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

NIST-GMI Seminar
Innsbruck, Austria

Resolving Challenging Mixtures Using Probabilistic Models of Interpretation

Michael D. Coble

NIST Applied Genetics Group

September 5, 2011

Official Disclaimer

The opinions and assertions contained herein are solely those of the author and are not to be construed as official or as views of the U.S. Department of Commerce, the U.S. Department of Justice, or the National Institute of Justice.

Commercial software, equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the U.S. Department of Commerce, the U.S. Department of Justice, or the National Institute of Justice nor does it imply that any of the software, materials, instruments or equipment identified are necessarily the best available for the purpose.

The Forensic Institute

Forensic e-Symposium™

The Forensic Institute

April 14, 2005

"If you show 10 colleagues a mixture, you will probably end up with 10 different answers."

- Dr. Peter Gill

"Don't do mixture interpretation unless you have to"

- Dr. Peter Gill (1998)

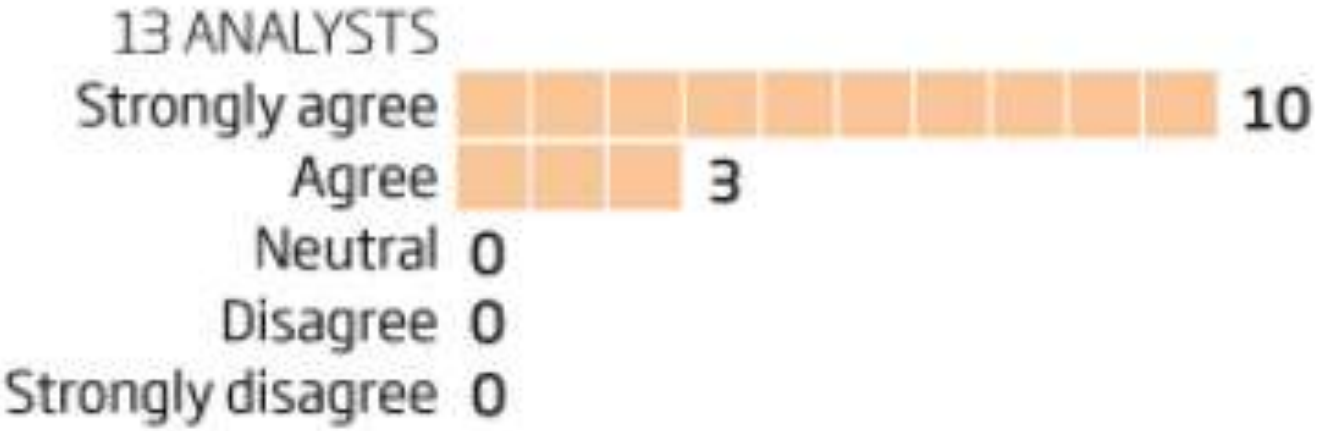
NewScientist

11 August 2010

Fallible DNA evidence can mean prison or freedom

http://www.newscientist.com/article/mg20727733.500-fallible-dna-evidence-can-mean-prison-or-freedom.html

Q: Lab staff need more training on how to deal with complex profiles such as mixtures and very small samples of DNA



Responses from Australia, Canada, India, New Zealand, UK, and US.

Gill and Buckleton *JFS* **55:** 265-268 (2010)

- "The purpose of the ISFG DNA commission document was to provide a way forward to demonstrate the use of ***probabilistic models to circumvent the requirement for a threshold*** and to safeguard the legitimate interests of defendants."

JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, 2011
doi: 10.1111/j.1556-4029.2011.01859.x
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Mark W. Perlin,[1] M.D., Ph.D.; Matthew M. Legler,[1] B.S.; Cara E. Spencer,[1] M.S.; Jessica L. Smith,[1] M.S.; William P. Allan,[1] M.S.; Jamie L. Belrose,[2] M.S.; and Barry W. Duceman,[3] Ph.D.

Validating TrueAllele® DNA Mixture Interpretation*,†

- Quantitative computer interpretation using Markov Chain Monte Carlo testing
- Models peak uncertainty and infers possible genotypes
- Results are presented as the Combined LR

“Markov Chain Monte Carlo Testing”

“Markov Chain Monte Carlo Testing”

True Allele Software (Cybergenetics)

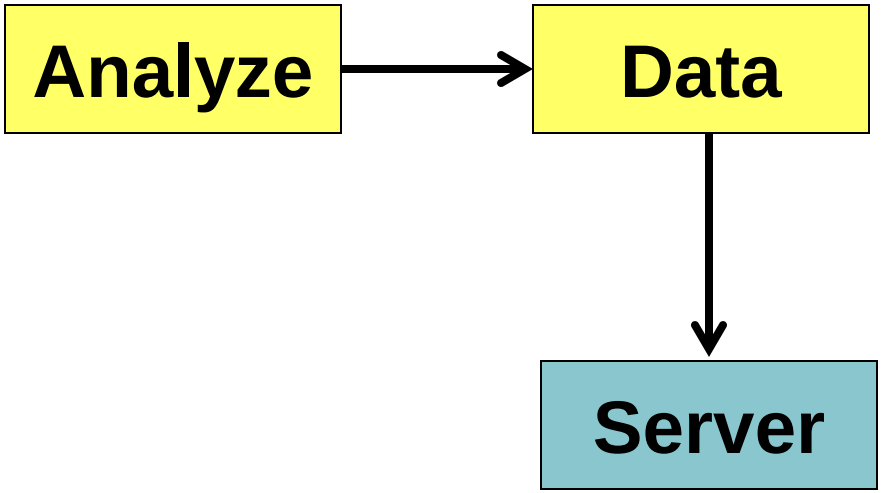
- We purchased the software in September 2010.
- Three day training at Cybergenetics (Pittsburgh, PA) in October.
- Software runs on a Linux Server with a Mac interface.

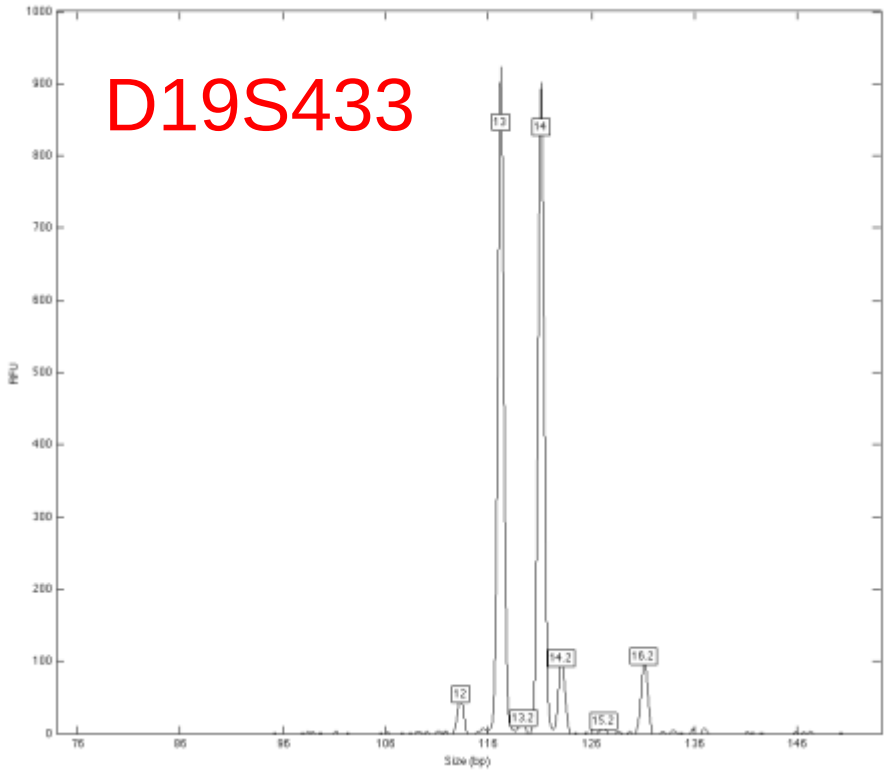
True Allele Casework Workflow
5 Modules

Analyze

.fsa files imported
Size Standard check
Allelic Ladder check
Alleles are called

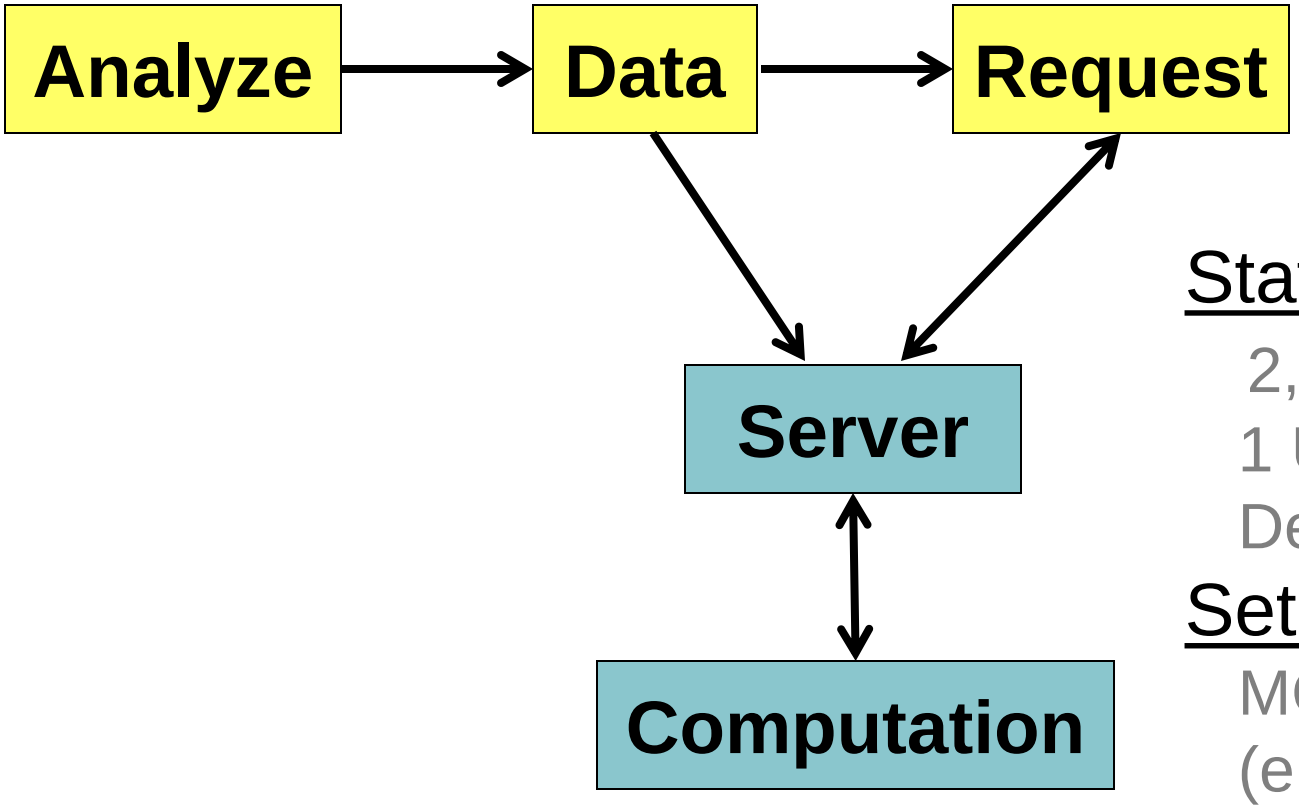
True Allele Casework Workflow
5 Modules





All Peaks above 10 RFU are considered

True Allele Casework Workflow
5 Modules



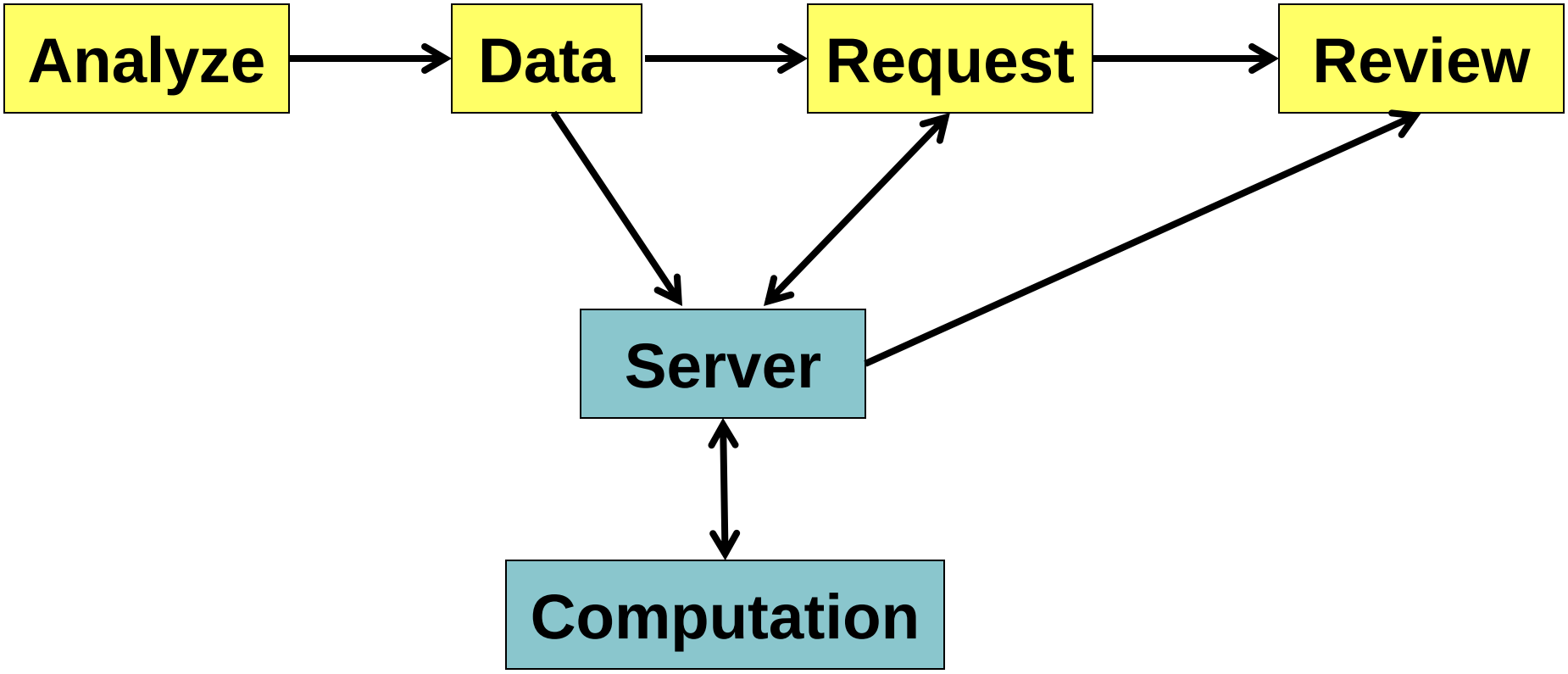
State Assumptions

2, 3, 4 unknowns
1 Unk with Victim?
Degradation?

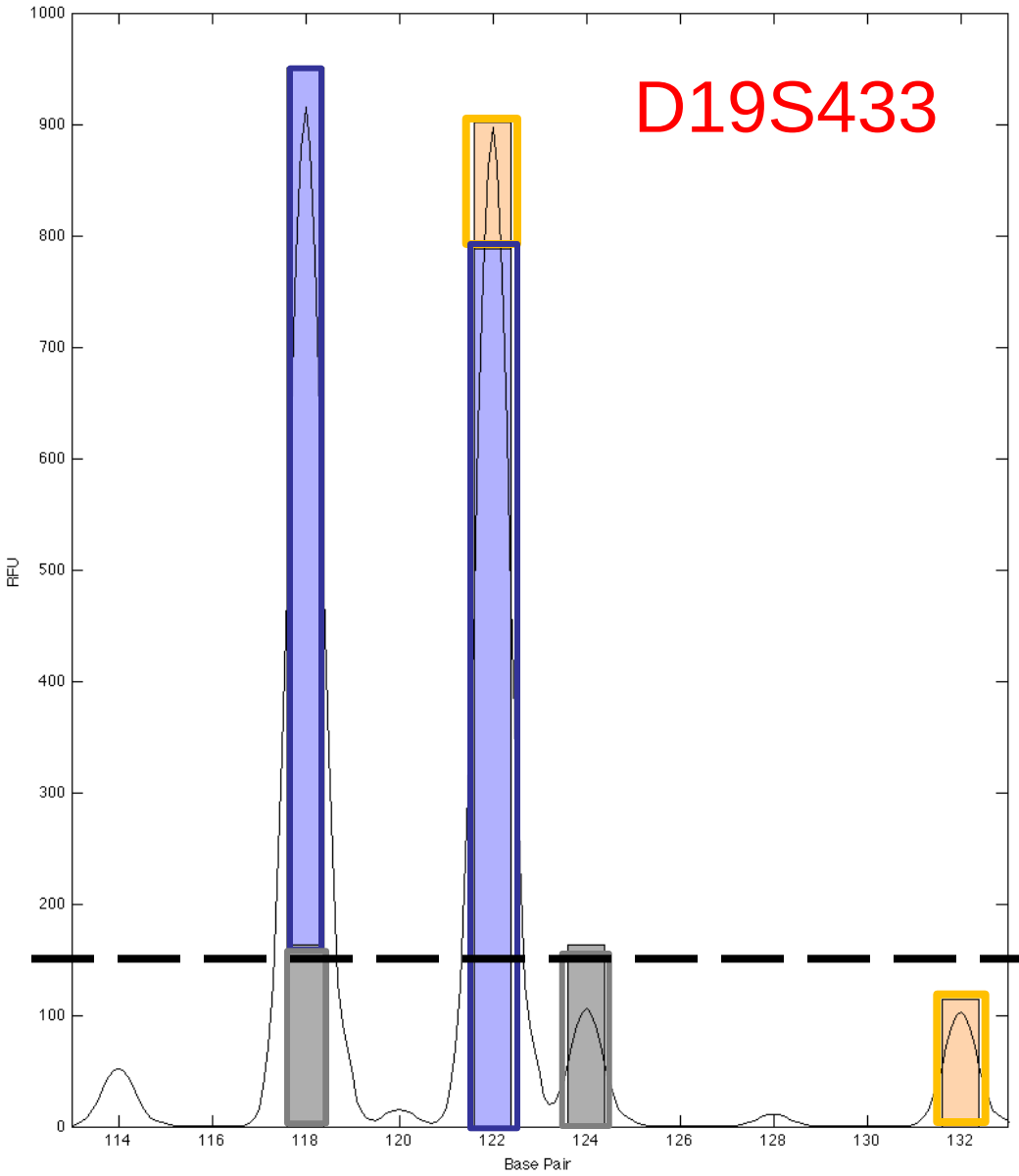
Set Parameters

MCMC modeling
(e.g.50K)

True Allele Casework Workflow
5 Modules



Review of One Replicate (of 50K)

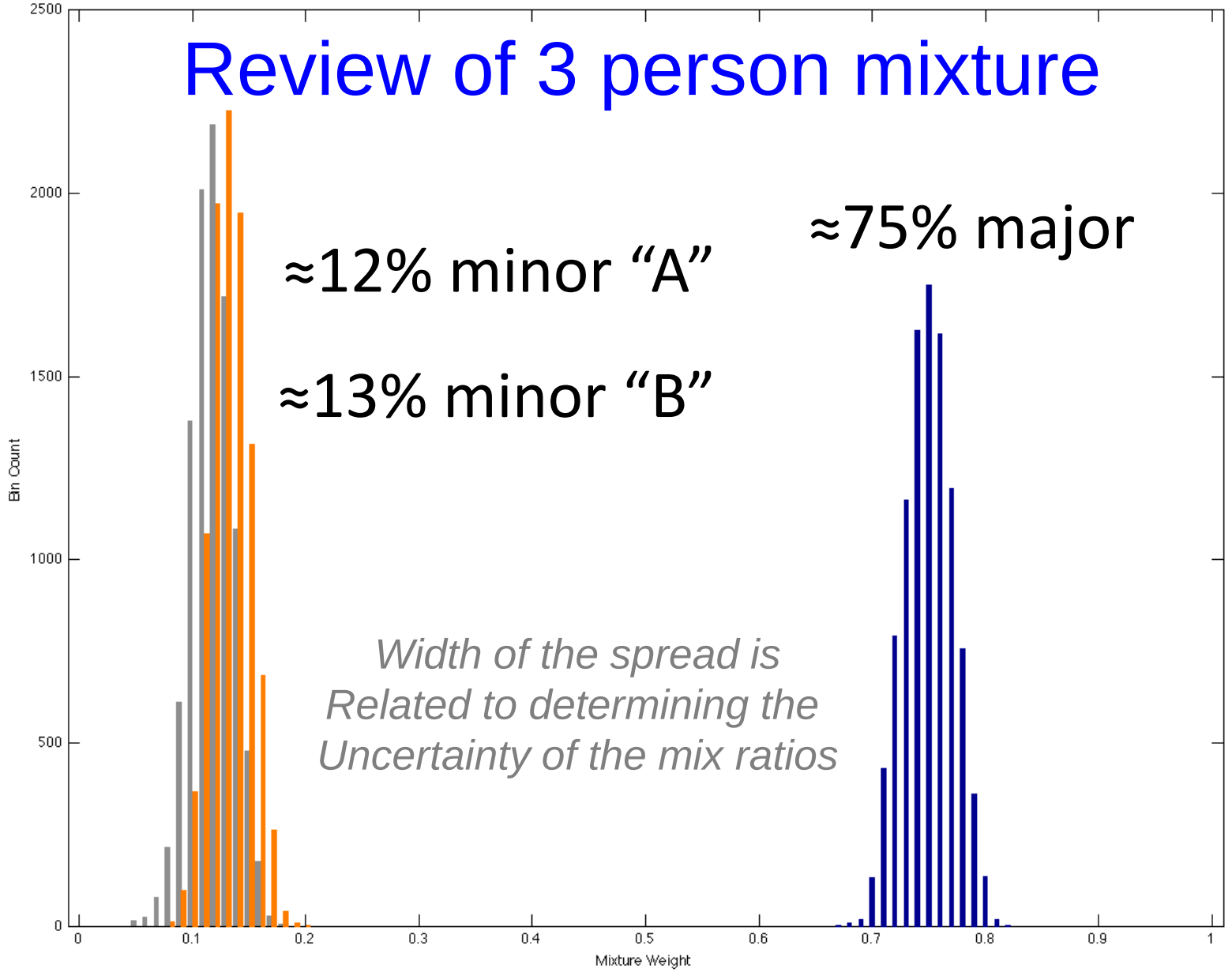


3P mixture,
2 Unknowns,

Conditioned on the Victim (major)

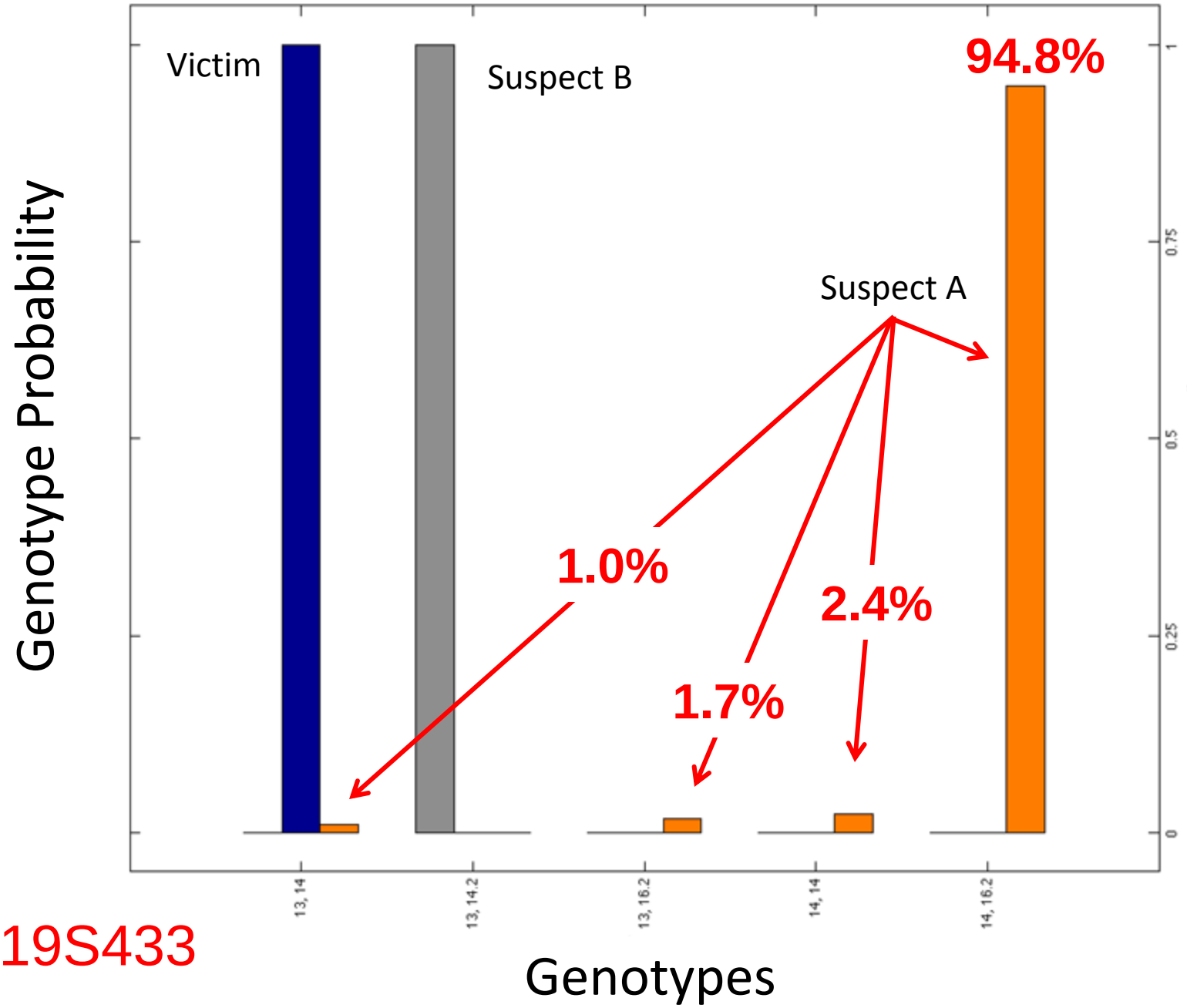
Good fit of the data to the model

Review of 3 person mixture



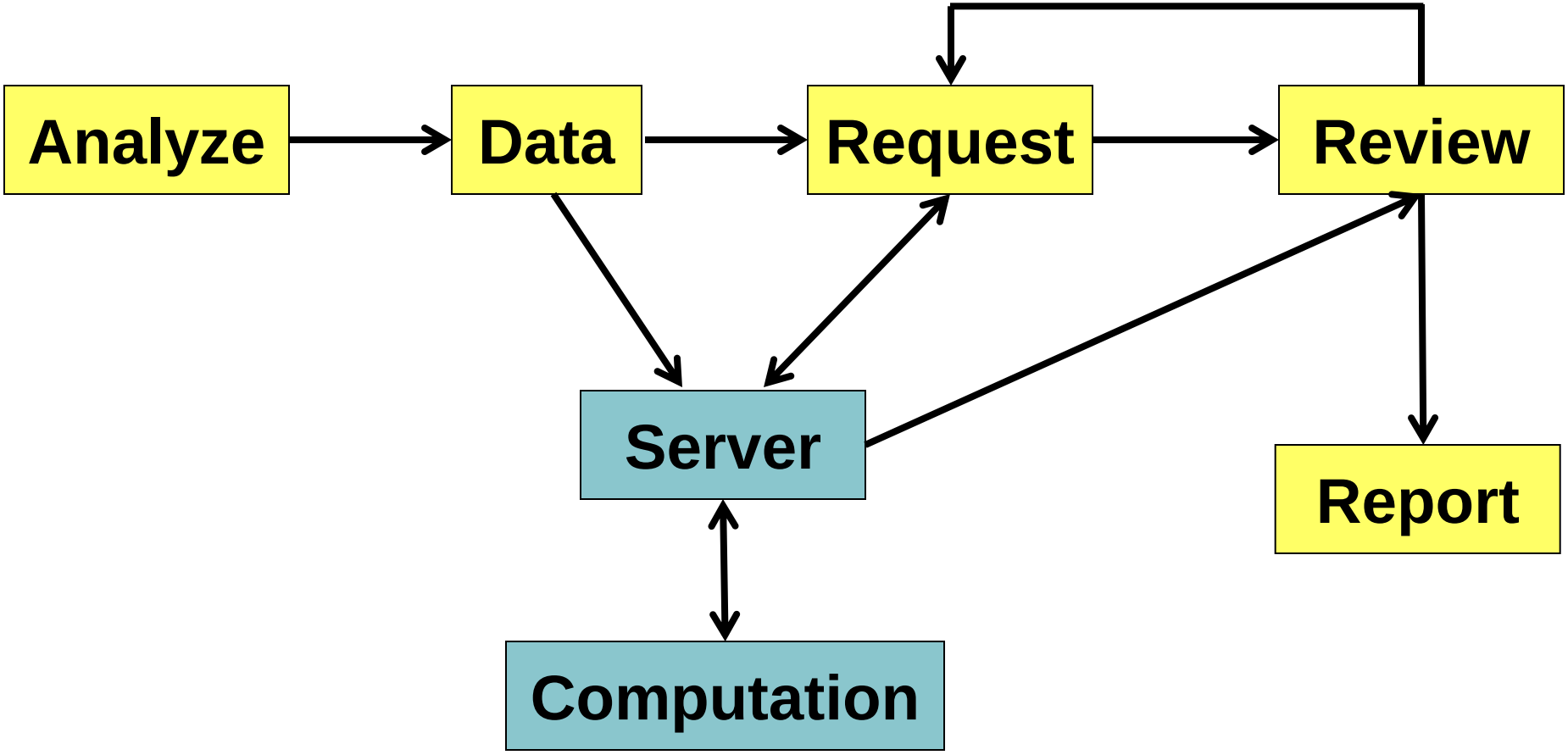
Bin Count

Mixture Weight


Victim
Suspect B
94.8%
Suspect A
1.0%
2.4%
1.7%
Genotype Probability
0
0.25
0.5
0.75
1
13, 14
13, 14.2
13, 16.2
14, 14
14, 16.2
Genotypes

D19S433

True Allele Casework Workflow
5 Modules



Determining the LR for D19S433

Suspect A = 14, 16.2

$$H_P = 1 * 0.967$$

Allele Pair	Probability Before Conditioning
14, 16.2	0.967
14, 14	0.003
13, 16.2	0.026
13, 14	0.001

$$LR = \frac{0.967}{\quad}$$

Determining the LR for D19S433

Suspect A = 14, 16.2

$H_P = 1 * 0.967$

Allele Pair	Probability Before Conditioning	Genotype Frequency	
14, 16.2	0.967	0.0120	
14, 14	0.003	0.0498	
13, 16.2	0.026	0.0131	
13, 14	0.001	0.1082	
		sum	0.0122

$$LR = \frac{0.967}{0.0122} = 79.26 \quad H_D$$

Combined LR = 5.6 Quintillion

			Genotype Probability Distribution			Weighted Likelihood	Likelihood Ratio		
	allele pair	Likelihood	Questioned	Reference	Suspect	Numerator	Denominator	LR	log(LR)
locus	x	l(x)	q(x)	r(x)	s(x)	l(x)*s(x)	l(x)*r(x)		
CSF1PO	11, 12	0.686	0.778	0.1448	1	0.68615	0.1292	5.31	0.725
D13S317	9, 12	1	1	0.0291	1	0.99952	0.02913	34.301	1.535
D16S539	9, 11	0.985	0.995	0.1238	1	0.98451	0.12188	8.036	0.905
D18S51	13, 17	0.999	1	0.0154	1	0.99915	0.01543	64.677	1.811
D19S433	14, 16.2	0.967	0.948	0.012	1	0.96715	0.01222	79.143	1.898
D21S11	28, 30	0.968	0.98	0.0872	1	0.96809	0.08648	11.194	1.049
D2S1338	23, 24	0.998	1	0.0179	1	0.99831	0.01787	55.866	1.747
D3S1358	15, 17	0.988	0.994	0.1224	1	0.98759	0.12084	8.14	0.911
D5S818	11, 11	0.451	0.394	0.0537	1	0.45103	0.07309	6.17	0.79
D7S820	11, 12	0.984	0.978	0.0356	1	0.98383	0.03617	27.198	1.435
D8S1179	13, 14	0.203	0.9	0.1293	1	0.20267	0.02993	6.771	0.831
FGA	21, 25	0.32	0.356	0.028	1	0.31986	0.01906	16.783	1.225
TH01	7, 7	0.887	0.985	0.1739	1	0.88661	0.15588	5.687	0.755
TPOX	8, 8	1	1	0.1375	1	1	0.13746	7.275	0.862
vWA	15, 20	0.998	0.996	0.0057	1	0.99808	0.00569	174.834	2.243

Results

- Results are expressed as logLR values

$$LR = 1{,}000{,}000 = 10^6$$

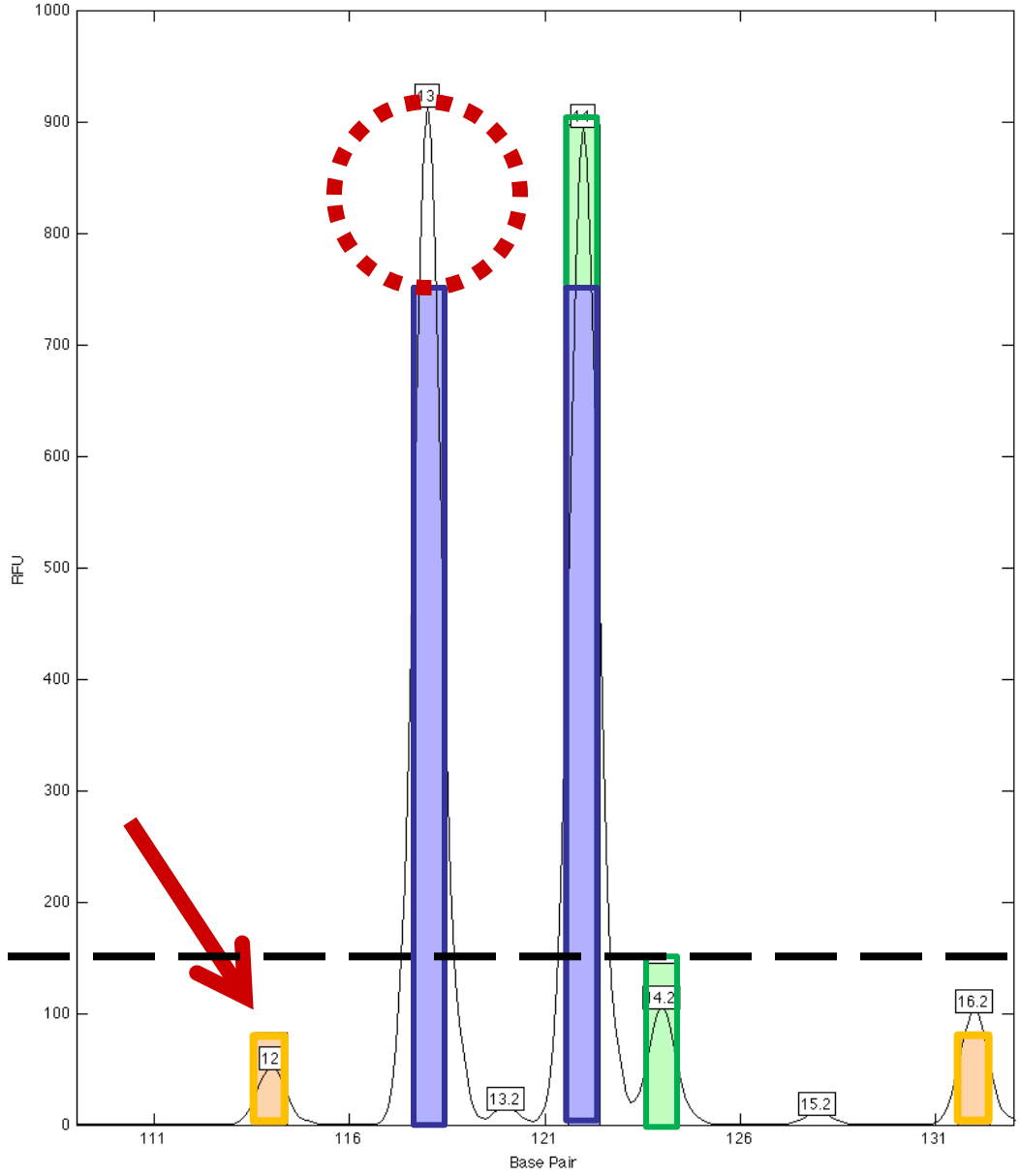
$$\log(LR) = \log 10^6$$

$$\log(LR) = 6 * \cancel{\log 10} \; (1)$$

$$\log(LR) = 6$$

Review of One Replicate (of 50K)

D19S433

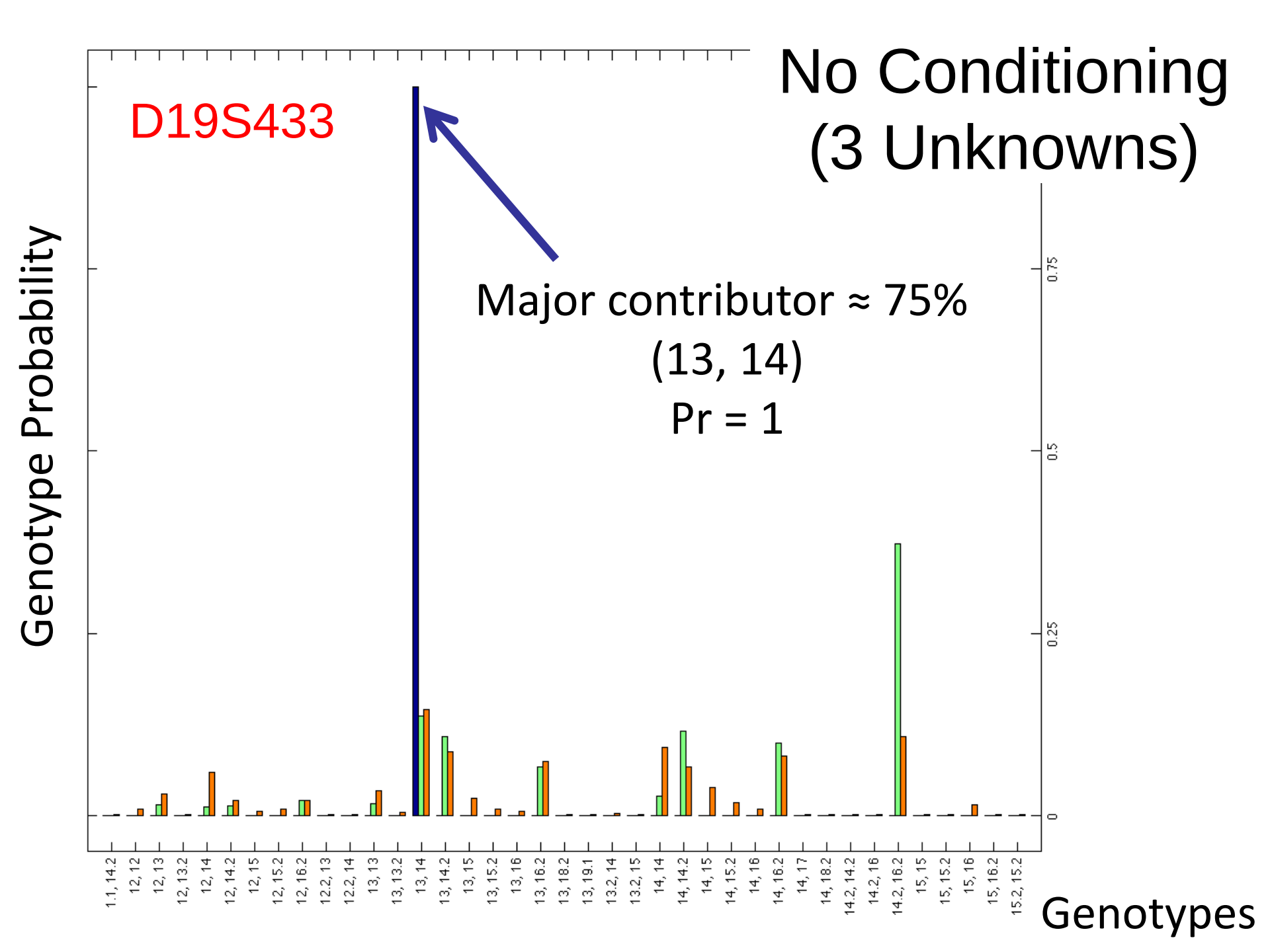


150 RFU

3P mixture,

3 Unknowns

Poor fit of the data to the model

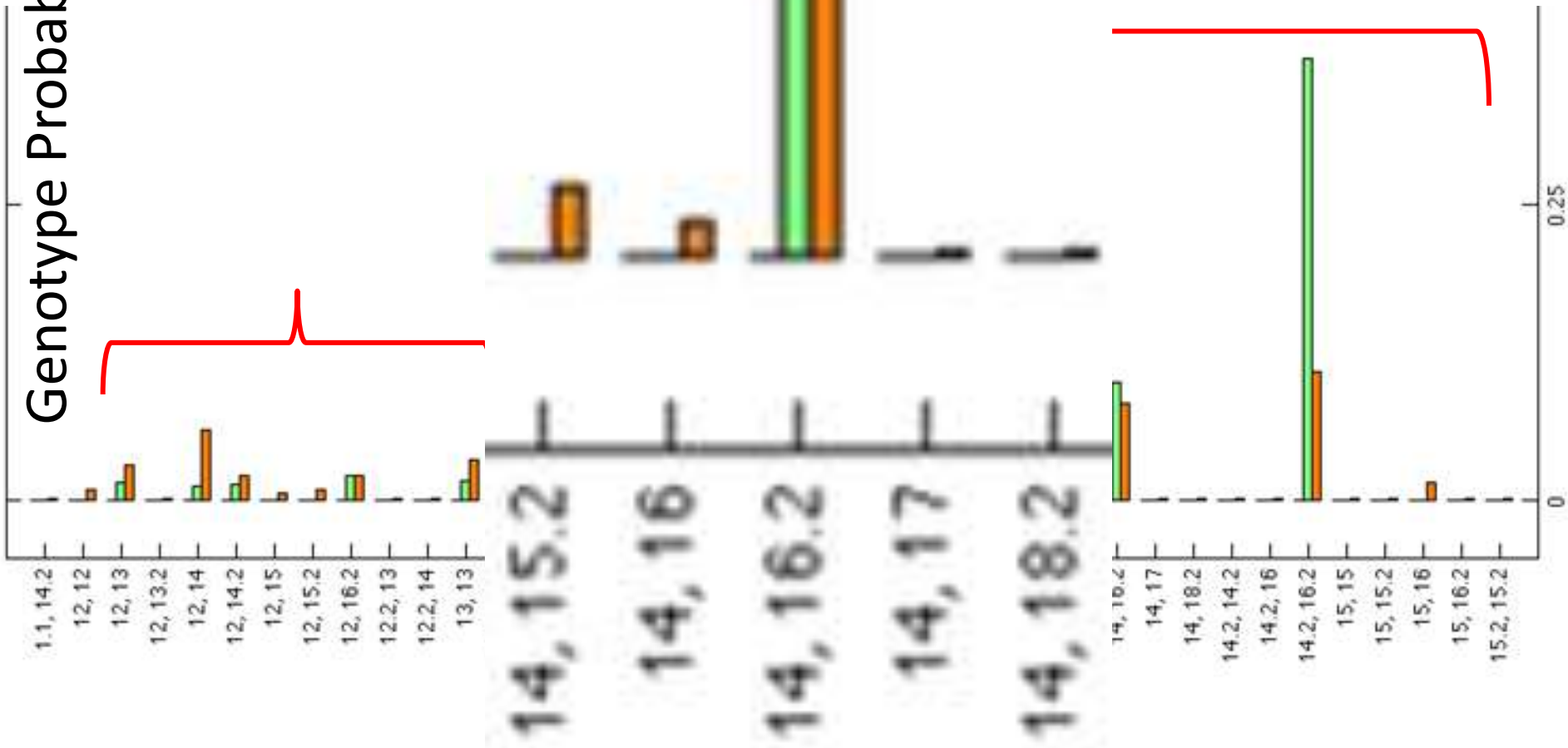

No Conditioning
(3 Unknowns)
D19S433
Major contributor ≈ 75%
(13, 14)
Pr = 1
Genotype Probability
Genotypes
0
0.25
0.5
0.75
1.1, 14.2
12, 12
12, 13
12, 13.2
12, 14
12, 14.2
12, 15
12, 15.2
12, 16.2
12.2, 13
12.2, 14
13, 13
13, 13.2
13, 14
13, 14.2
13, 15
13, 15.2
13, 16
13, 16.2
13, 18.2
13, 19.1
13.2, 14
13.2, 15
14, 14
14, 14.2
14, 15
14, 15.2
14, 16
14, 16.2
14, 17
14, 18.2
14.2, 14.2
14.2, 16
14.2, 16.2
15, 15
15, 15.2
15, 16
15, 16.2
15.2, 15.2

No Conditioning (3 Unknowns)

D19S433

8.1%

ns for the two
ibutors



locus	allele pair	L	Q	R	S	L*S	L*R	LR	log(LR)
D19S433	13 , 14	0.002	0.146	0.1082			0.00020		
	14.2, 16.2	0.270	0.109	0.0044			0.00118		
	14 , 14	0.002	0.093	0.0498			0.00008		
	13 , 14.2	0.017	0.088	0.0392			0.00068		
	14 , 16.2	0.013	0.081	0.0120	1	0.01295	0.00016		
	13 , 16.2	0.018	0.074	0.0131			0.00023		
	14 , 14.2	0.009	0.067	0.0361			0.00031		
	12 , 14	0.002	0.059	0.0498			0.00012		
	14 , 15	0.001	0.038	0.0343			0.00002		
	13 , 13	0.001	0.034	0.0587			0.00007		
	12 , 13	0.002	0.029	0.0541			0.00010		
	13 , 15	0.001	0.024	0.0373			0.00002		
	12 , 16.2	0.017	0.021	0.0060			0.00010		
	12 , 14.2	0.013	0.020	0.0180			0.00023		
	14 , 15.2	0.001	0.018	0.0275			0.00003		
	15 , 16	0.002	0.015	0.0006			0.00000		
	13 , 15.2	0.001	0.009	0.0299			0.00003		
	12 , 15.2	0.003	0.009	0.0137			0.00004		
	14 , 16	0.000	0.009	0.0017			0.00000		
	12 , 12	0.004	0.009	0.0125			0.00004		
	12 , 15	0.001	0.006	0.0172			0.00001		
	13 , 16	0.000	0.006	0.0019			0.00000		
	13 , 13.2	0.001	0.004	0.0261			0.00003		
	13.2, 14	0.001	0.003	0.0240			0.00002		
	13.2, 15	0.001	0.002	0.0083			0.00001		
	14 , 18.2	0.002	0.002	0.0017			0.00000		
	13 , 19.1	0.019	0.002	0.0000			0.00000		
	12 , 13.2	0.002	0.002	0.0120			0.00003		
	14.2, 16	0.001	0.002	0.0006			0.00000		
	12.2, 13	0.001	0.002	0.0168			0.00002		
	13 , 18.2	0.002	0.001	0.0019			0.00000		
	12.2, 14	0.001	0.001	0.0155			0.00001		
	14.2, 14.2	0.004	0.001	0.0065			0.00003		
	15 , 15	0.000	0.001	0.0059			0.00000		
	15 , 15.2	0.000	0.001	0.0095			0.00000		
	14 , 17	0.001	0.001	0.0000			0.00000		
	15 , 16.2	0.000	0.001	0.0042			0.00000		
	15.2, 15.2	0.001	0.001	0.0038			0.00000		
	1.1, 14.2	0.072	0.001	0.0097			0.00069		
						0.01295	0.00385	3.367	0.527

Suspect “A” Genotype

39 probable genotypes

D19S433

Suspect A = 14, 16.2

$H_P = 1 * 0.013$

Allele Pair	Probability	Genotype Frequency	Prob * GenFreq
13,14	0.002	0.1082	0.00020
14.2, 16.2	0.270	0.0044	0.00118
14, 14	0.002	0.0498	0.00008
13, 14.2	0.017	0.0392	0.00068
14, 16.2	0.013	0.0120	0.00016
13, 16.2	0.018	0.0131	0.00023
etc…	etc…	etc…	etc…
		Sum	**0.00385**

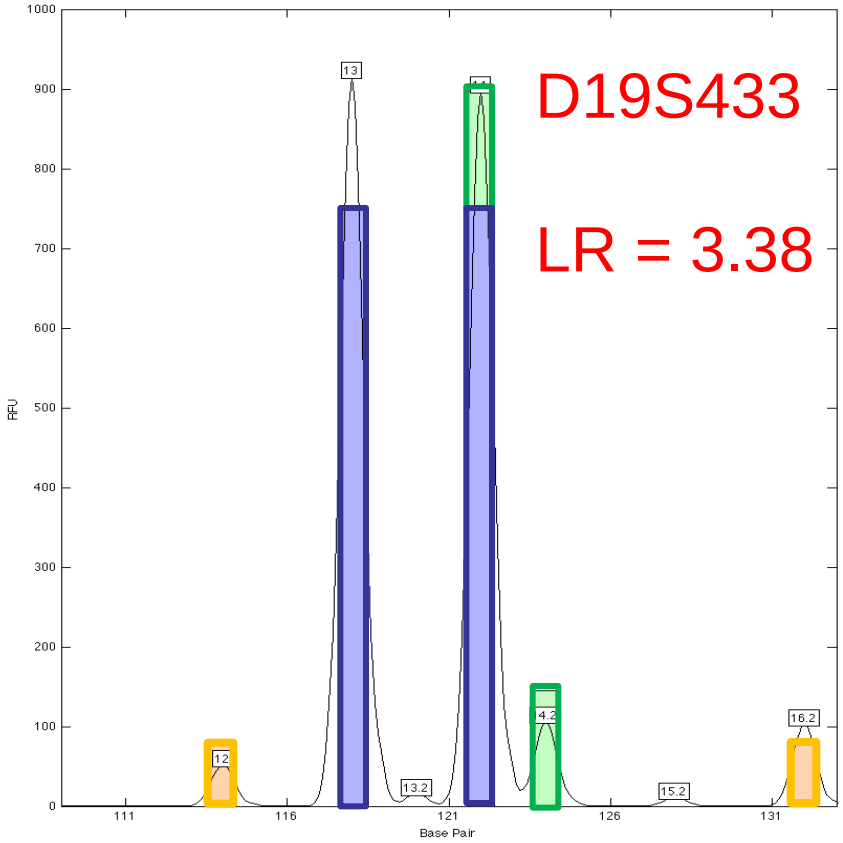
H_D

$$LR = \frac{0.013}{0.00385} = 3.38$$

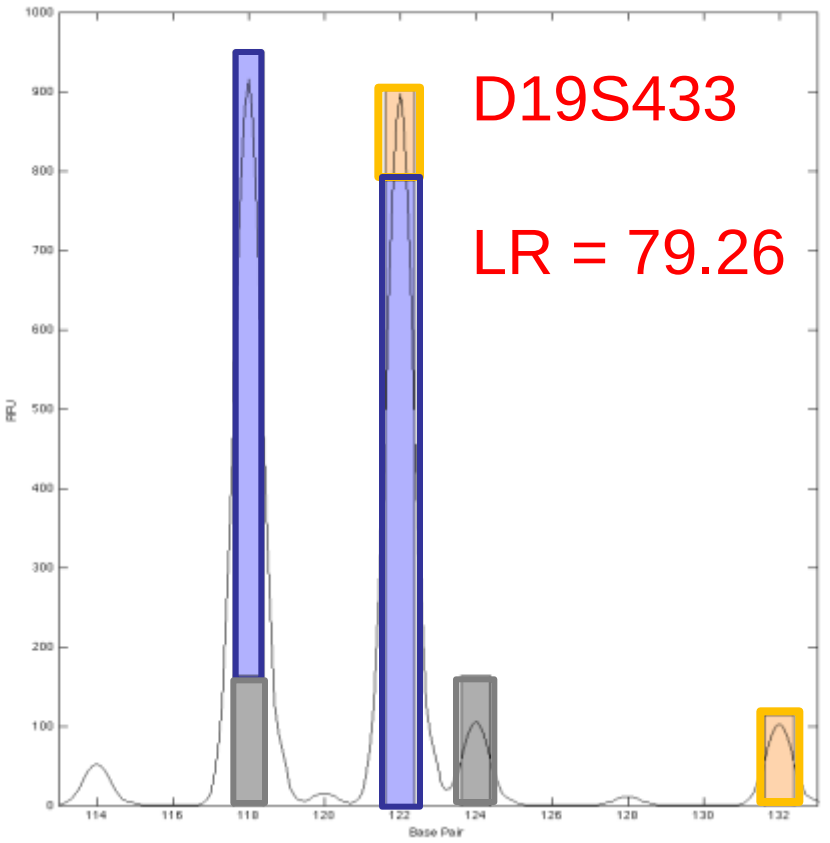
D19S433

No Conditioning (3 Unknowns)

No Conditioning



Conditioned on Victim



Profile - Combined log(LR)

Suspect A log(LR) = 8.03

Suspect B log(LR) = 7.84

Profile - Combined log(LR)

Suspect A log(LR) = 18.72

Suspect B log(LR) = 19.45

Exploring the Capabilities

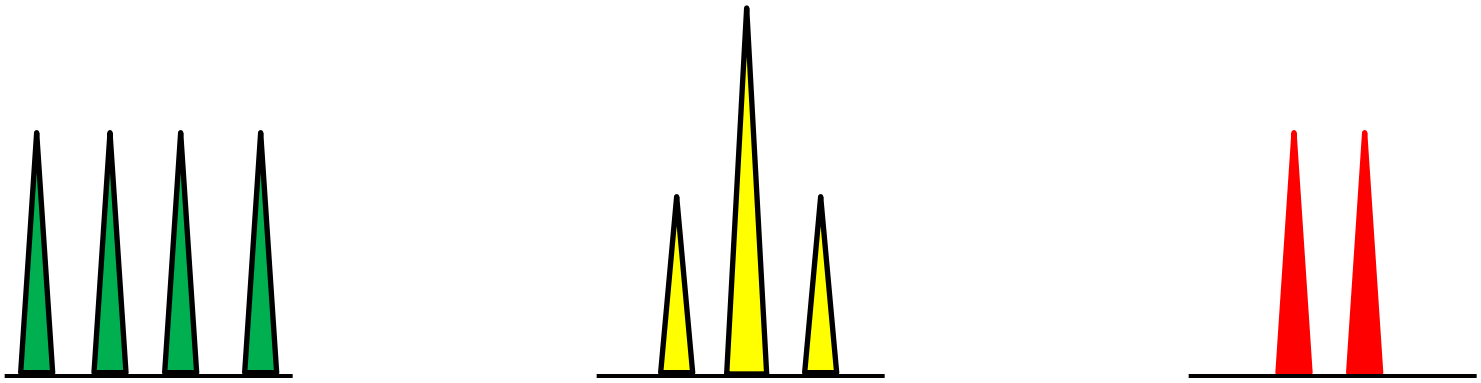
- **Degree of Allele Sharing**
- **Mixture Ratios**
- DNA Quantity

Mixture Data Set

- Mixtures of pristine male and female DNA amplified at a total concentration of 1.0 ng/μL using Identifiler (standard conditions).
- Mixture ratios ranged from 90:10, 80:20, 70:30 60:40, 50:50, 40:60, 30:70, 20:80, and 10:90
- Each sample was amplified twice.

Mixture Data Set

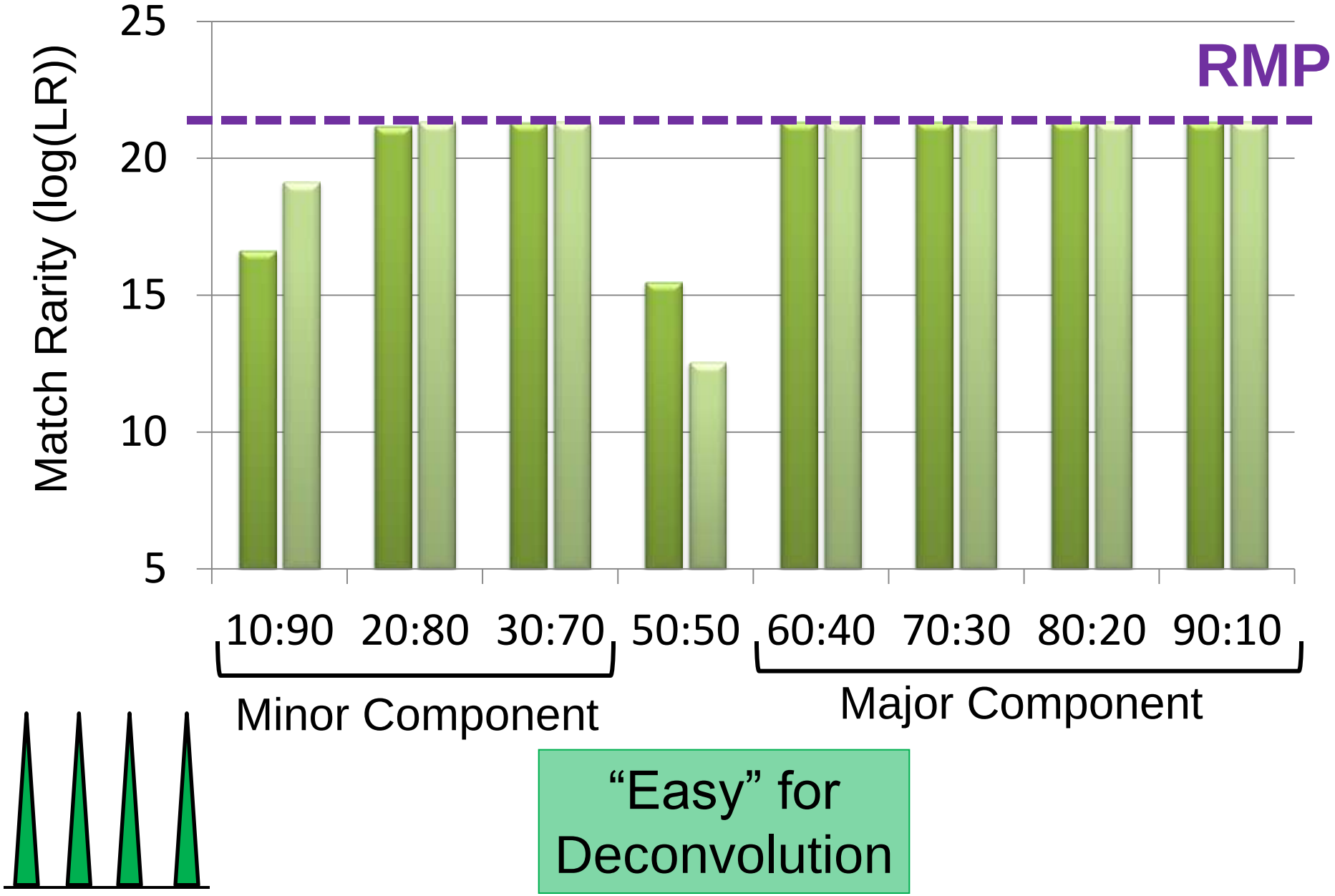
- Three different combinations:



"Low" Sharing	"Medium" Sharing	"High" Sharing
4 alleles – 10 loci	4 alleles – 3 loci	4 alleles – 0 loci
3 alleles – 5 loci	3 alleles – 8 loci	3 alleles – 6 loci
2 alleles – 0 loci	2 alleles – 4 loci	2 alleles – 8 loci
1 allele – 0 loci	1 allele – 0 loci	1 allele – 1 loci

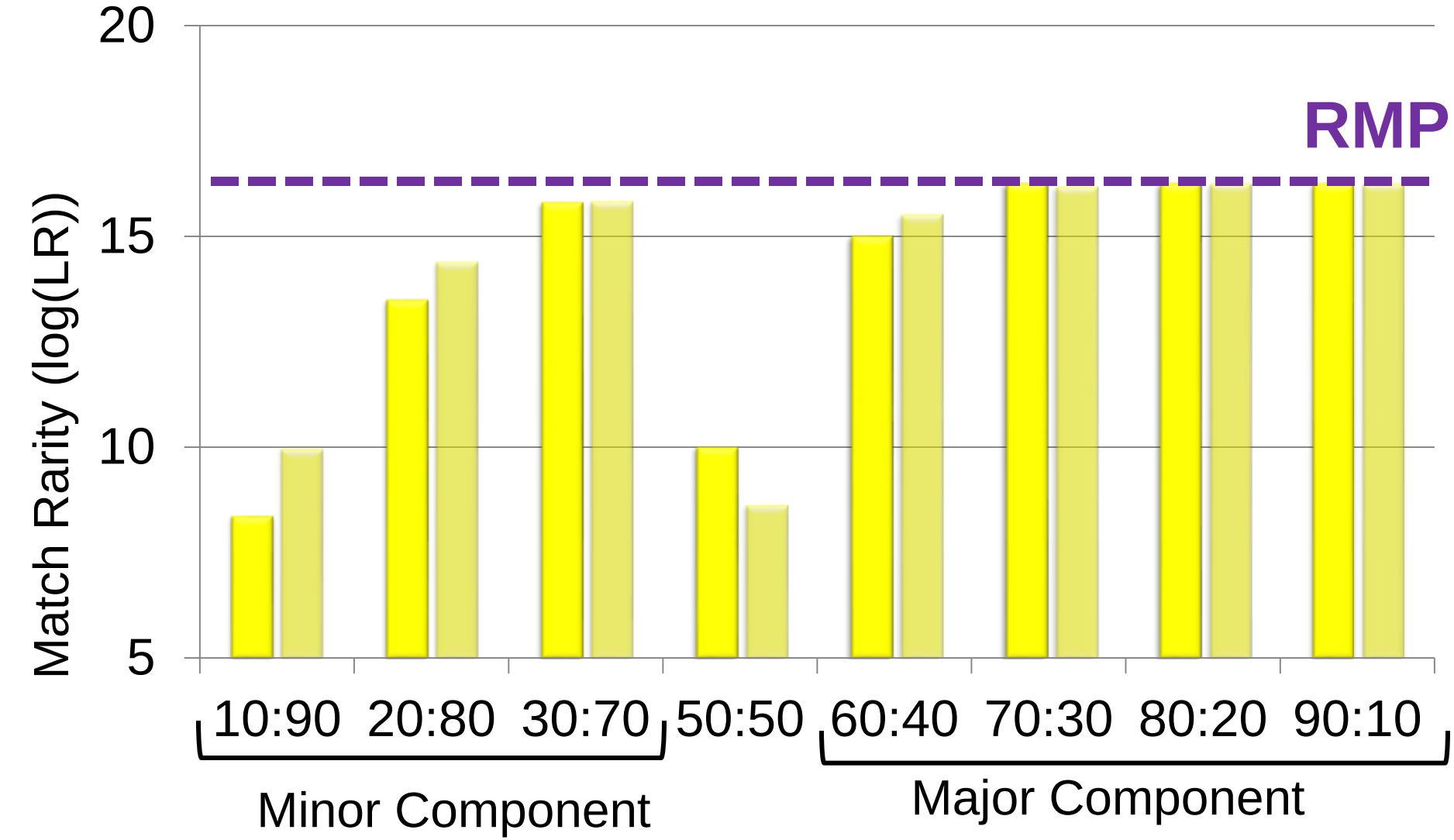
Virtual MixtureMaker - http://www.cstl.nist.gov/strbase/software.htm

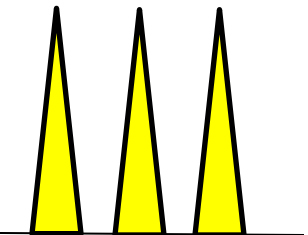
Match Score in Duplicate Runs



"Easy" for Deconvolution

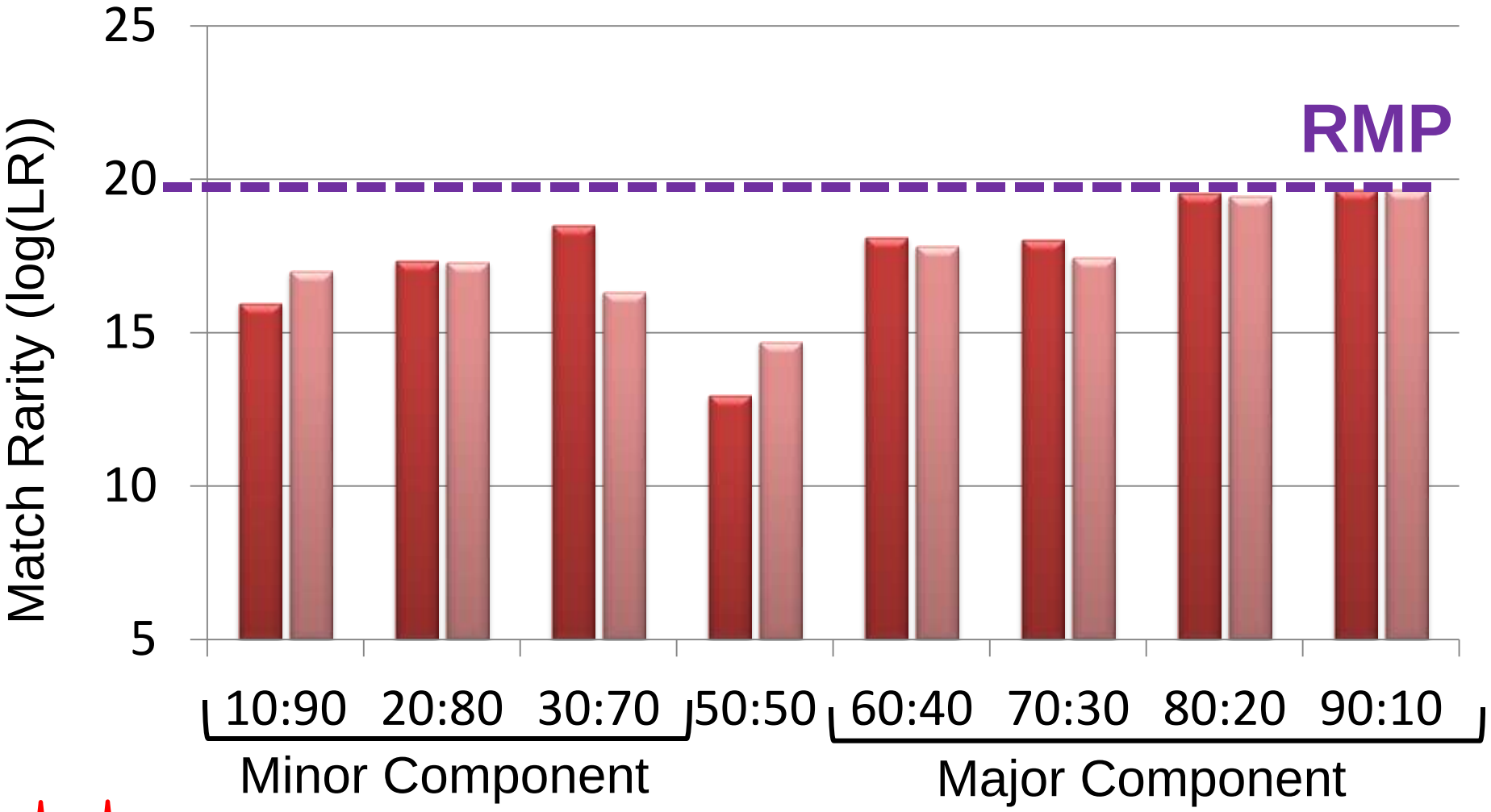
Match Score in Duplicate Runs



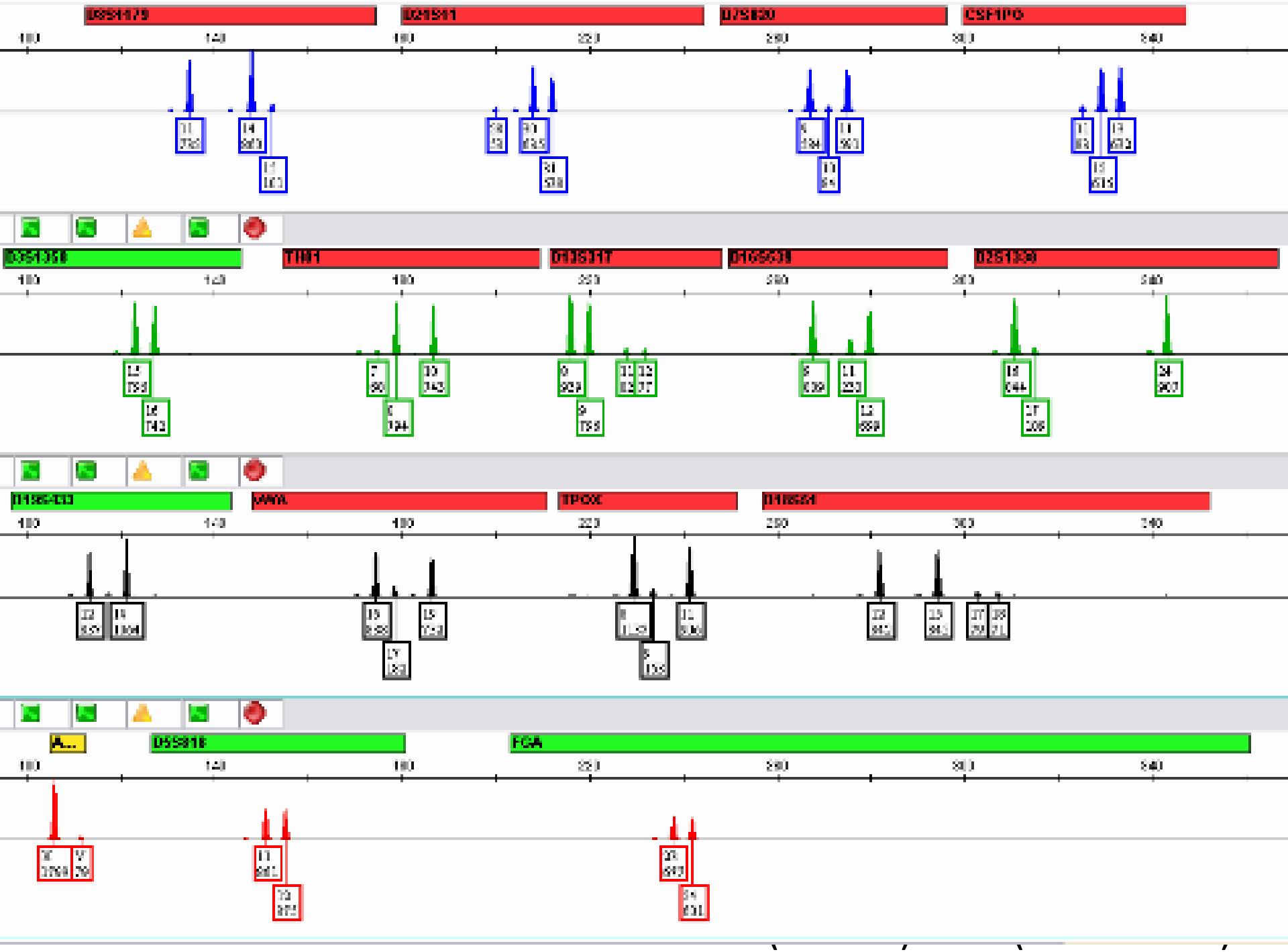


"Challenging" for Deconvolution

Match Score in Duplicate Runs



"Difficult" for Deconvolution


D5S818
D21S11
D7S820
CSF1PO
TH01
D13S317
D16S539
D2S1338
AMEL
TPOX
FGA

Exploring the Capabilities

- Degree of Allele Sharing
- Mixture Ratios
- **DNA Quantity**

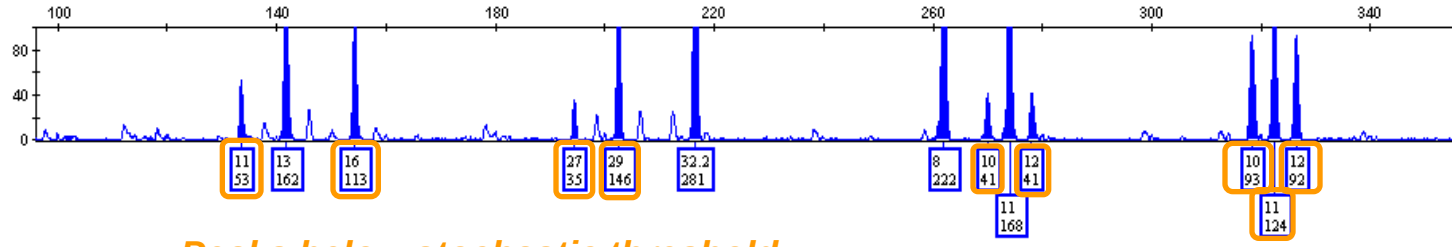
Identifiler

125 pg total DNA

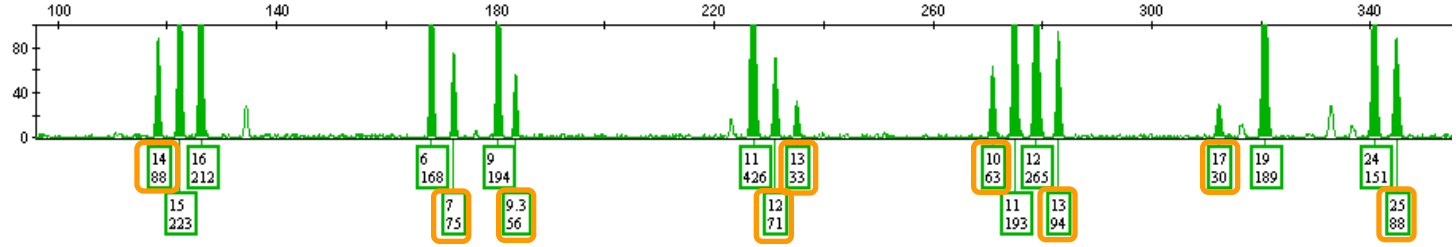
AT = 30 RFU

ST = 150 RFU

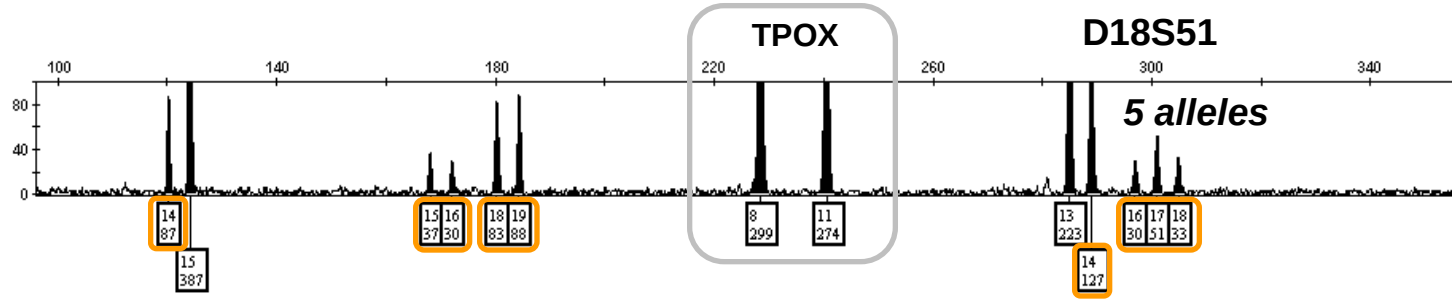
Stutter filter off

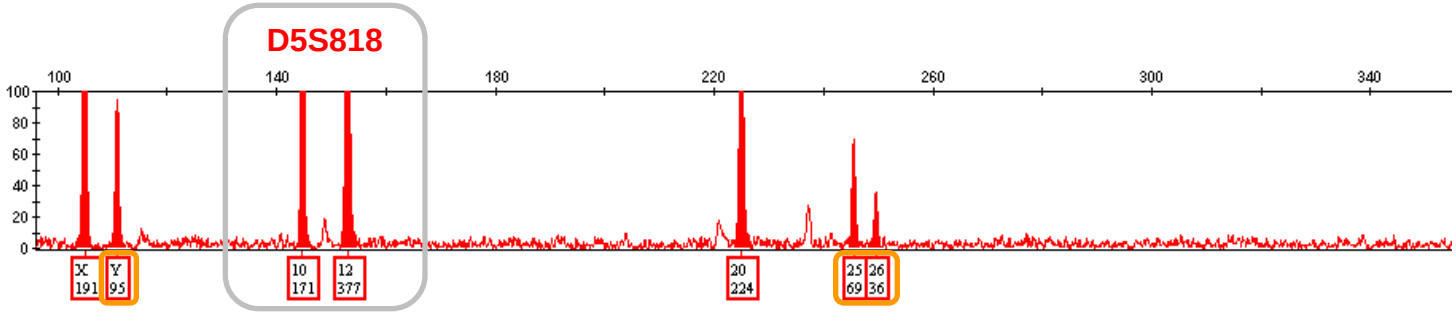


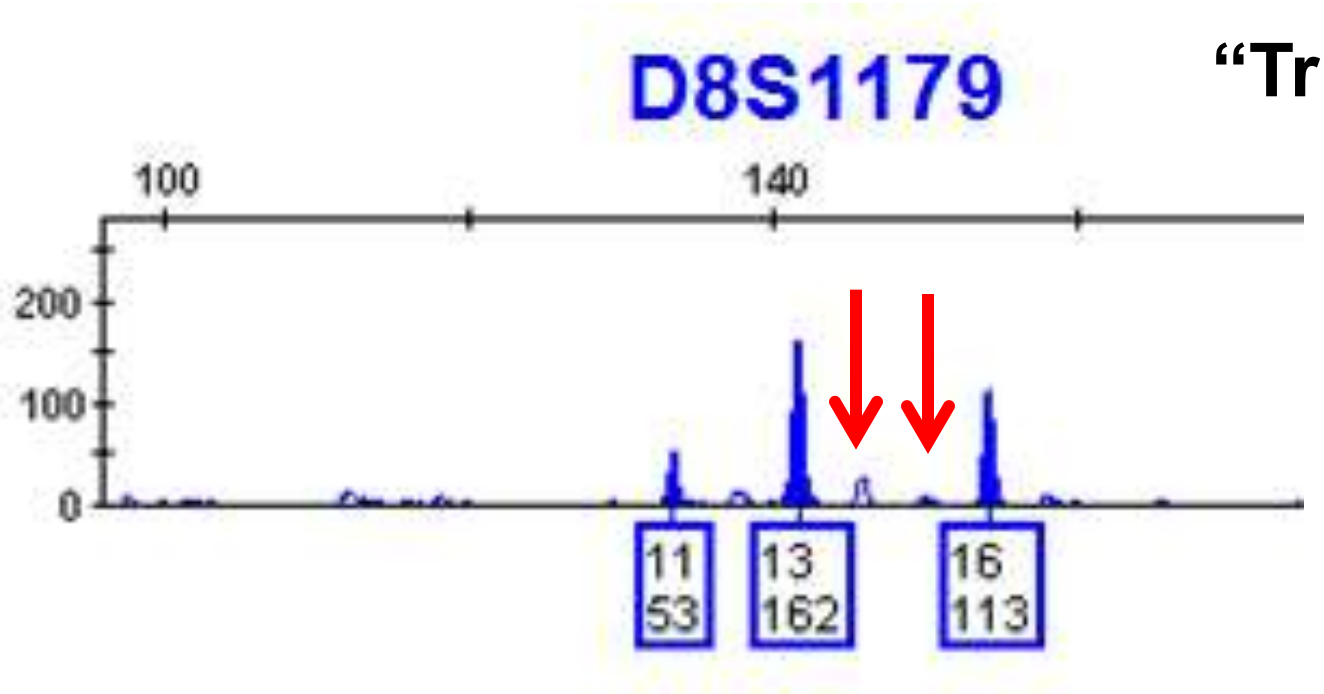
Peaks below stochastic threshold



y-axis zoom to 100 RFU







“True Genotypes”

A = 13, 16

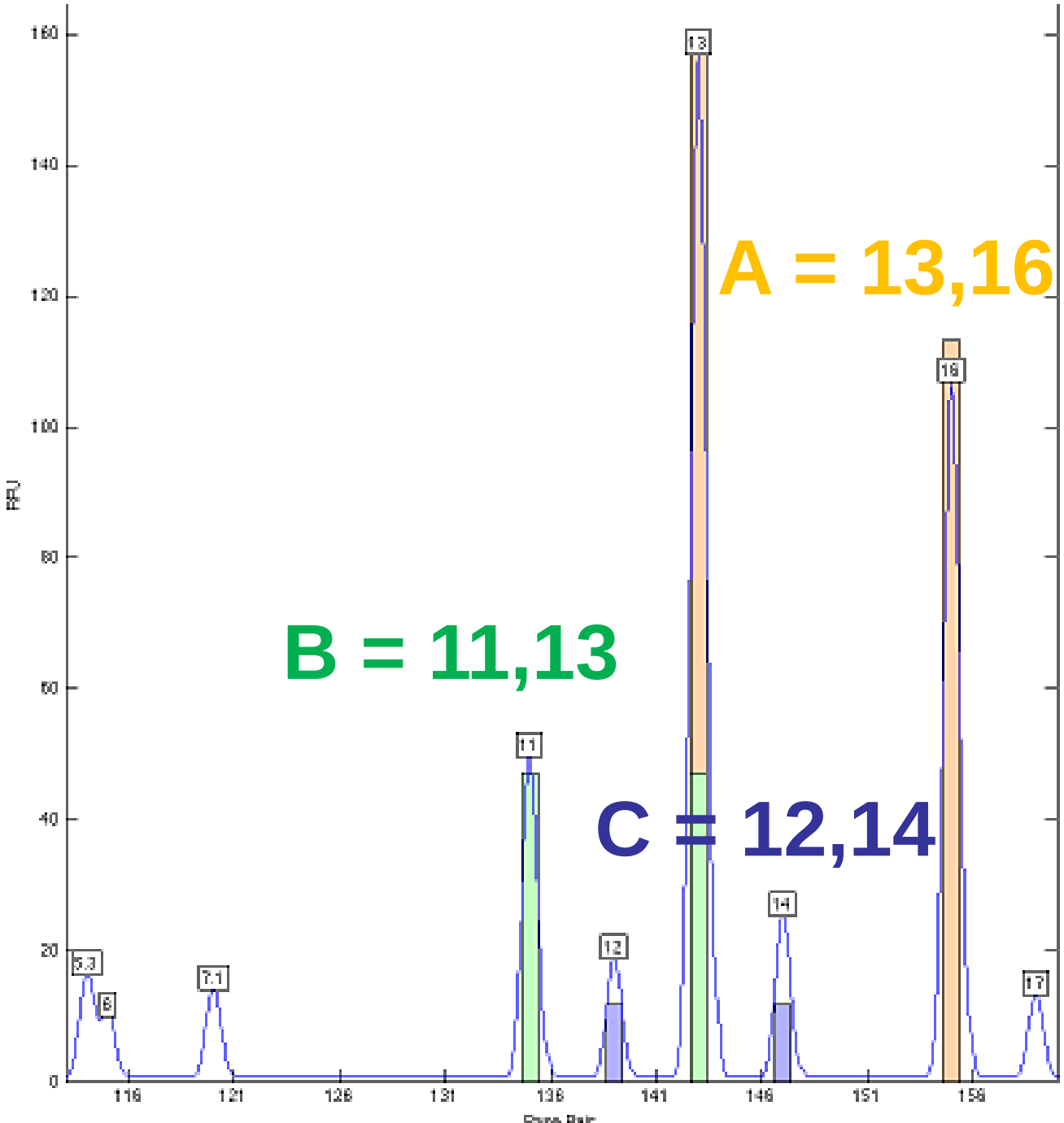
B = 11, 13

C = 14, 15

3 person Mixture – No Conditioning

Major Contributor ≈ 83 pg input DNA

2 Minor Contributors ≈ 21 pg input DNA

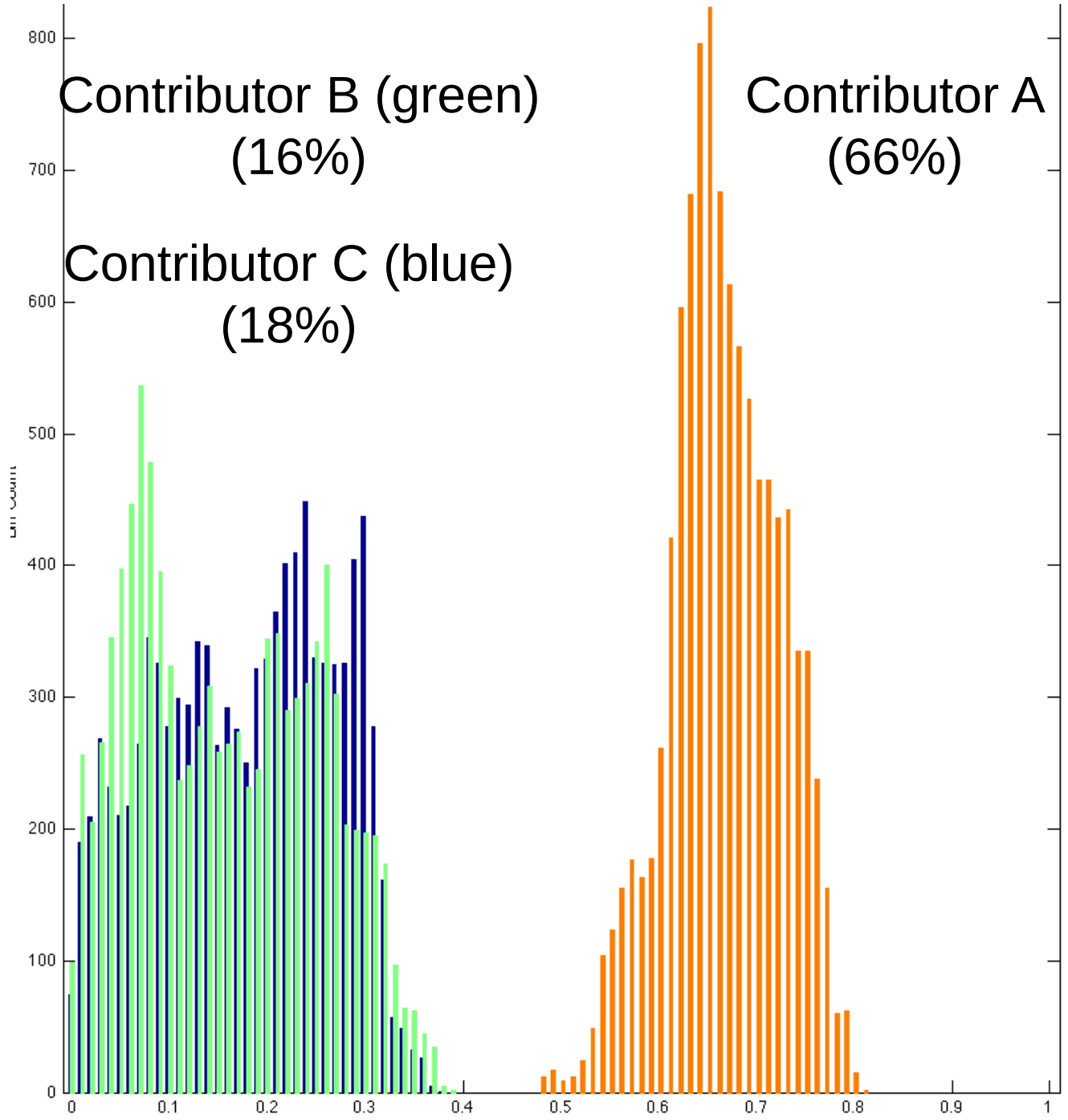


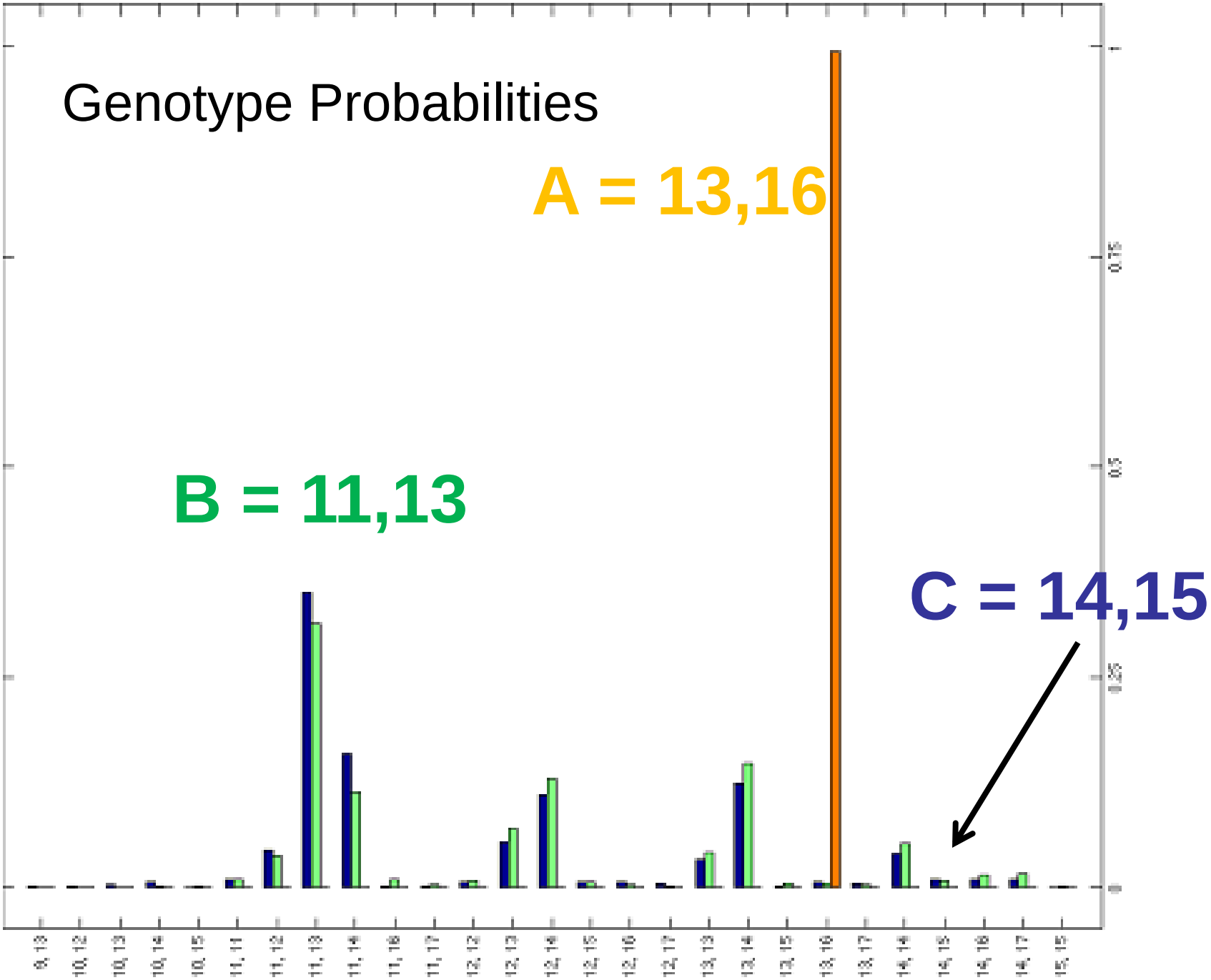
"True Genotypes"

A = 13,16

B = 11,13

C = 14,15


Contributor B (green)
(16%)
Contributor C (blue)
(18%)
Contributor A
(66%)
800
700
600
500
400
300
200
100
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Genotype Probabilities
A = 13,16
B = 11,13
C = 14,15
0.75
0.5
0.25
0
9, 13
10, 12
10, 13
10, 14
10, 15
11, 11
11, 12
11, 13
11, 14
11, 16
11, 17
12, 12
12, 13
12, 14
12, 15
12, 16
12, 17
13, 13
13, 14
13, 15
13, 16
13, 17
14, 14
14, 15
14, 16
14, 17
15, 15

Results for Contributor A (male)

Locus	Allele Pair	Probability Likelihood	Genotype Frequency	Suspect	H_p Numerator	H_d Denominator	LR
CSF1PO	10, 11	0.572	0.1292			0.07395	
	11, 12	0.306	0.2133	1	0.30563	0.0652	
	10, 12	0.12	0.1547			0.01861	
					0.30563	0.15791	1.935
D13S317	11, 11	1	0.1149	1	1	0.11488	8.704
D8S1179	13, 16	0.998	0.0199	1	0.99786	0.0199	49.668

The match rarity between the evidence and suspect is 1.21 quintillion

Results for Contributor B (female)

Locus	Allele Pair	Probability Likelihood	Genotype Frequency	Suspect	H_p Numerator	H_d Denominator	LR
D8S1179	11, 13	0.073	0.0498	1	0.07338	0.00366	
	11, 14	0.034	0.0271			0.00092	
	13, 14	0.006	0.0996			0.00065	
	12, 14	0.011	0.0606			0.00068	
	12, 13	0.005	0.1115			0.0006	
	11, 12	0.018	0.0303			0.00054	
	14, 14	0.004	0.0271			0.00012	
	13, 13	0.003	0.0916			0.00031	
	14, 16	0.003	0.0108			0.00003	
	14, 15	0.001	0.0379			0.00003	

etc…

9.197

The match rarity between the evidence and suspect is 1.43 million

Results for Contributor C (male)

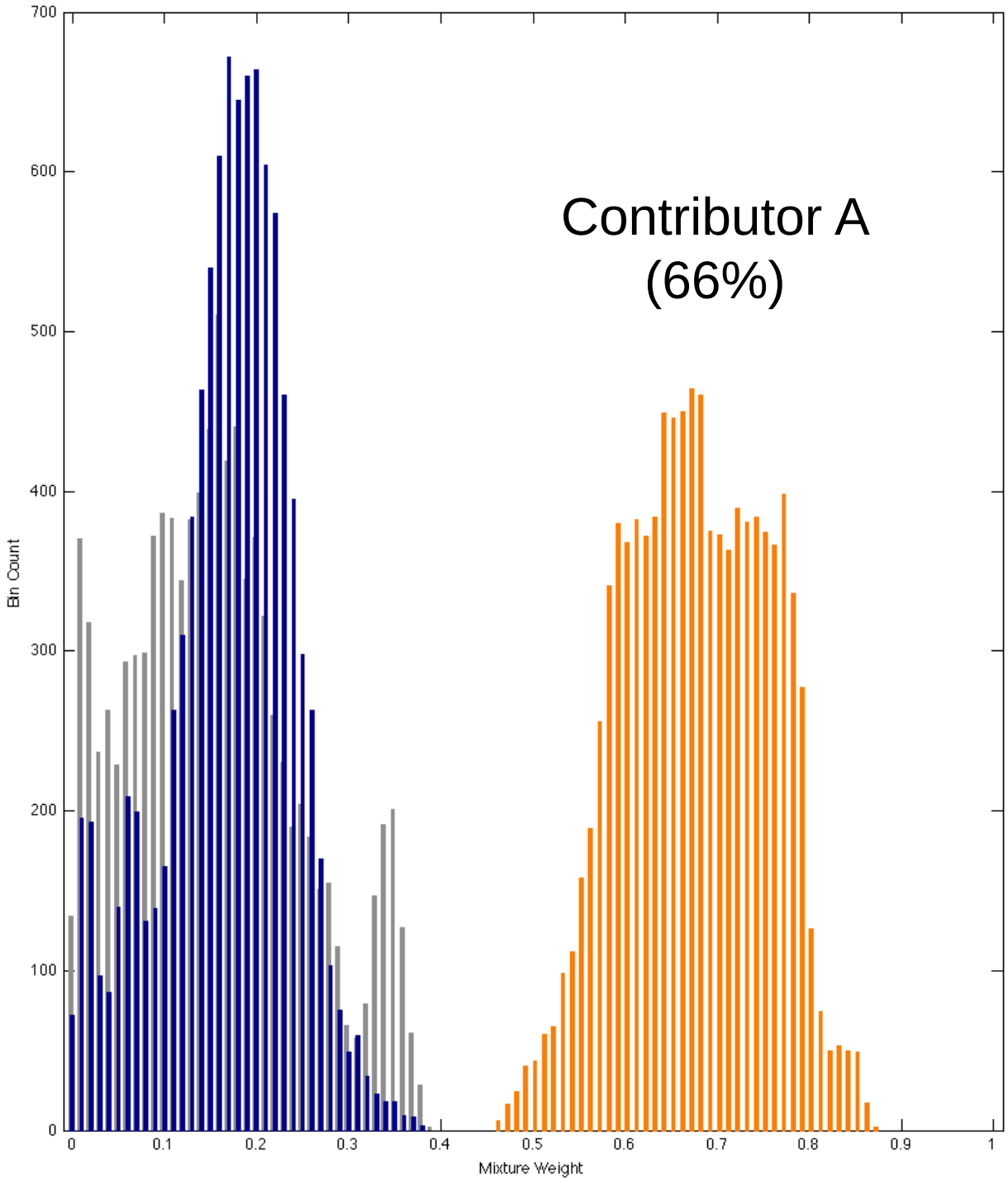
Locus	Allele Pair	Probability Likelihood	Genotype Frequency	Suspect	H_p Numerator	H_d Denominator	LR
D8S1179	11, 13	0.056	0.0498			0.00279	
	13, 14	0.007	0.0996			0.00066	
	12, 14	0.011	0.0606			0.00068	
	11, 14	0.021	0.0271			0.00056	
	12, 13	0.006	0.1115			0.00066	
	14, 14	0.005	0.0271			0.00013	
	etc…	etc…	etc…			etc…	
	14, 15	0.001	0.0379	1	0.00056	0.00002	
	12, 15	0.001	0.0424			0.00003	
	etc…	etc…	etc…			etc…	
	10, 15	0	0.0227			0.00001	
					0.00056	0.00665	**0.084**

The match rarity between the evidence and suspect is 9.16 thousand

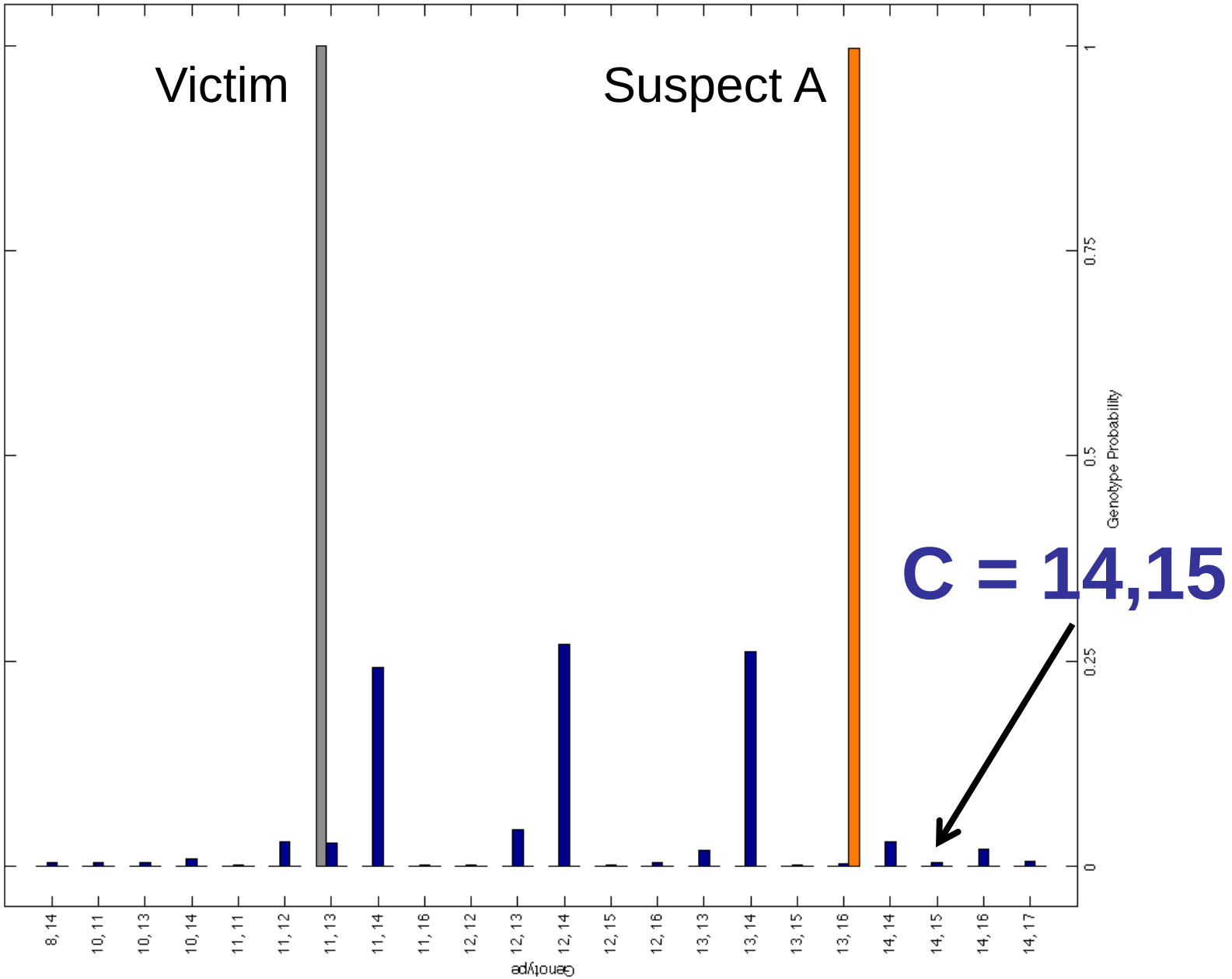
Contributor B (gray)
(16%)

Contributor C (blue)
(18%)

Conditioned on the Victim



The Power of Conditioning



The Power of Conditioning

	LR (no conditioning, 3unk)
Contributor A	**1.21 Quintillion**
Contributor B (victim)	**1.43 Million**
Contributor C	**9.16 Thousand**

	LR (conditioned on victim + 2unk)
Contributor A	**1.32 Quintillion**
Contributor B (victim)	**2.19 Million**
Contributor C	**59.8 Thousand**

Ranged from 1.13 to 800K

Summary

- True Allele utilizes probabilistic genotyping and makes better use of the data than the RMNE approach.

- However, the software is computer intensive. On our 4 processor system, it can take 12-16 hours to run up to four mixture samples.

Summary

- **Allele Sharing:** Stacking of alleles due to sharing creates more uncertainty.

- **Mixture Ratio:** With “distance” between the two contributors, there is greater certainty. Generally, True Allele performs better than RMNE and the classic LR with low level contributors.

Summary

- **DNA Quantity:** Generally, with high DNA signal, replicates runs on True Allele are very reproducible.
- However, with low DNA signal, higher levels of uncertainty are observed (as expected).
- There is a need to determine an appropriate threshold for an inclusion log(LR).

Future Work

- More work will be performed with low level, complex (3 and 4 person) mixtures.

Thank You!

Forensic DNA Team


John Butler


Mike Coble


Becky Hill


Margaret Kline

Data Analysis Support


Dave Duewer

DNA Biometrics Team


Pete Vallone


Erica Butts


Kristen Lewis O'Connor

Funding from the **National Institute of Justice (NIJ)** through NIST Office of Law Enforcement Standards

Funding from the **FBI S&T Branch** through NIST Information Access Division

ILM – Walther Parson and Richard Scheithauer

Cybergenetics – Dr. Mark Perlin, Dr. Ria David, Bill Allan, Matt Legler, and Erin Turo