

### Exploring DNA Interpretation Software Using the PROVEDIt Dataset

Sarah Riman<sup>1</sup>; Hari Iyer<sup>2</sup>; Peter M. Vallone<sup>1</sup>

<sup>1</sup>Applied Genetics Group, National Institute of Standards and Technology

<sup>2</sup>Statistical Design, Analysis, and Modeling Group, National Institute of Standards and Technology



1

---

---

---

---

---

---

---

---

#### Acknowledgments and Disclaimers

I would like to thank: Øyvind Bleka (Oslo University Hospital), Zane Kerr and Judi Morawitz (ESR), and Steven Myers (CAL DOJ) for meaningful discussions on data analysis using the software.

**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 or the U.S. Department of Commerce.

**NIST Disclaimer** Certain commercial products and instruments are identified in order to specify experimental procedures as completely as possible. In no case does such an identification imply a recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that any of these products are necessarily the best available for the purpose.



2

---

---

---

---

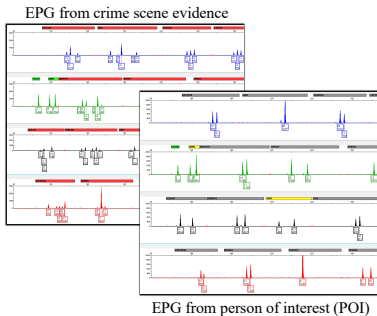
---

---

---

---

#### DNA mixture interpretation



3

---

---

---

---

---

---

---

---

**DNA mixture interpretation**

EPG from crime scene evidence

**Likelihood Ratio**

EPG from person of interest (POI)

4

---

---

---

---

---

---

---

---

**Likelihood ratio (LR)**

$$LR = \frac{Pr(E|H_p, I)}{Pr(E|H_d, I)}$$

$H_p$ : the DNA from the POI **IS** in the mixture  
 $H_d$ : the DNA from the POI **IS NOT** in the mixture  
 $I$ : background information  
 $E$ : evidence

5

---

---

---

---

---

---

---

---

**Different approaches to assess LR**

Binary → Semi-Continuous → Continuous → LR

Probabilistic genotyping software

FST  
Lab Retriever  
LiRa  
LRmix/LRmix studio

(Proprietary) STRmix  
(Open source) EuroForMix

DNAmixtures  
DNA Mixture Solution  
Bullet and BulletProof  
GenoProof Mixture 3

6

---

---

---

---

---


---

---

---

**Motivation**

Understand similarities/differences between two LR systems, by applying two fully continuous PROBGEN models (STRmix and EFM) to ground truth known mixture profiles available publicly (PROVEDIt data archive – Catherine Grgicak).




---

---

---

---

---

---

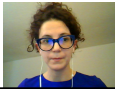
---

---

7

**Overview**

- Description of PROVEDIt dataset
- Defining LR system
- Discrimination performance check of the LR systems
- Evaluation of discrepancies between the LR systems




---

---

---

---

---

---

---

---

8

**PROVEDIt database**

Large publicly available database

Allows examination of probabilistic genotyping systems

Examine effect of analytical thresholds and peak detection parameters on downstream analysis

Assess approaches to evaluate STR signal (genotyping software packages and validation software)

Analyzed with different CE instrument types and injection times

Amplified with different STR kits DNA quantity (0.007-1 ng)

1 to 5 person mixtures varying


- contributor ratios
- DNA quality

Contains over 25,000 STR profiles

<https://fbi.com/forensic-science/brs/provedit>

Project Research Openness for Validation with Empirical Data

**PROVEDIt**




---

---

---

---

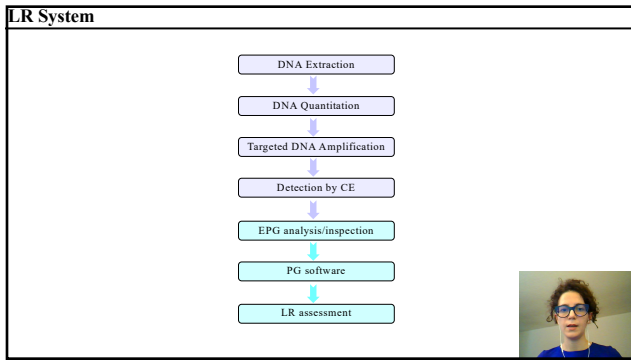
---

---

---

---

9



10

---

---

---

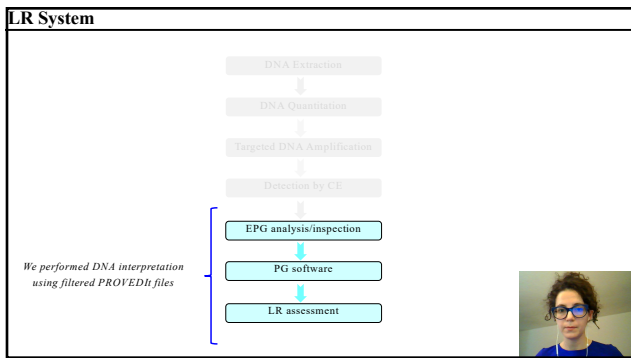
---

---

---

---

---



11

---

---

---

---

---

---


---

---

**LR System**

Data processing of PROVEDIt filtered files

- Provedit filtered files were analyzed in GeneMapper at an AT = 1 RFU
- Artefacts (pull-ups, minus A, and -2bp stutters at SE33) were filtered according to defined criteria set by the creators of the database
- Analyzed the filtered files using per dye ATs
- Removed OLS/-2bp stutters at DIS1656



12

---

---

---

---

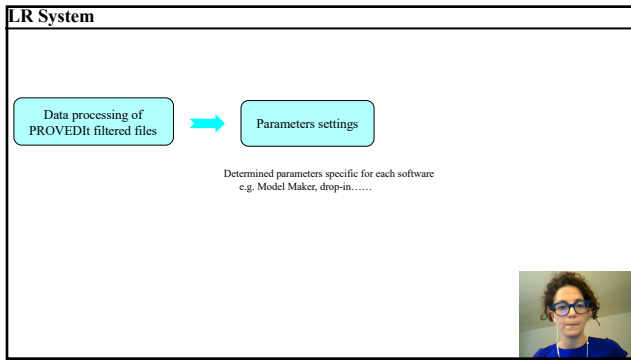
---

---

---

---





13

---

---

---

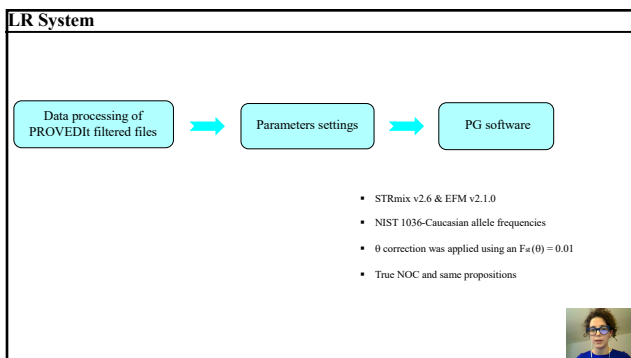
---

---

---

---

---



14

---

---

---

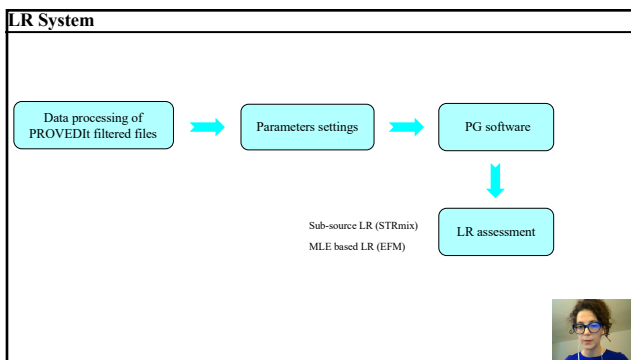
---

---

---

---

---



15

---

---

---

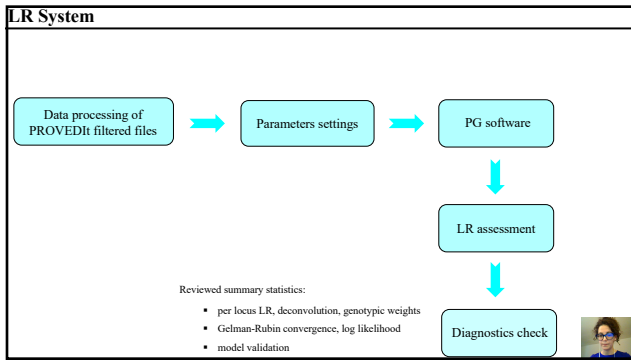
---

---

---

---

---



16

---

---

---

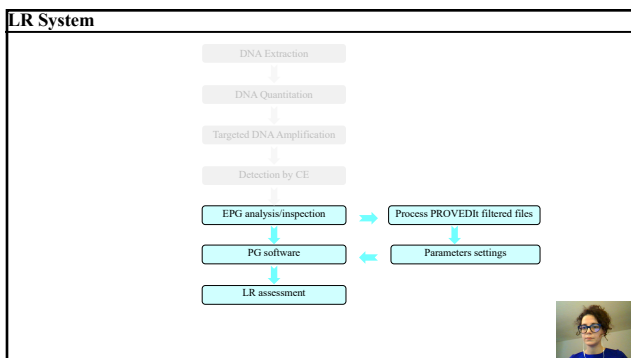
---

---

---

---

---



17

---

---

---

---

---

---

---

---

### Dataset used in our study

Number of contributors	Kit (PCR cycle no.)		CE instrument (injection time)				Minor Contributor DNA amount (µg)
	Pristine DNA	Degraded DNAse I	Degraded Sonication	Damaged UV	Inhibited Humic Acid		
<b>2P</b>		<b>GlobalFiler (29 cycles)</b>					<b>3500 (15 s)</b>
(16 unique individuals)	1:1	x	x	x	x	x	15; 30; 62; 125
	1:2	x	x				15; 30; 62; 125
	1:4	x	x	x	x	x	15; 30; 62; 125
	1:9	x	x	x	x	x	15; 30; 54; 62; 75
<b>Sum</b>	<b>88</b>	<b>228</b>	<b>44</b>	<b>104</b>	<b>108</b>	<b>572</b>	
<b>3P</b>		<b>GlobalFiler (29 cycles)</b>					<b>3500 (15 s)</b>
(21 unique individuals)	1:1:1	x	x	x	x	x	15; 30; 62; 125
	1:2:1	x	x				15; 30; 62; 125
	1:2:2	x	x				15; 30; 62; 125
	1:4:1	x	x	x	x	x	15; 30; 62; 125
	1:4:4	x	x	x	x	x	15; 30; 62; 83
	1:9:1	x	x				15; 30; 45; 62
1:9:9	x	x				15; 26; 30; 40	
<b>Sum</b>	<b>114</b>	<b>324</b>	<b>72</b>	<b>138</b>	<b>162</b>	<b>810</b>	

18

---

---

---

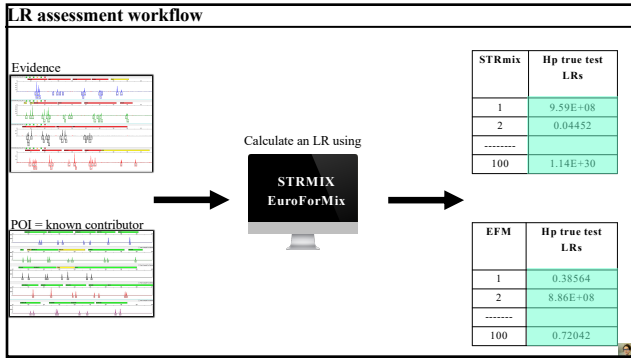
---

---

---

---

---



19

---

---

---

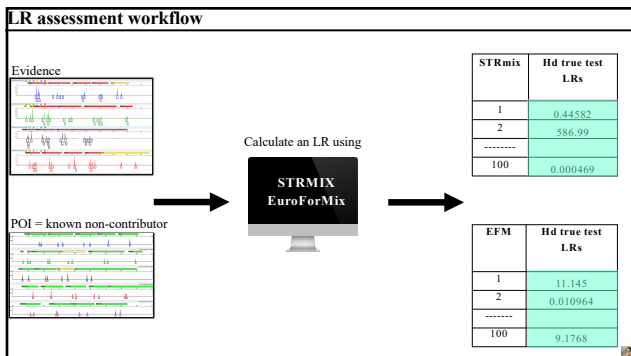
---

---

---

---

---



20

---

---

---

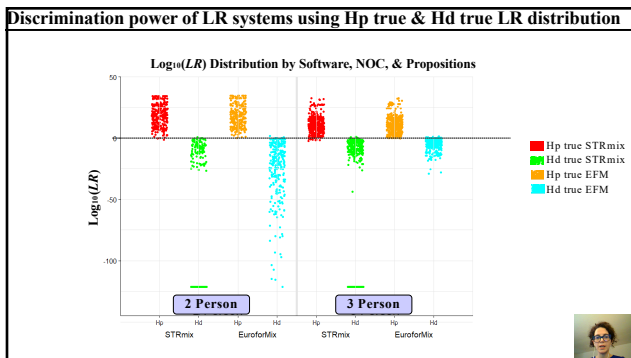
---

---

---

---

---



21

---

---

---

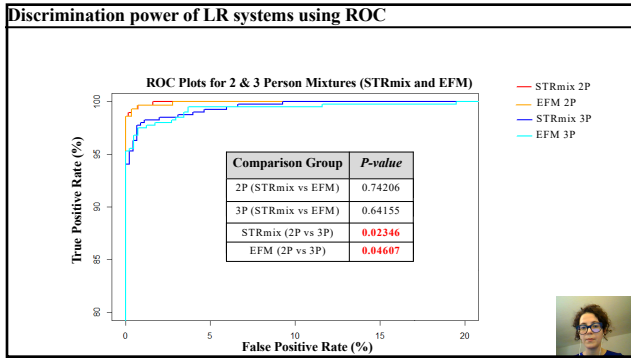
---

---

---

---

---



22

---

---

---

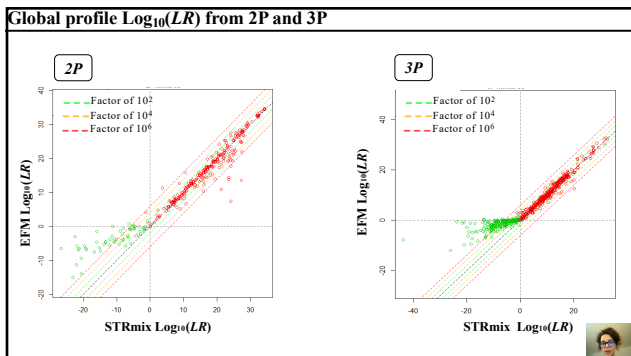
---

---

---

---

---



23

---

---

---

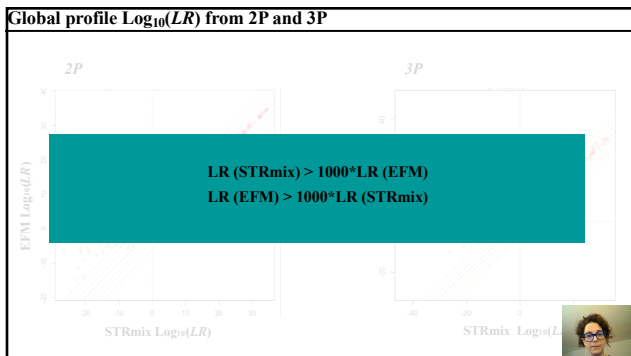
---

---

---

---

---



24

---

---

---

---

---

---

---

---

### Evaluation of discrepancies in $\text{Log}_{10}(LR)$ values between software

Differences observed in LR values can occur due to a combination of the following reasons:

- Nonconvergence of the Markov Chain Monte Carlo algorithms and maximum likelihood estimators
- Analyst decisions on what peaks to leave in and/or what peaks to remove from the EPG
- Different modeling assumptions
- Choice of parameters settings

**STRmix**

**Model Maker**

Model your laboratory's data

- Allelic variance
- Scatter variance
- Locus amplification variance

**Profiling Kits**

Manage multiple DNA profiling kits

- Detection thresholds
- Drop-in frequency's log-distribution

**EFM**

EFM Settings

Default detection threshold: 0.01

Default drop-in frequency: 0.001


Default amplification variance: 0.001

Default allele frequency: 0.01

Default mutation rate: 0.001

Default mutation model: (default, 1)

Max. iterations: 10



25

---

---

---

---

---

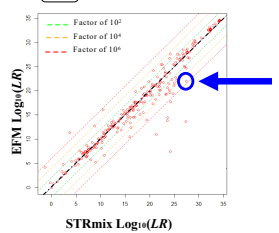
---

---

---


### LR (STRmix) > LR (EFM)

2P



Pristine DNA of total template amount 315 pg; Ratio 1:4

Software	$\text{Log}_{10}(LR)$ C1 (major)	$\text{Log}_{10}(LR)$ C2 (minor)
STRmix	27.61	27.428
EFM	27.90	21.992
STRmix - EFM	-0.29	<b>5.436</b>



26

---

---

---

---

---


---

---

---

### Per Locus LR of STRmix and EFM

Locus	STRmix	EFMv2.1	STRmix/EFM
D3S1338	10.00	13.42	0.745
vWA	3.84	3.396	1.131
D16S539	24.90	12.98	1.918
CSF1PO	15.50	14.44	1.073
TPOX	7.43	7.982	0.931
D8S1179	22.30	11.02	2.024
D21S11	21.20	15.49	1.369
D18S51	23.10	11.05	2.090
D20441	6.29	2.552	6.5
D19S433	8.84	7.011	6.1
TH01	3.80	11.07	1.251
FGA	6.53	6.77	0.963
<b>D22S1045</b>	<b>19.79</b>	<b>0.601149</b>	<b>17295.874</b>
D5S818	57.20	87.8	0.651
D18S317	4.33	4.103	1.055
D7S820	6.06	5.355	1.132
SE33	122.00	129.9	0.939
D10S1248	10.50	11.79	0.891
D1S1656	80.10	53.16	1.507
D12S591	137.00	135.4	1.012
D2S1338	21.90	23.73	0.923



27

---

---

---

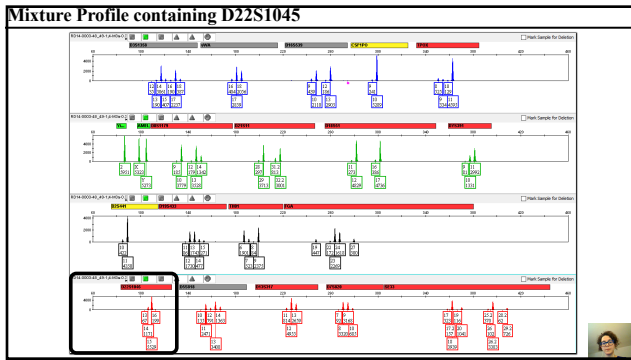
---

---

---

---

---



28

---

---

---

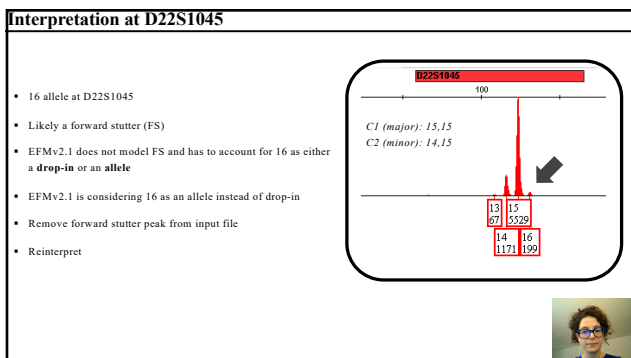
---

---

---

---

---



29

---

---

---

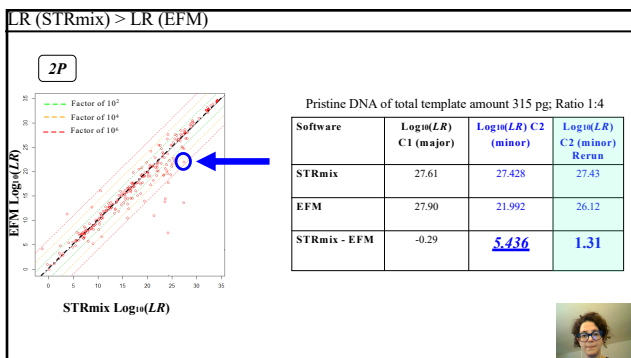
---

---

---

---

---



30

---

---

---

---

---

---

---

---

**Per Locus LR after EFM rerun**

Locus	STRmix	EFMv2.1	EFMv2.1 rerun
D3S1358	10.00	13.42	13.4
VWA	3.84	3.396	3.524
D16S539	24.90	12.98	13.68
CSF1PO	15.50	14.44	14.47
TPOX	7.43	7.982	7.995
D8S1179	22.30	11.02	10.85
D21S11	21.20	15.49	15.78
D18S51	23.10	11.05	11.9
D2S441	4.29	2.552	2.4
D19S433	8.84	7.011	6.96
TH01	13.80	11.03	10.83
FGA	15.00	673.5	640.3
<b>D22S1045</b>	<b>19.70</b>	<b>0.001139</b>	<b>16.22</b>
D5S818	57.20	87.8	80.52
D13S317	4.33	4.103	4.188
D7S820	6.06	5.355	5.144
SE33	122.00	129.9	117.5
D10S1248	10.50	11.79	12.44
D181656	80.10	53.16	54.72
D12S391	137.00	135.4	132.2
D2S1338	21.90	23.73	22.6

31

---

---

---

---

---

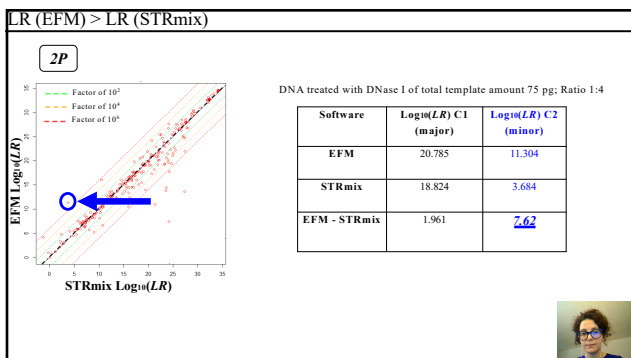
---

---

---

---

---



32

---

---

---

---

---

---

---

---

---

---

**Per Locus LR of EFM and STRmix**

Locus	EFMv2.1	STRmix	EFM/STRmix
D3S1358	1.272	0.076	16.73684
VWA	3.423	2.070	1.653623
D16S539	11.71	4.240	2.761792
CSF1PO	4.35	12.300	0.353669
TPOX	2.003	1.590	1.259748
D8S1179	2.874	3.150	0.912381
D21S11	3.555	7.950	0.44717
D18S51	4.432	10.800	0.41037
D2S441	5.009	4.370	1.146224
D19S433	4.667	3.150	1.481587
TH01	1.187	0.223	5.32287
FGA	4.181	0.197	10.53149
D22S1045	7.287	13.300	0.547895
D5S818	0.4505	0.055	8.190909
D13S317	<b>Inclusion</b> 18	0.077	<b>Exclusion</b> 697523
D7S820	0.007	4.040	~905198
SE33	0.004	1.220	3.24918
D10S1248	7.455	5.680	1.31252
<b>D181656</b>	<b>1.36</b>	<b>0.003</b>	<b>451.8272</b>
D12S391	33.76	41.600	0.811538
D2S1338	6.772	11.500	0.58887

33

---

---

---

---

---

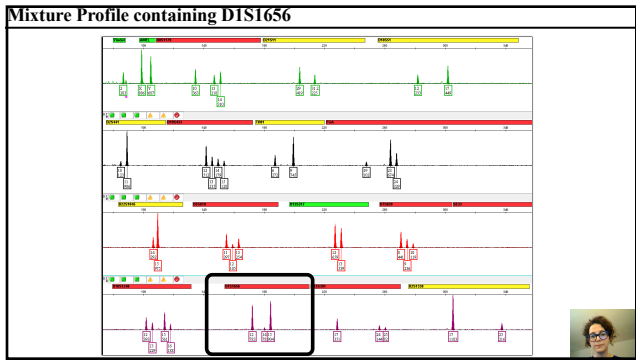
---

---

---

---

---



34

---

---

---

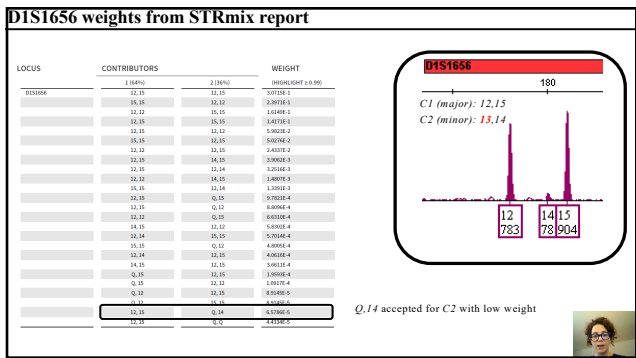
---

---

---

---

---



35

---

---

---

---

---

---

---

---

**Summary**

- Both LR systems have equal ability in discriminating between known contributors and known non-contributors.
- However, that does not imply that both LR systems are producing equal LR values or agreeing when the same profile is being interpreted.
- Differences observed in LR values can occur due to a combination of the following reasons:
  - Analyst decisions on what peaks to leave in and/or what peaks to remove from the EPG
  - Different modeling assumptions
  - Choice of parameters settings
  - Nonconvergence of the algorithms

36

---

---

---

---

---

---

---

---