NIST National Institute of Standards and Technology U.S. Department of Commerce

NEAFS
Probabilistic DNA Mixture Interpretation Workshop
Allentown, PA
September 25-26, 2015

Probabilistic Genotyping Software

Simone Gittelson, Ph.D., simone.gittelson@nist.gov
Michael Coble, Ph.D., michael.coble@nist.gov

Acknowledgement

I thank Simone Gittelson, Todd Bille, John Butler, and John Buckleton for their helpful discussions.

Disclaimer

Points of view in this presentation are mine and do not necessarily represent the official position or policies of the National Institute of Standards and Technology.

Why go to probabilistic genotyping?

- CPI was never meant to be used for higher order, complex DNA mixtures.
- 2p with the binary LR may actually be anti-conservative.
- Probabilistic genotyping uses all of the data – no need to drop loci.

Electrophoresis 2014, *35*, 3125–3133

Todd W. Bille[1]
Steven M. Weitz[1]
Michael D. Coble[2]
John Buckleton[3]
Jo-Anne Bright[3]

[1]Bureau of Alcohol, Tobacco, Firearms and Explosives,

Research Article

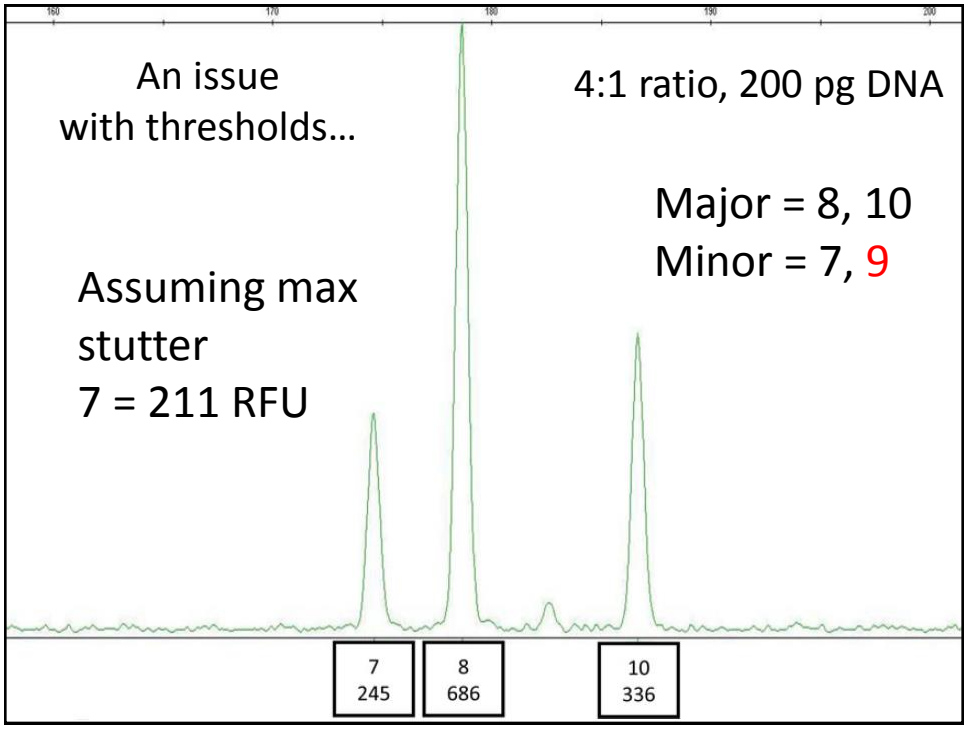
Comparison of the performance of different models for the interpretation of low level mixed DNA profiles

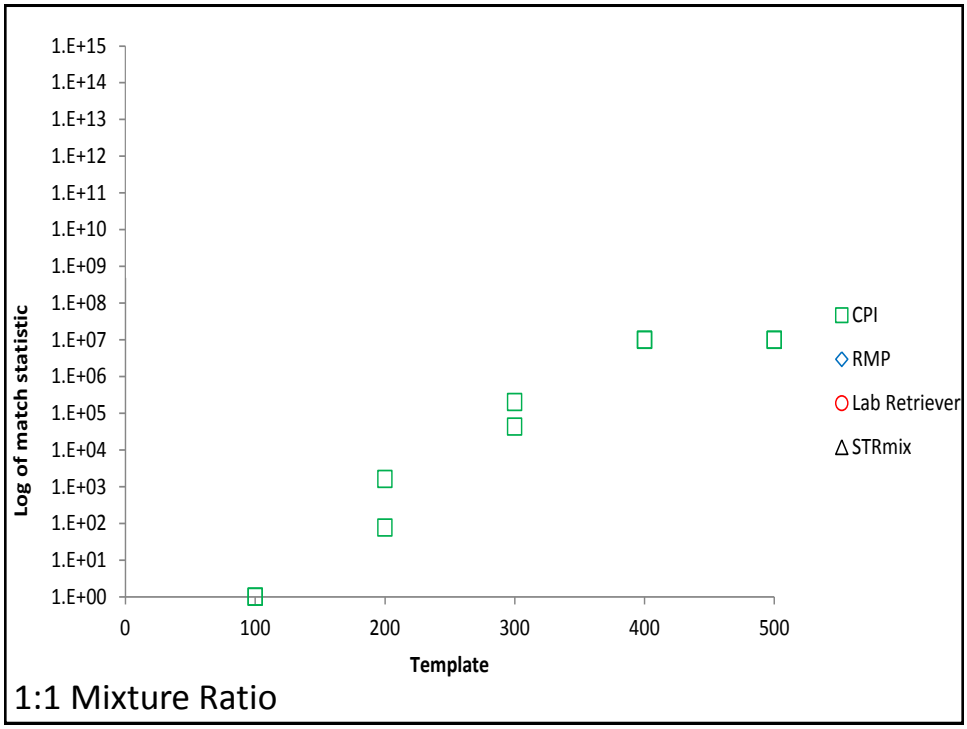
Advantages of Probabilistic Models

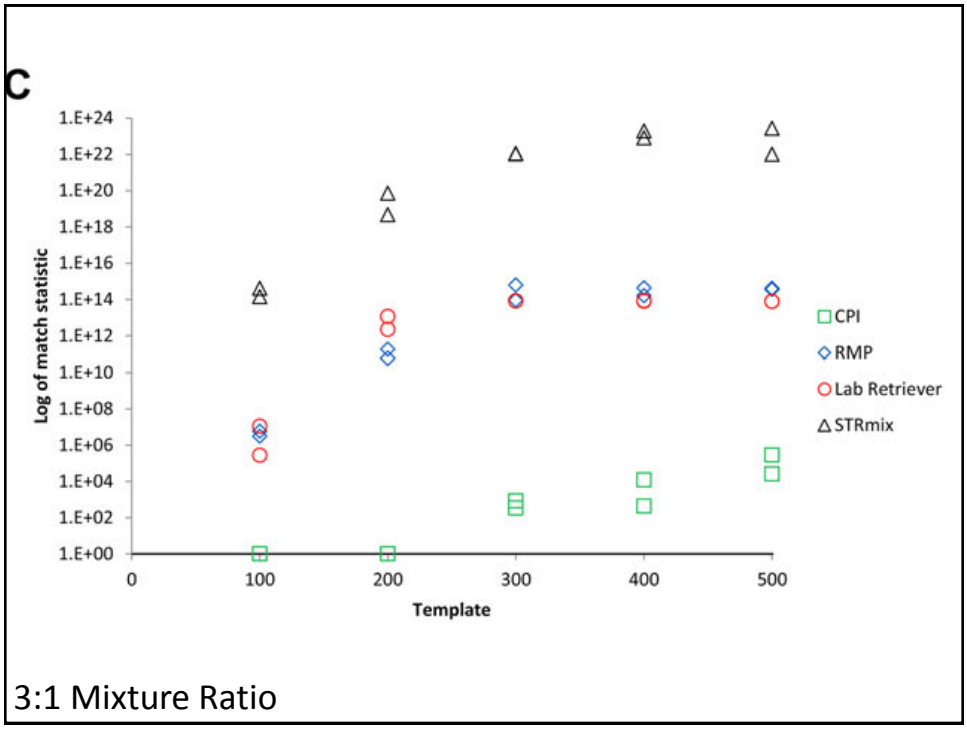
- Bille et al. *Electrophoresis*
- Used two samples with low allele sharing (10 markers – 4 alleles, 5 markers – 3 alleles). 2 PCR amplifications.
- 1:1, 2:1, 3:1, 4:1, and 5:1
- 500, 400, 300, 200, 100 pg input DNA
- CPI, RMP (2p), Lab Retriever, STRmix

Table 1. Average number of loci used for CPI and RMP calculations for mix ratios of 1:1, 2:1, and 3:1

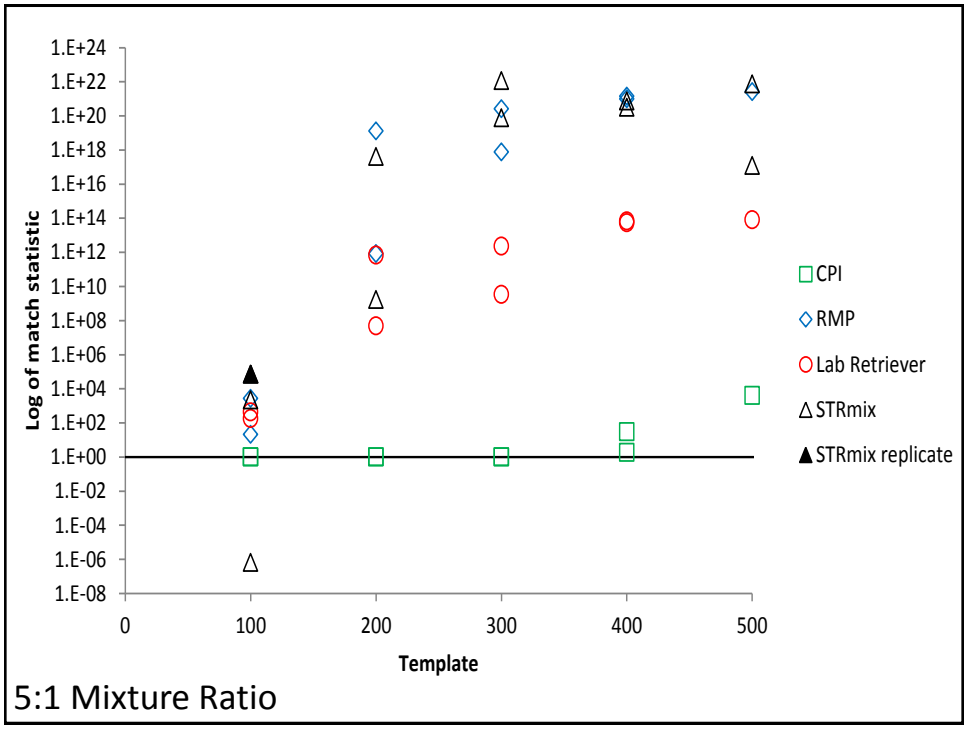
	Total DNA template				
	100 pg	200 pg	300 pg	400 pg	500 pg
CPI—Avg. loci used	0.0	3.2	7.7	12.2	14.0
RMP—Avg. loci used	13.8	14.8	15.0	15.0	15.0


An issue with thresholds…
4:1 ratio, 200 pg DNA
Major = 8, 10
Minor = 7, 9
Assuming max stutter
7 = 211 RFU
7
245
8
686
10
336

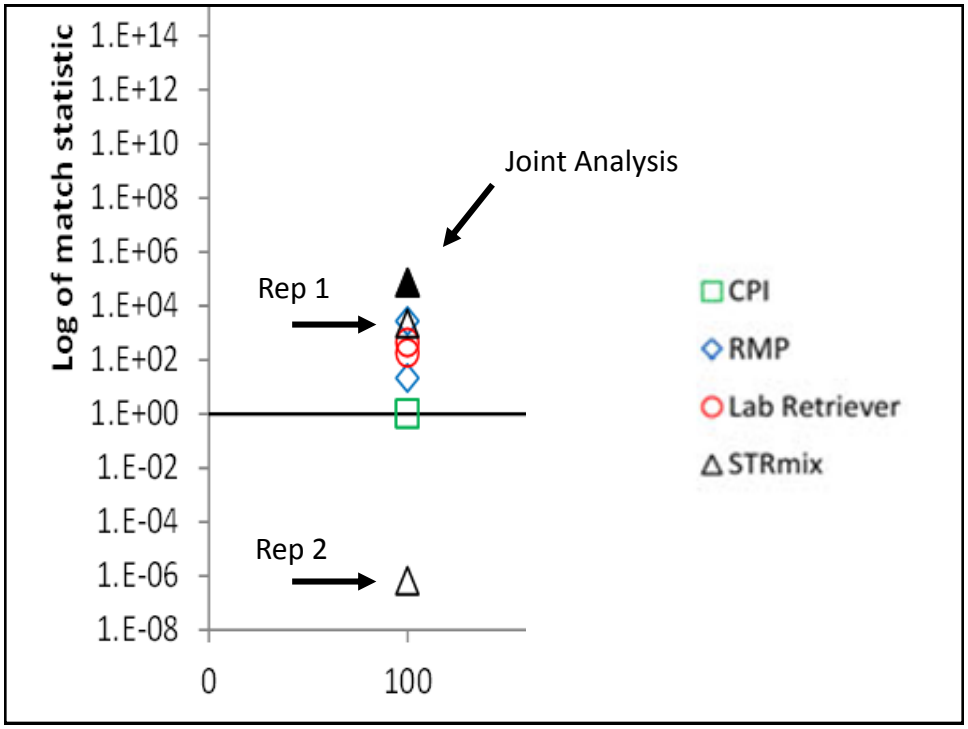

Log of match statistic
1.E+15
1.E+14
1.E+13
1.E+12
1.E+11
1.E+10
1.E+09
1.E+08
1.E+07
1.E+06
1.E+05
1.E+04
1.E+03
1.E+02
1.E+01
1.E+00
0
100
200
300
400
500
Template
CPI
RMP
Lab Retriever
STRmix
1:1 Mixture Ratio



3:1 Mixture Ratio



5:1 Mixture Ratio



General Trends

- Both Semi-continuous and fully-continuous models typically make more use of the data than the binary LR and CPI.
- This is especially true for low level mixtures.
- This exercise was an “ideal” mixture that decreased the uncertainty with masking alleles.

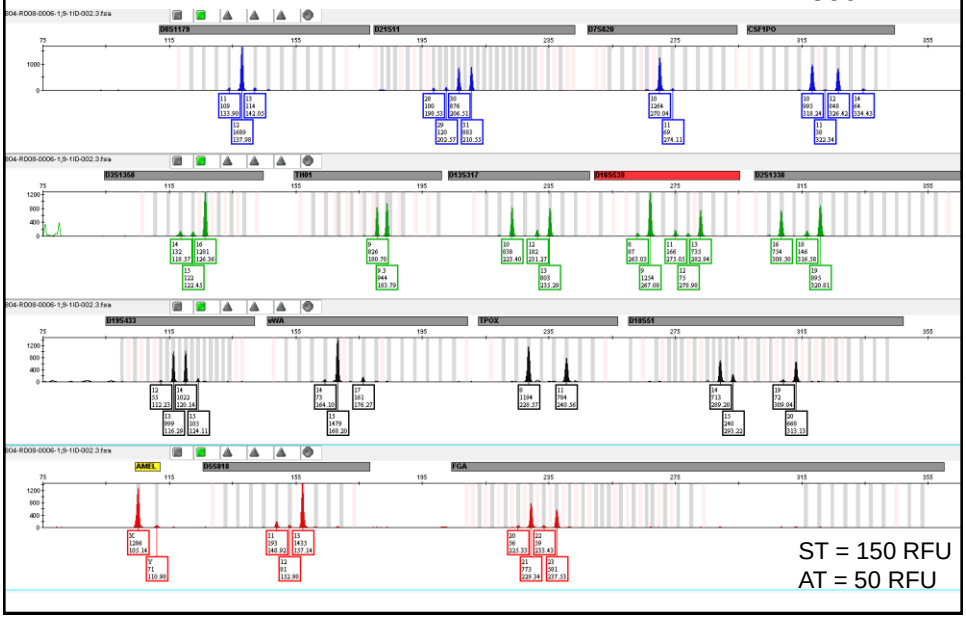
Software Examined

- LR Mix (Gill and Haned) – open source R program with GUI.
- Lab Retriever (Rudin, Lohmueller, Inman) – free software, based on the Balding/Buckleton approach.
- TrueAllele (Cybergenetics) – Continuous approach, publications and presentations by Perlin *et al.*
- STRmix (Australia/NZ) – Continuous approach, publications by Taylor, Bright and Buckleton.

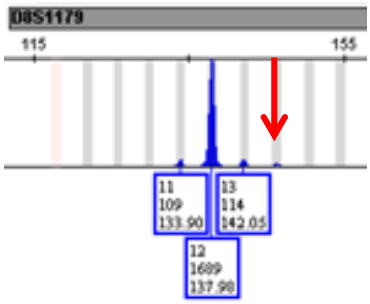
Some Ground Rules

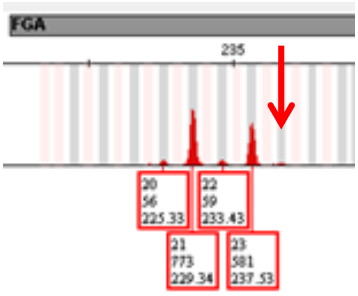
- For LRmix and Lab Retriever, the same values for $Pr(D_{out})$ and $Pr(D_{in})$ were used.
- The NIST (2003) allele frequencies for Western Europeans were used for all systems.
- TrueAllele analyses were performed at both 10 RFU (default) and 50 (2p mixtures) or 30 (3p mixture) RFUs.

Example 1 – Low-level 2p with D_{out}



Loci with Drop-out

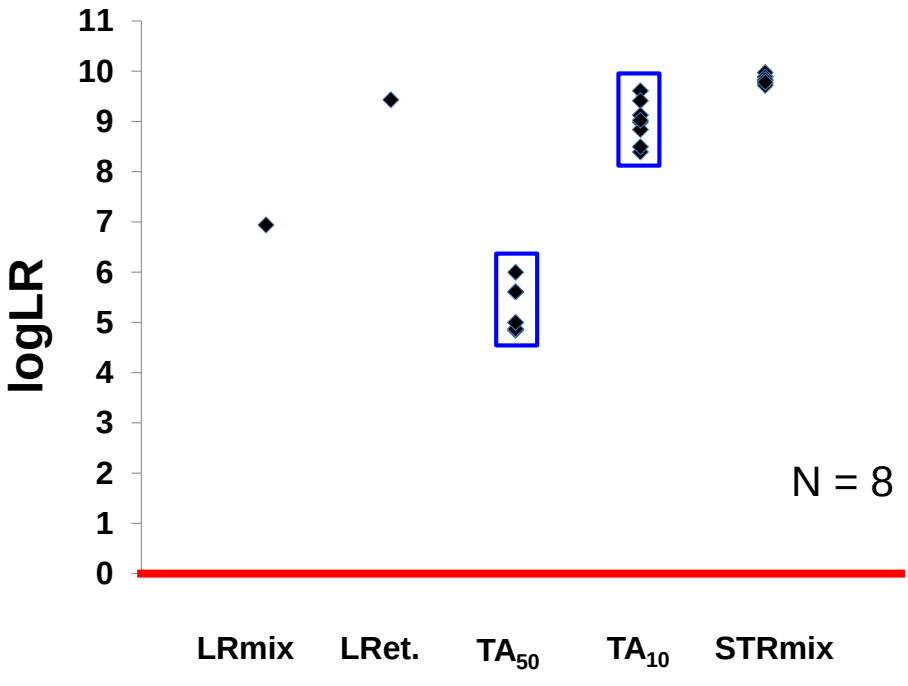




POI = 13, 14

POI = 21, 24

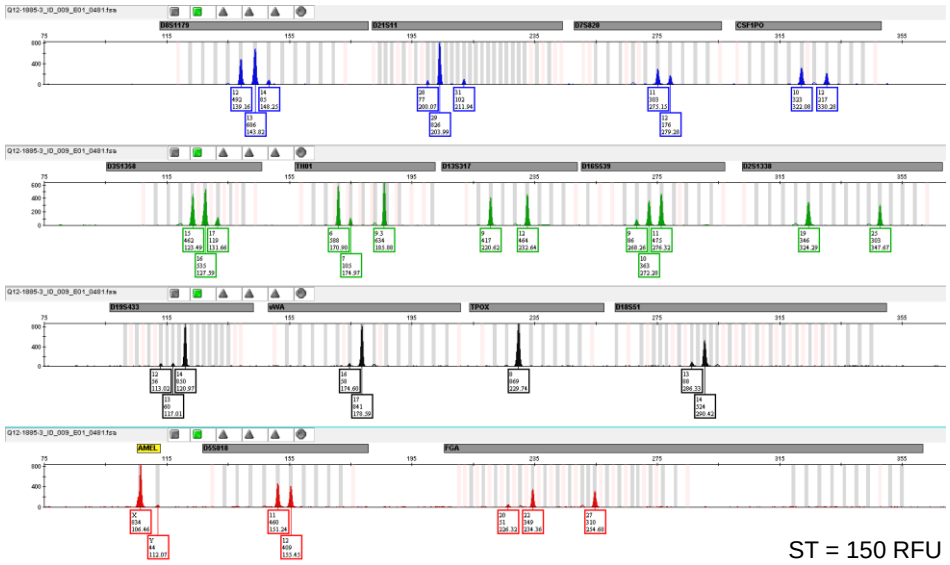
2 loci with allele drop-out



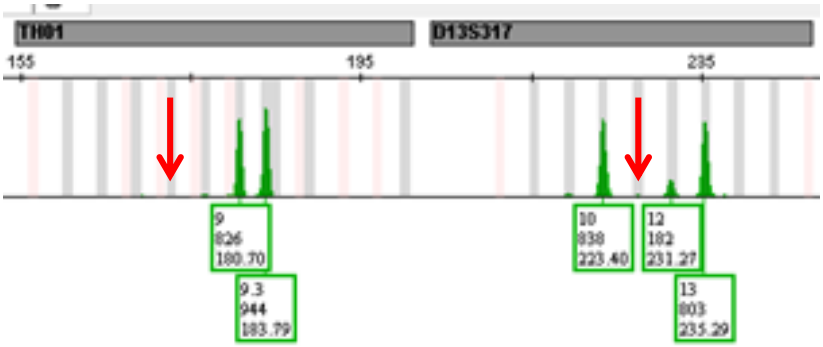
Time of Analysis

	2p - 2DO
LRmix	< 1 sec.
Lab Ret.	< 1 sec.
TrueAllele	16+ hours
STRmix	25.2 sec.

Example 2 – Low-level 2p more D_{out}

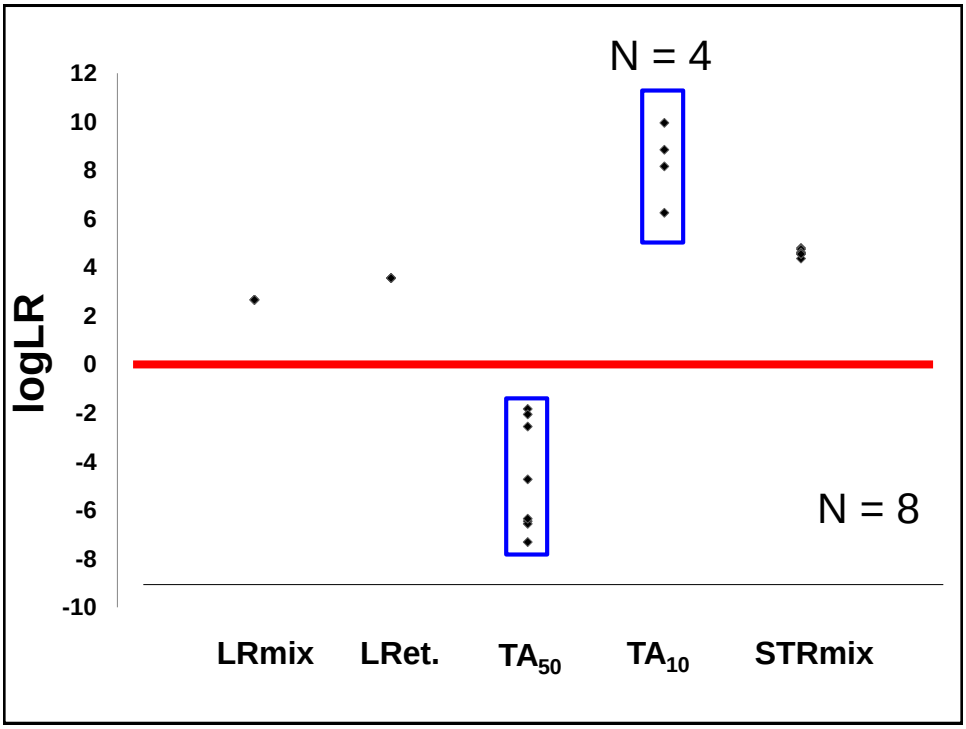


Loci with Drop-out



POI = 7, 9 POI = 11, 12

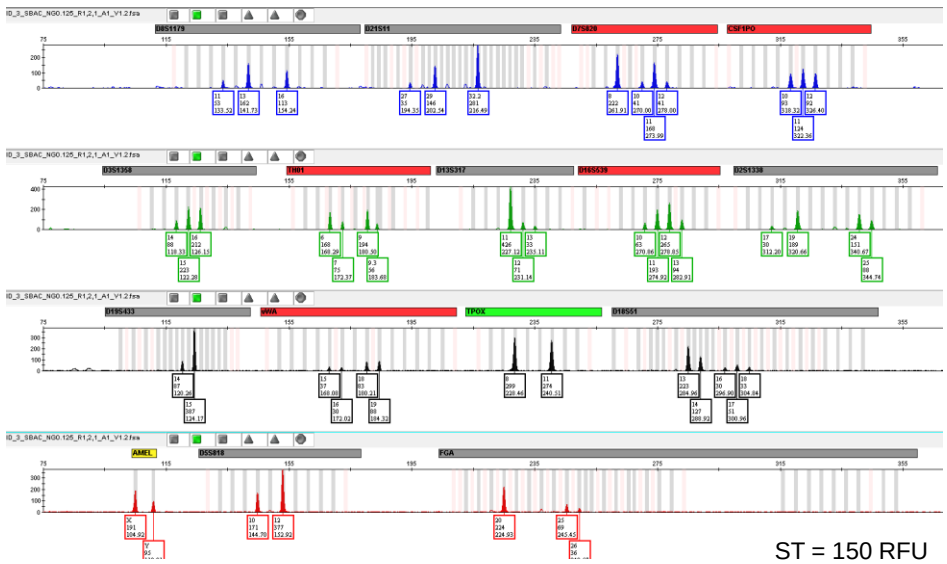
5 loci with allele drop-out; 1 locus drop-out (CSF)



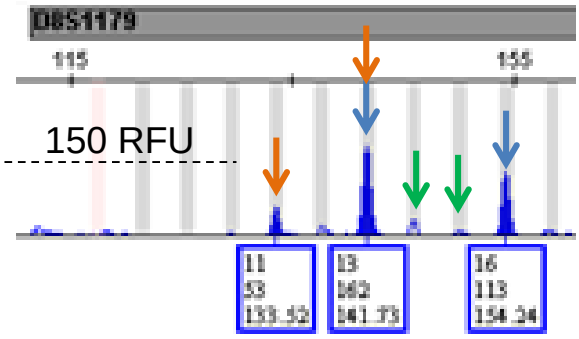
Time of Analysis

	2p - 2DO	2p - 6DO
LRmix	< 1 sec.	< 1 sec.
Lab Ret.	< 1 sec.	1 sec.
TrueAllele	16+ hours	16+ hours
STRmix	25.2 sec.	14.8 sec.

Example 3 – Low-level 3 person mixture



Example 3 – Low-level 3 person mixture



A = 13,16

B = 11,13

C = 14,15

125 pg input DNA
1:2:1 ratio

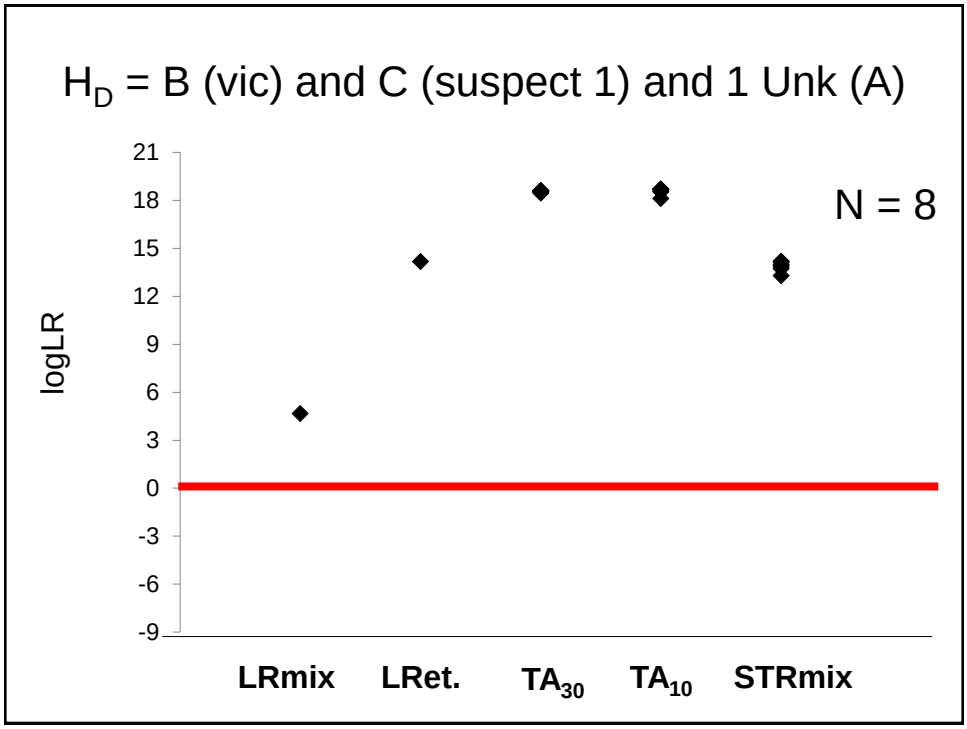
B = female

Conditioning

- H_P = B (vic) and C (suspect 1) and A (suspect 2)
- (1) H_D = B (vic) and C (suspect 1) and 1 Unk
- (2) H_D = B (vic) and A (suspect 1) and 1 Unk

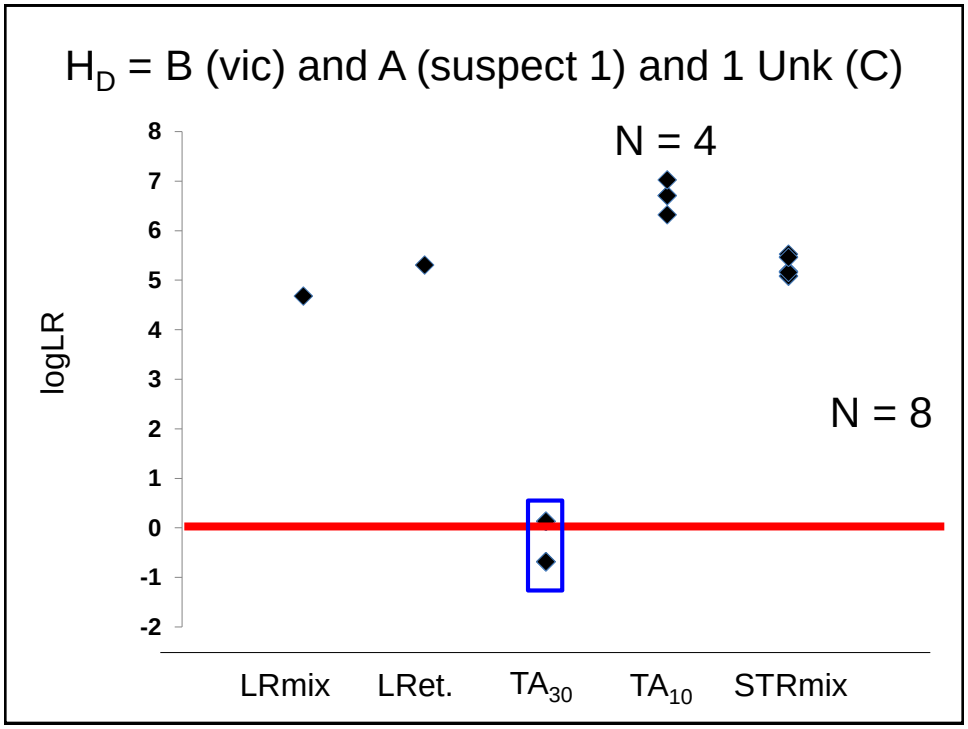
Suspect A, Pr(DO) = 0.02

Suspect C, Pr(DO) = 0.529



Time of Analysis

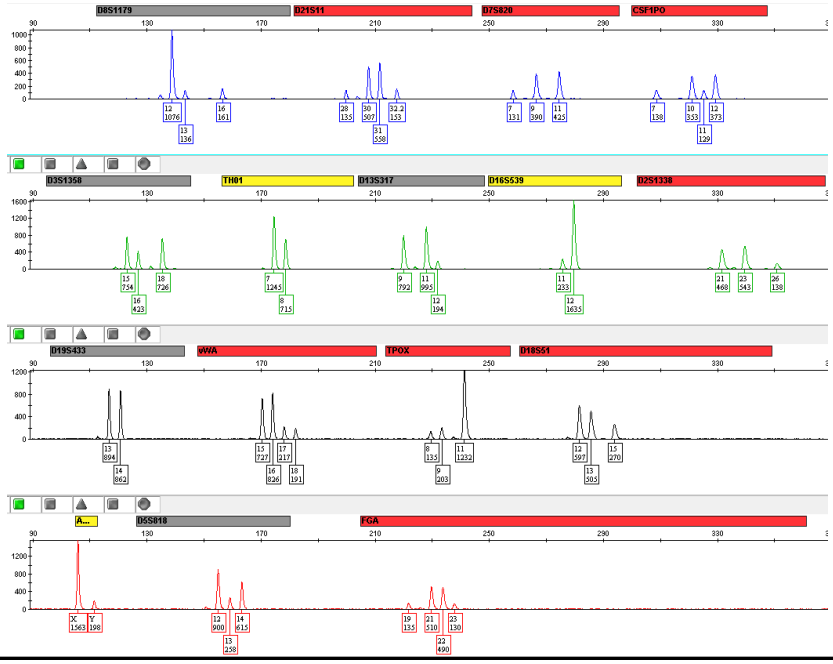
	2p - 2DO	2p - 6DO	3p - A unk
LRmix	< 1 sec.	< 1 sec.	3 sec.
Lab Ret.	< 1 sec.	1 sec.	8 sec.
TrueAllele	16+ hours	16+ hours	48+ hours
STRmix	25.2 sec.	14.8 sec.	63.1 sec.



Time of Analysis

	2p - 2DO	2p - 6DO	3p - A unk	3p - C unk
LRmix	< 1 sec.	< 1 sec.	3 sec.	3.4 sec.
Lab Ret.	< 1 sec.	1 sec.	8 sec.	7.5 sec.
TrueAllele	16+ hours	16+ hours	16+ hours	84+ hours
STRmix	25.2 sec.	14.8 sec.	63.1 sec.	50.5 sec.

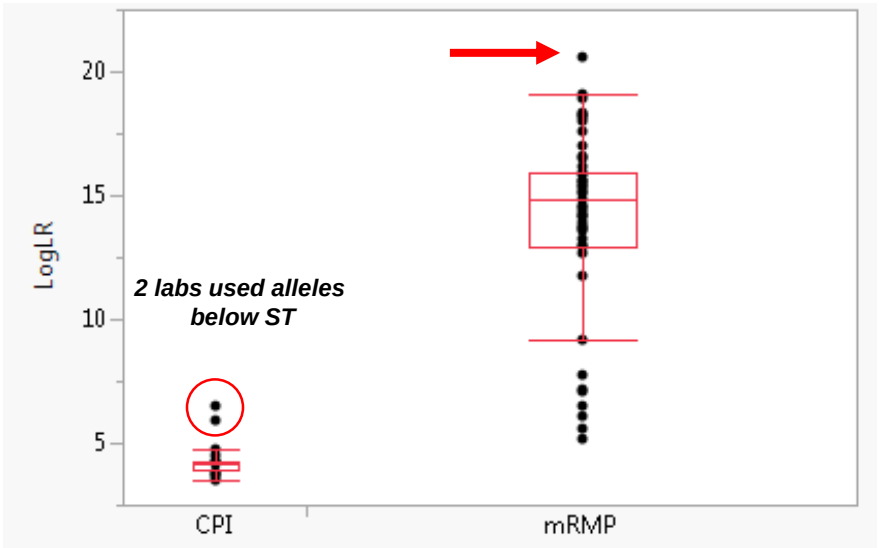
Case 04 – IDPlus



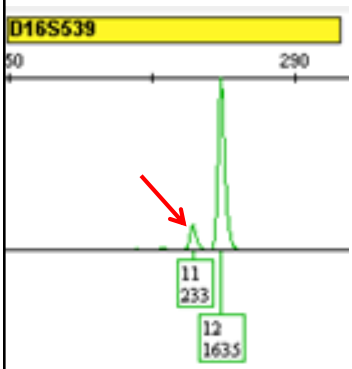
MIX13 Study (Case 04)

- Summary – Mock sexual assault, 2 person **3.5**:**1** mixture, minor component has alleles below the ST of 150 (required by all labs!)
- Purpose – How many labs would attempt to separate the two components?
- Good News! No false exclusions.

Statistical Concerns



Focus on Uncertainty with D16 (stutter)

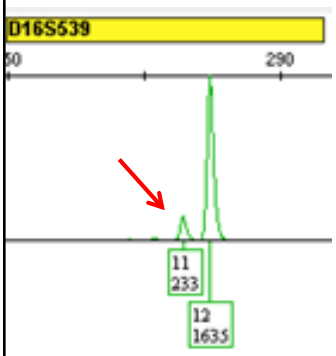


If 10% stutter from the 12 allele (163 RFU) is part of the 11 allele, then the remaining peak (70 RFU) is below the ST

No CPI labs excluded D16 from the stat

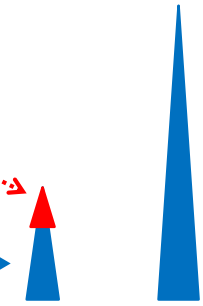
POI = 11, 12

Uncertainty with D16



70 RFU
(POI)

163 RFU
(Stutter)



11
233

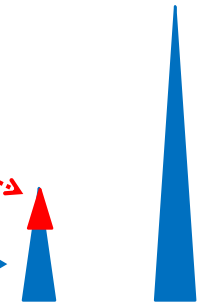
12
1633

POI = 11, 12

Uncertainty with D16
Observed Approaches

70 RFU
(POI)

163 RFU
(Stutter)



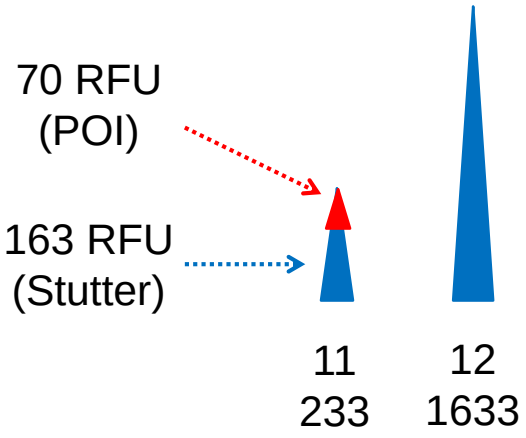
11
233

12
1633

Drop the locus
(Below ST)

POI = 11, 12

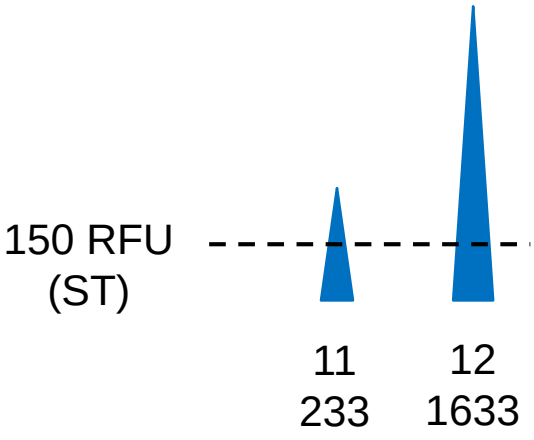
Uncertainty with D16 Observed Approaches



Use 2p
(Below ST)

POI = 11, 12

Uncertainty with D16 Observed Approaches



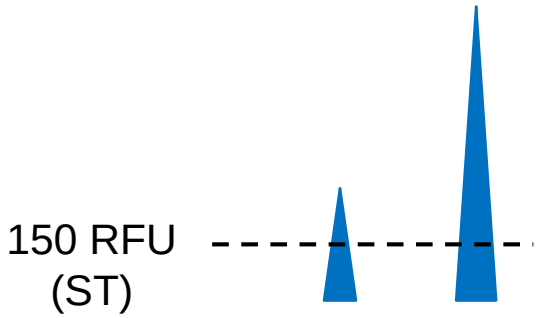
Use 2pq or p^2

(to infer genotypes)

11,12 or 11,11

POI = 11, 12

Uncertainty with D16 Observed Approaches



Use 2pq

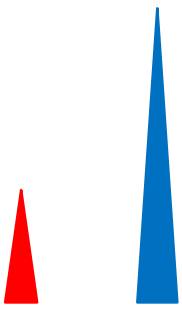
(to infer genotype)

11,12

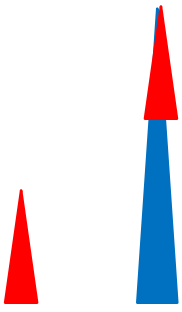
(matches POI)
"Certainty"

POI = 11, 12

Ratio = 7 : 1

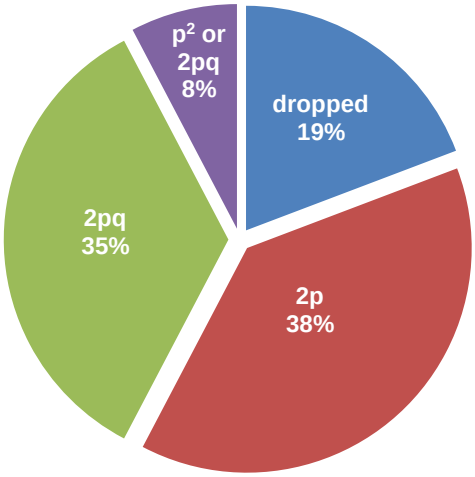


Ratio = 3 : 1



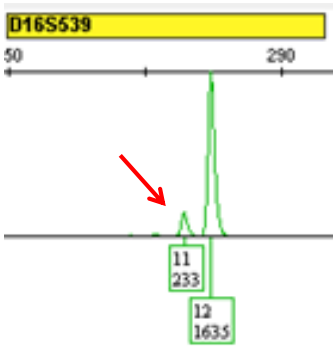
Assumes 0% Stutter

Statistics with D16 (mRMP/LR)



Probabilistic approaches to interpretation can be useful to reduce uncertainty and subjectivity

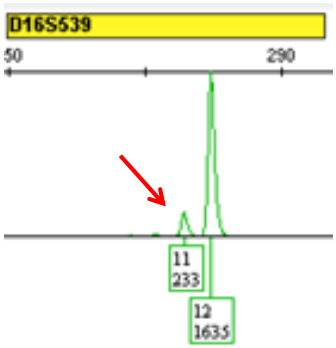
Probabilistic Weights



D16S539

[12,12]	[11,11]	0.0441
[12,12]	[-1,11]	0.0007
[12,12]	[11,12]	0.7428
[12,12]	[12,12]	0.2123
[12,12]	[-1,12]	0.0001

Stats



$$\frac{1}{2pq} = 4.77$$

$$\frac{1}{2pq \text{ or } p^2} = 3.20$$

$$\frac{1}{CPI} = 2.38$$

$$\frac{1}{2p} = 1.15$$

STRmix = 4.30

Summary

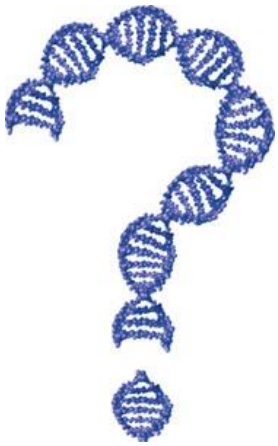
- The goal of the software programs should not be to simply “get bigger numbers” but to understand the details of these approaches and not treat the software as a “black box.”
- Semi-continuous approaches will produce a LR that could be replicated by hand if necessary.

Summary

- Each approach has it's own advantages and disadvantages.
- "When analysed using a discrete model such as that of [Balding and Buckleton], it is necessary to rely on an analyst designations of low peaks as allelic, stutter, or masking and hence ambiguous."

  - Puch-Solis *et al.* (2013)

Thank you for your attention

Thanks!!
Dr. Larry Quarino
Sheila Heller
Cedar Crest College

NEAFS!