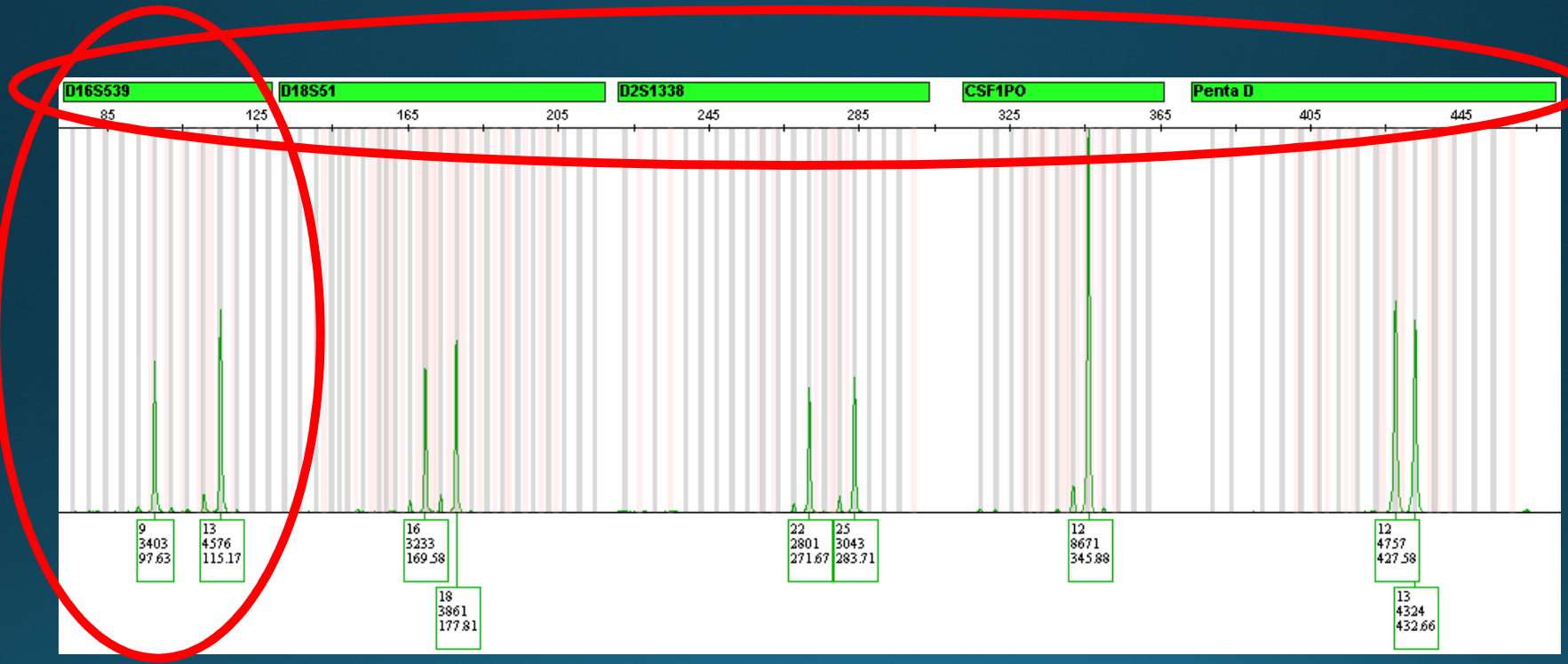
STRmix™ Training

Motivation for Change

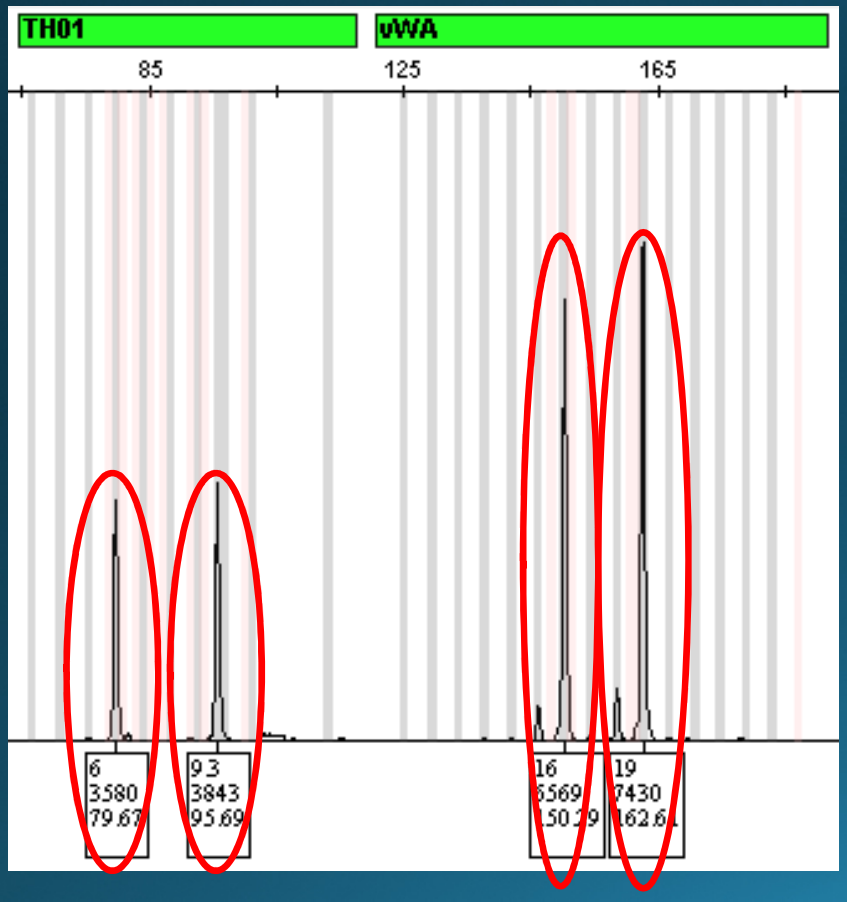
Electropherogram: picture of the DNA



Loci (Locus): segment of DNA

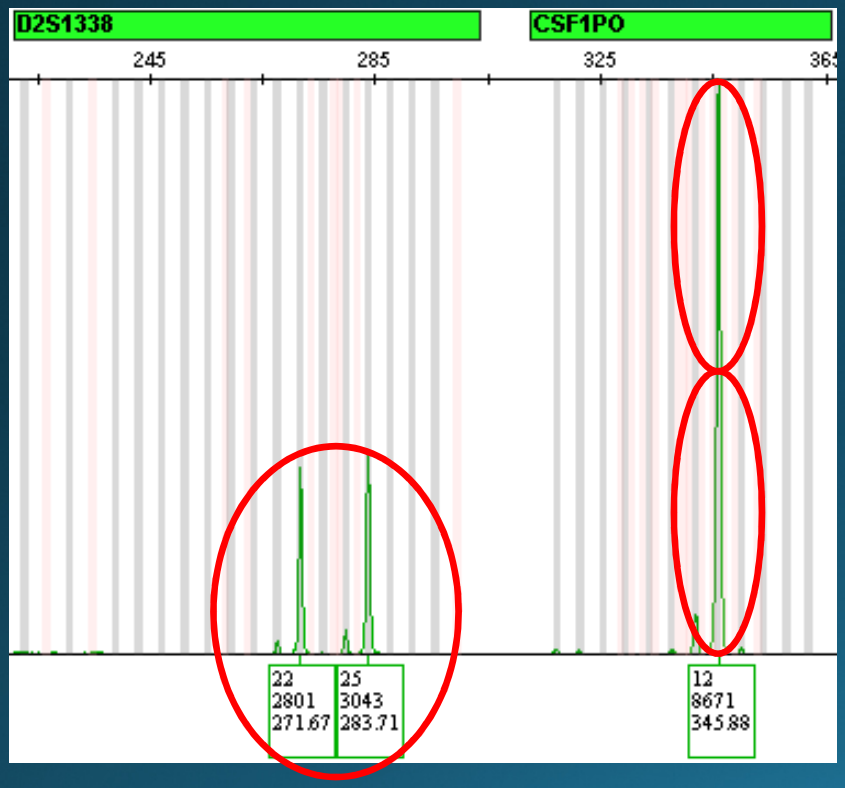


Allele: different copies of a gene



- 2 Loci: THo1 and vWA
- Alleles at Tho1: 6, 9.3
- Alleles at vWA: 16, 19

Genotype: set of alleles



- Loci: D2 and CSF
- Genotype at D2 = [22, 25]
- Genotype at CSF = [12, 12]

The biological process that leads to a DNA profile: PCR

- The PCR process allows us to make lots of copies of only the segments of DNA in which we are interested
- This process is not perfect
- DNA could be degrading, so that we are getting lots of information from smaller pieces, but no information from the larger pieces.
- There will be different amounts of DNA from each person, which means that one person may be hidden by the other
- The process itself creates artifacts that can interfere with analysis

Profile Interpretation

- Number of contributors
  - How many people do we think contributed DNA to the sample?
- Mixture proportions
  - If there is more than 1 person, how much do we think each is contributing?
- Genotype combinations of each contributor
  - How do those alleles pair together for each individual in a profile?

How do we make these determinations?

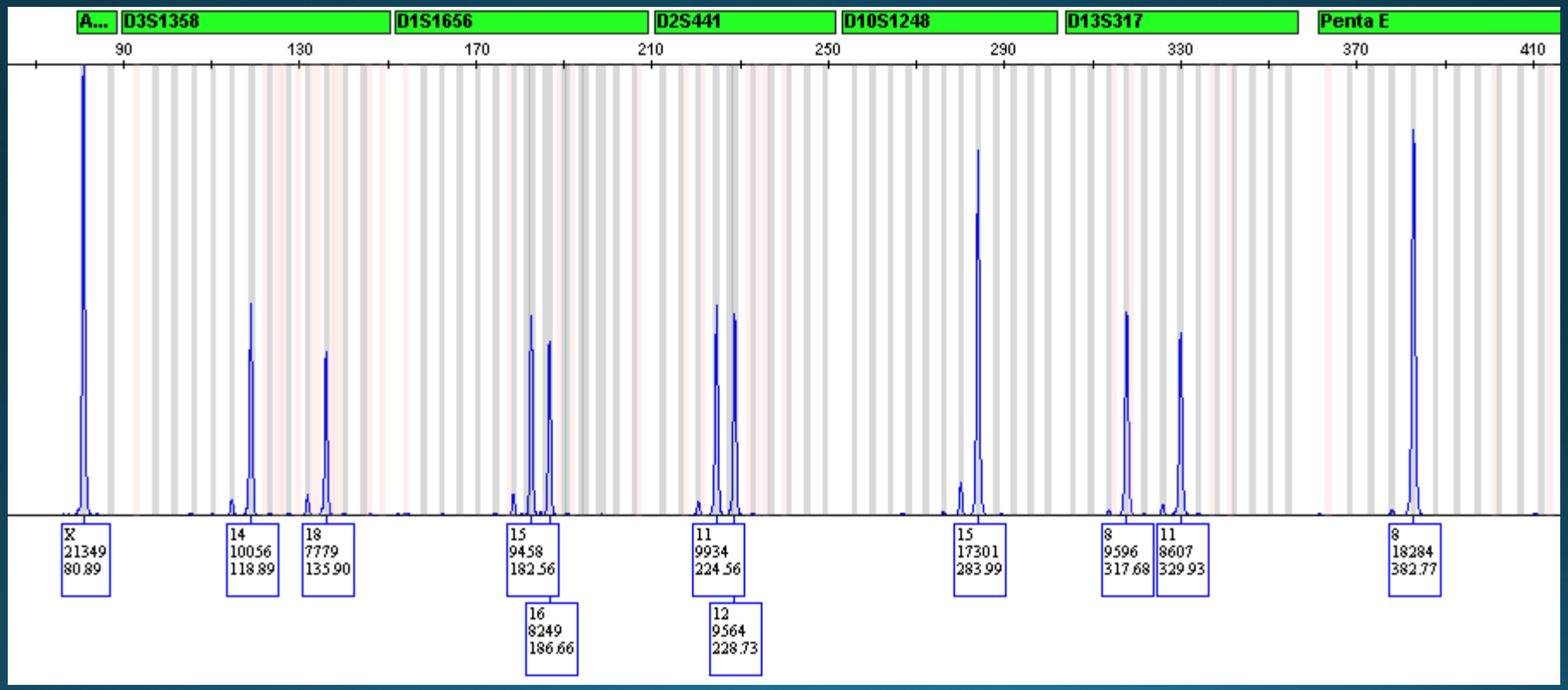
Core Concepts of DNA Interpretation

As long as we have been using the CE platform, we have known certain things about the way DNA behaves:

- Peak height is proportional to the amount of DNA
- If alleles are present from different sources, they add "stack" onto each other
- There is more variability with lower amounts of DNA

The goal of any DNA interpretation is to determine the *possible genotypes of the individual contributors*

Single Source – all from 1 individual



Summary Sheet

Evidence Profile

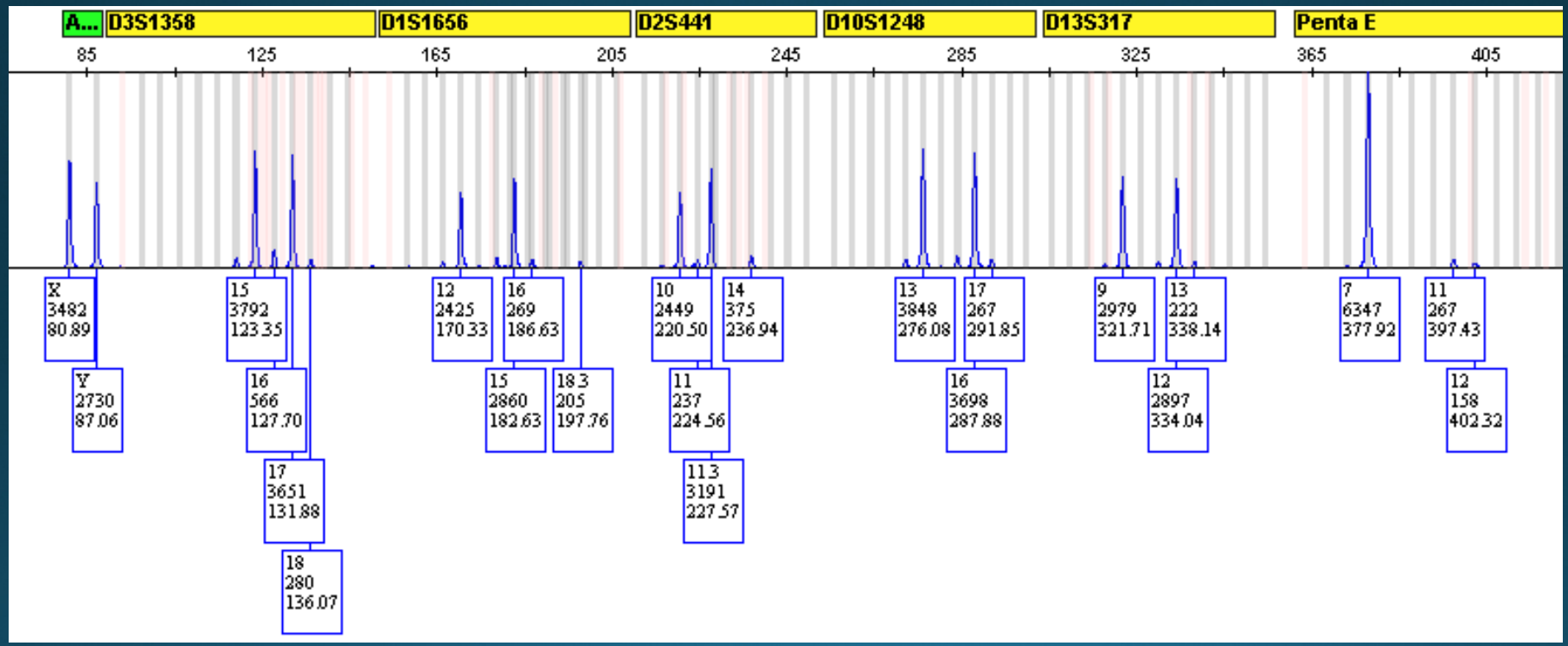
Alleles Detected	
Amelogenin	X
D3S1358	14, 18
D1S1656	15, 16
D2S441	11, 12
D10S1248	15
D13S317	8, 11
Penta E	8

Reference Profile (Known Person)

Sample	001A
Description (Name)	Jane Smith
Amelogenin	X
D3S1358	14, 18
D1S1656	15, 16
D2S441	11, 12
D10S1248	15
D13S317	8, 11
Penta E	8

- Sir
- Sir
  - tus
- Str arison

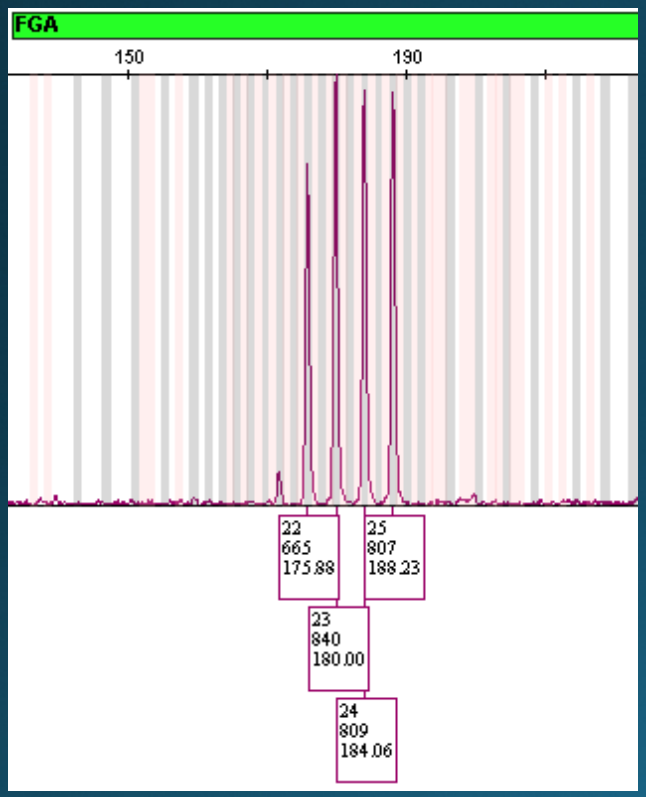
Mixture – 2 Person, Major/Minor



Genotype Combinations

	Mixture Alleles				Major Profile	Minor Profile
Amelogenin	X	Y			X, Y	-
D3S1358	15	[16]	17	[18]	15, 17	[16], [18]
D1S1656	12	15	[16]	[18.3]	12, 15	[16], [18.3]
D2S441	10	[11]	11.3	[14]	10, 11.3	[11], [14]
D10S1248	13	16	[17]		13, 16	[17], A
D13S317	9	12	[13]		9, 12	[13], A
Penta E	7	[11]	[12]		7	[11], [12]

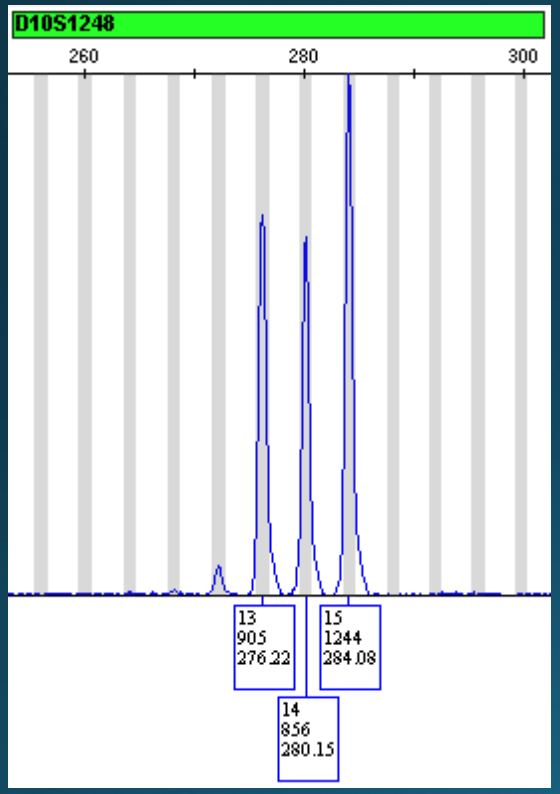
Indistinguishable Mixture



FGA
22
23
24
25

4 alleles		
22, 23	Y	24, 25

Indistinguishable with Sharing



D10S1248
13
14
15

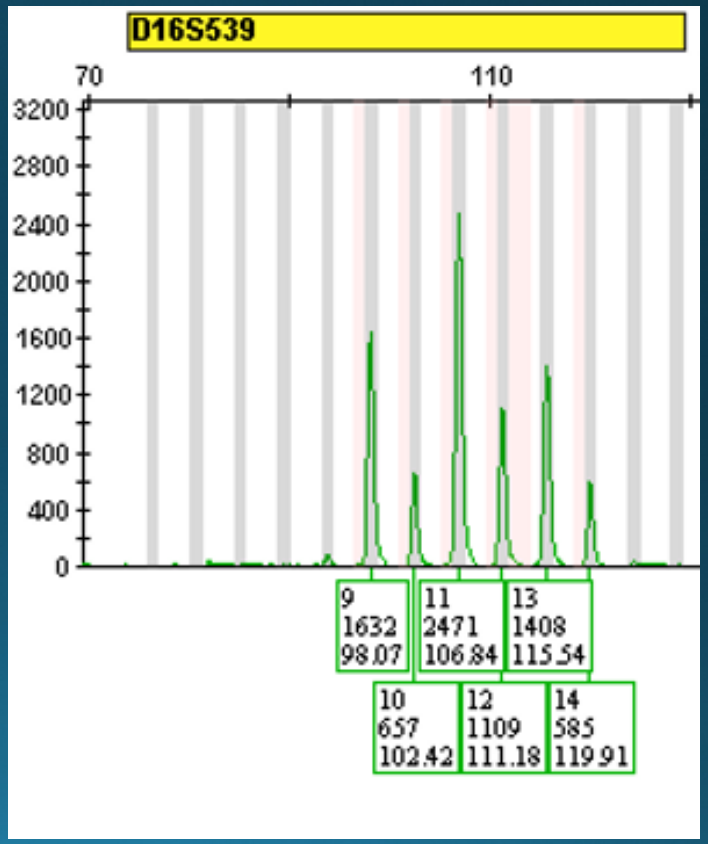
3 alleles		
13		14, 15
14		13, 15
15	Y	13, 14
13, 14		13, 15
14, 15	Y	13, 15
13, 14		14, 15
13, A		14, 15
14, A		13, 15
15, A		13, 14

Limitations of manual interpretation

- With our manual method, we used our judgement to account for variability.
  - Generally we include more genotype combinations with low level samples, and less combinations with higher quantity / higher quality samples.
- The problem with this approach is that it is hard to be consistent between samples and between analysts.
- Our previous methods were limited, because we didn't have a way to standardize each interpretation.
- Often low level samples and complex mixtures were deemed inconclusive.

Low level or complex mixtures

- Assuming 3 contributors (no allelic drop-out)
  - 9,10 11,12 13,14
  - 9,10 11,13 12,14
  - 9,10 11,14 12,13
  - 9,11 10,12 13,14
  - 9,11 10,13 12,14
  - 9,11 10,14 12,13
  - 9,12 10,11 13,14
  - 9,12 10,13 12,14
  - 9,12 10,14 11,13
  - 9,13 10,11 12,14
  - 9,13 10,12 11,14
  - 9,13 10,14 11,12
  - 9,14 10,11 12,13
  - 9,14 10,12 11,13
  - 9,14 10,13 11,12



Enter Probabilistic Genotyping

- In the last few years, new computer software options have become available to help standardize interpretation.
- Using a process of random sampling, computer programs can now test hundreds of thousands of possibilities.
- Probabilistic genotyping uses complex math and what we know about the way DNA behaves to determine reasonable genotype combinations of possible contributors to a DNA profile.
- Simply, probabilistic genotyping is using computer software to separate out different components of a DNA profile.

What is STRmix™?

- Fully-continuous Probabilistic Genotyping software
- STRmix™ uses scientific understanding of biological processes to build conceptual profiles
- It then grades those profiles by how closely they match the observed profile
- STRmix™ assigns high probability to profiles that best represent the observed data and a low probability to profiles that do not
  - Not a simple yes/no as used in manual interpretation

STRmix™ Interpretation

A short video explaining STRmix is available at:

https://strmix.esr.cri.nz/

Example STRmix Results

GENOTYPE PROBABILITY DISTRIBUTION

LOCUS	CONTRIBUTORS		WEIGHT
	1 (87%)	2 (13%)	(HIGHLIGHT ≥ 0.99)
D3S1358	15, 17	15, 18	4.42817E-1
	15, 17	17, 18	3.11119E-1
	15, 17	16, 18	2.32470E-1
	15, 17	15, 17	5.84032E-3
	15, 17	16, 17	2.28199E-3
	15, 17	15, 16	2.11842E-3
	15, 17	15, 15	1.86183E-3
	15, 17	17, 17	1.07971E-3
	15, 17	18, 18	3.40799E-4
	15, 17	14, 18	5.90777E-5
	15, 17	16, 16	1.19037E-5

Determining Weights

- Uses a process similar to a game of "Hot/Cold" or "Battleship"
- It makes a random guess of what the profile could look like, given the possible genotype combinations
- It will guess values for how much DNA is present, how the sample is degrading, and the efficiency of the reaction
- Using these values (parameters) it will build a conceptual profile
- The software then compares this conceptual profile to the observed profile
- If the guess explains the profile well it is a good guess "Accept"
  - Weights are determined by the percentage of "Accepts" with a genotype set

Biological Modelling

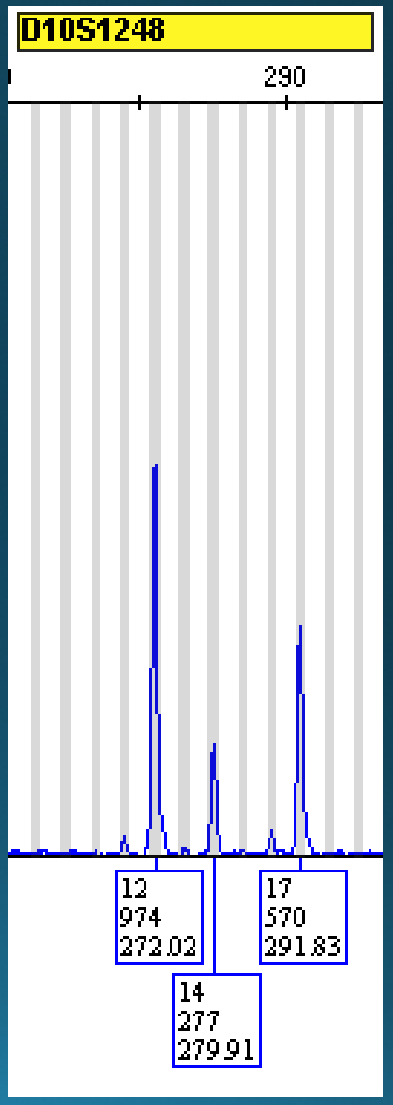
- Probabilistic Genotyping systems use biological modelling to determine reasonable genotype combinations.
- A model is a mathematical description of a biological process that leads to a DNA profile
- The models used by STRmix™ account for the following parameters
  - Genotypes for each contributor
  - Amount of DNA per contributor
  - Degradation of each contributor profile
  - Amplification efficiency for each locus
  - Amplification artifacts (stutter)
  - Peak variance

Genotype Combinations

3 alleles			MR
12	Y	[14], [17]	1.14
[14]		12, [17]	5.57
[17]		12, [14]	2.19
12, [14]	Y	12, [17]	
[14], [17]		12, [17]	
12, [14]		[14], [17]	
12, A		[14], [17]	
[14], A	Y	12, [17]	
[17], A	Y	12, [14]	

STRmix™ Weightings

0.16
0.14
0.05
0.48
0.12
0.003
0.0003
0.04
0.001



Benefits to STRmix™ Analysis

- Uses more information in the profile to make better determinations about reasonable genotype combinations
- Increases consistency between samples and analysts
- Allows us to interpret more complex samples
- Provides statistical weight to a profile within the context of the case – based on the relevant question to the court

Moving forward

- Using STRmix™ we will be able to interpret low level samples and mixtures of up to 4 contributors
- Each analyst will be reviewing casework dating back to August 1, 2016
  - Property crimes will not be reviewed or reanalyzed without request
  - STRmix™ was validated with our current kit chemistry, reanalysis of evidence requires TL approval
- If samples are identified that could be run in the software, we will automatically reanalyze and issue supplemental reports
- This should be completed by March 2018
- If you have a specific case that you are interested in, please contact the reporting analyst with questions

Other Probabilistic Systems

- Probabilistic Genotyping describes a class of software programs, some semi-continuous, others fully continuous
- We chose STRmix™, but you should know there are others that are used in forensic DNA analysis
- ISP will not reanalyze samples that were previously outsourced and interpreted with other probabilistic genotyping software unless specifically requested and approved by Laboratory Command (per policy).

STRmix™ Software

- Deconvolution
  - Using the core concepts of interpretation, determines possible genotype combinations for each of the proposed contributors
  - Calculates the probability of each combination – these are called "*weights*"
  - Weights are established prior to comparison to persons of interest
- Statistical Evaluation
  - Using the weights determined in deconvolution, STRmix™ calculates the likelihood of the observed profile given two competing hypotheses (propositions)

STRmix™ Training

Statistical Evaluation

Comparison of Statistics

- Previous statistic: Random Match Probability (RMP)
  - Estimates how often you would expect to find someone in the general population who could have contributed this profile
- New statistic: Likelihood Ratio (LR)
  - Determines the weight of a profile given two possible explanations for the origin of the DNA (propositions / hypotheses)
  - Propositions will be what most reasonably describes the evidence
    - Based on the profile itself along with relevant case information
  - Several different propositions can be tested, but the court must decide which of those proposition sets best represents the issue of interest

How are these statistics similar?

- Both provide weight (*meaningful information*) to a DNA profile
- They are a way to distinguish between profiles with a lot of information and profiles with very little information
- They are only estimates
  - In both cases, the actual values will vary slightly

How are these statistics different?

- With random match probability (RMP), the probability of a profile is *completely independent* of the possible contributor(s) or prior information
- Using likelihood ratios (LR), the probability of a profile is *entirely contingent* on the possible contributors and the specific hypotheses that have been proposed
  - *In an LR: the numerical value WILL change if you change the propositions*
  - *For this reason, we want to explain how we set those propositions, and how important case information can be in this process*

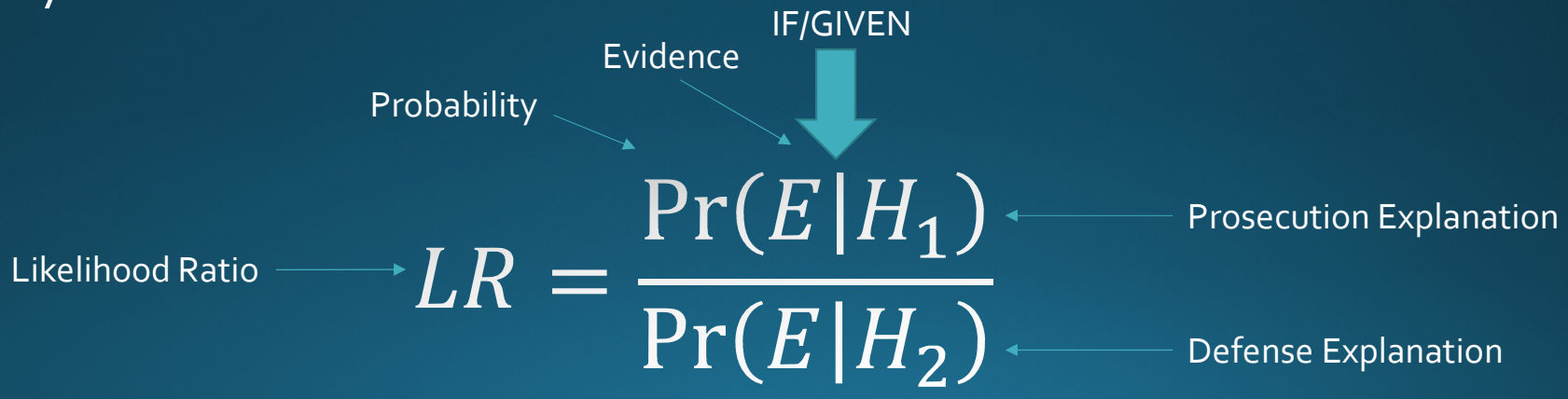
Constructing Propositions

- H_1 Inclusionary Proposition (Prosecution Hypothesis)
  - Explains the prosecutor's most reasonable explanation of the evidence
  - Includes the person of interest
  - *H1: The DNA profile originated from John Doe*

- H_2 Exclusionary Proposition (Defense Hypothesis)
  - Explains the defense's most reasonable explanation of the evidence
  - Excludes the person of interest
  - *H2: The DNA profile originated from an unknown, unrelated individual*

Likelihood Ratio in DNA

- Evaluate the evidence (E) relative to competing hypotheses
- Should be what most reasonably describes the evidence
- Several different hypotheses can be tested but the equation is always the same:



What does the likelihood ratio mean?

- The likelihood ratio is the relationship between the two probabilities
  - LR greater than 1 is support for H_1
  - LR equal to 1 is neutral regarding H_1 and H_2
  - LR less than 1 is support for H_2

Constructing Propositions

- In order to determine appropriate propositions, we will consider the following information:
  - Where was the item located?
  - To whom does it belong?
  - Is it reasonable to expect an individual's profile on the item?
  - What question(s) could be answered by DNA?
- Propositions must be mutually exclusive in order to calculate
  - *If all parties agree that the DNA profile originated from the same individual(s), the likelihood ratio would equal 1. A likelihood ratio of 1 is uninformative.*

Why is case information important?

- Likelihood ratios are driven by the major contributor(s).
  - If an item is located in the victim's home, and her DNA is present on that item, that information is probably not helpful to the court / trier of fact
  - Most likely, court interested in a possible foreign contributor (someone *other than* the victim)
- When constructing a conservative likelihood ratio, it is often important to remove the portion of the profile which is not in dispute and focus the calculation on the portion which is in question.
  - In the above example, including the victim in both H1 and H2 will generally result in a lower LR and better representation of the portion of the profile *someone else* could have contributed

STRmix™ Training

Testimony

Communicating Likelihood Ratios

- Likelihood ratios can be easily misconstrued, so it is important that we stick to specific statements in order to accurately reflect what has been calculated
  - Likelihood ratios are only relevant to the specific propositions which have been calculated
  - Remember that changing the propositions will change the numerical value
- If asked to summarize, we will most likely simply restate our conclusions as written in the report
  - Attempts to simplify could misstate the value

Reporting Likelihood Ratios

The DNA profile developed from the item was interpreted as originating from a single individual. The probability of the evidence has been calculated by considering the following propositions:

H1: The profile originated from John Doe.

H2: The profile originated from an unknown, unrelated individual.

The DNA profile is at least 1 trillion times more likely if it originated from John Doe than if it originated from an unknown individual. This analysis provides strong support for the proposition that John Doe is a contributor to the DNA profile.

Results that support exclusion

- An LR less than 1 supports the exclusionary proposition (H2), but decimals can be difficult to understand when reported
- LRs less than one will be inverted and the propositions will be re-ordered
- Example: an LR value of 0.1 would be reported as $\frac{1}{0.1} = 10$
  - The DNA profile is 10 times more likely if it originated from an unknown individual than if it originated from John Doe. This analysis provides weak support for the proposition that John Doe is excluded as a contributor to the DNA profile.

Verbal equivalency statement

- Our verbal scale is intended to provide a sense of relative weight to a likelihood ratio value
- Numbers without context can be difficult to understand
- For example: batting averages can range from 0 to 1. If you just look at it as a number it seems like 0.3 would be a relatively bad number. However, when put in to the context of baseball, that is a great batter. The verbal equivalent gives context to the number presented.

Likelihood Ratio (or 1/LR)	Verbal equivalent
1 ≤ LR< 10	uninformative
10 ≤ LR < 100	provides weak support
100 ≤ LR < 1000	provides moderate support
1000 ≤ LR	provides strong support

Why does wording matter?

Consider the following conditional probability:

- The probability that an animal has 4 legs IF it is an elephant is very high
  - $\Pr(4\ legs \mid elephant)$
- The probability that an animal is an elephant IF it has 4 legs is fairly low
  - $\Pr(elephant \mid 4\ legs)$

$$\Pr(\boldsymbol{Results|Proposition}) \neq \Pr(\boldsymbol{Proposition|Results})$$

Common mistakes

- Stating the results as the probability of the proposition, instead of the probability of the profile
  - "It is 1000 times more likely that John Doe is a contributor."
- Likelihood ratios do not translate directly into the frequency of a profile in the general population
  - "Probability of observing this profile is 1 in #"
  - "If there are 5 million people in Indiana, we could expect # of them to also be included in this mixture."

General Acceptance of STRmix™

- Core concepts of interpretation are widely accepted in the forensic DNA community and have been used for years in manual interpretation
  - Peaks are approximately proportional to template (amount of DNA)
  - Contributions from two sources add (stack onto each other)
  - Variability increases as template decreases
- Numerous laboratories throughout the United States are currently either using or in the process of validating STRmix™
  - Including the FBI and USACIL

Standards and Controls

- In 2015 the Scientific Working Group on DNA Analysis Methods (SWGDAM) published guidelines for the validation of probabilistic genotyping systems
- STRmix™ maintains software and documentation version control with a requirement of training prior to use

Testing and Validation

- STRmix™ has undergone rigorous developmental validation and is internally validated by each laboratory utilizing it in casework, including the Indiana State Police, for their specific laboratory system (i.e. chemistry, equipment, and procedures).
  - Developmental validation and FBI internal validation are both published in peer-reviewed scientific publications
- The Indiana State Police validation of STRmix™ followed the recommendations of the developers of the program as well as the 2015 SWGDAM Guidelines for validation of Probabilistic Genotyping Systems.

Peer Review & Publications

- Papers describing the biological model, mathematics, performance and validation of STRmix™ have been published in various peer-reviewed forensic journals
- STRmix™ source code is available with a non-disclosure agreement
  - Terms and further information available on the STRmix™ website
- References available

STRmix™ Acceptance in Court

- Accepted after admissibility hearings in New York, Michigan, Texas and Florida
- Evidence has been used in more than 65 other cases in NY, California, Idaho, Michigan, Texas, Georgia, Wyoming, South Carolina, Wisconsin and Florida
- USACIL 20+ trials, ~500 reports (as of April 2016)
- A non-exhaustive list of U.S. cases is included on the references and general information sheet

New York vs Oral "Nick" Hillary

- DNA Testimony not allowed
- New York had not performed an internal validation
- Ruling specifically stated that "Based upon a review of the record, this court finds that STRmix has been developmentally validated and is generally accepted as reliable within the scientific community."
- A copy of the ruling is available

Have STRmix™ results been accepted in Indiana courts?

- Not to our knowledge
- STRmix™ was implemented for use in the Indiana State Police Laboratory system on November 1, 2017. It would only have been accepted if testing was performed by another public or private laboratory that uses STRmix™ unbeknownst to us.

Have any probabilistic genotyping results been accepted in Indiana courts?

- Yes. There are several other probabilistic genotyping systems available. One of them is Cybergenetics' TrueAllele® system.
  - Cybergenetics will perform a free evaluation of some evidence profiles upon request, official results will need to be paid for, but there have been a few counties in the state that taken it upon themselves to pursue this analysis.

1. **Indiana v Dugniqio Dishay Forest** (6/3/2016)
   82D03-1501-F2-000566 – Vanderburgh County
   Defendant withdrew objection prior to conclusion of Daubert – court found TrueAllele® scientifically reliable
2. **Indiana v Malcolm Bryan Wade** (8/1/2016)
   53C02-1411-F3-001042 – Monroe County
   No apparent challenge
3. **Indiana v Randal L. Coalter** (8/3/2017)
   62C01-1703-MR-000192 – Perry County
   Motion to exclude denied

Questions?