Defense Forensic Science Center

Probabilistic Genotyping and the Bench Scientist – Using STRmix in Everyday Casework

Tim Kalafut
Forensic Biologist
Green Mountain Conference, August, 2016

Disclaimer

The opinions or assertions contained herein are the private views of the author and are not to be construed as official or as reflecting the views of the Department of the Army or the Department of Defense.

Names of commercial manufacturers or products included are incidental only, and inclusion does not imply endorsement by the authors, DFSC, OPMG, DA or DoD.

Unless otherwise noted, all figures, diagrams, media, and other materials used in this presentation are created by the respective author(s) and contributor(s) of the presentation and research.

Overview

- The process
- A bit of Bayes' Theorem
- What STRmix does
- The mixture and results
- Building the LR result by hand

Goal of Joint Presentation

- Lots of buzz right now in Forensic DNA on probabilistic genotyping/statistics
- At least one state is doing wholesale do-overs on data that would be best solved by a probabilistic approach
- But what does this mean to the average bench scientist
- This talk will show how STRmix™ and TrueAllele™ are integrated into the work flow at two different laboratories

Ground Rules

- This is not a competition between labs or software
- There are other software packages available – only you can decide what best fits your needs
- This is not intended to be a training session on probabilistic software, just a glimpse into what it does and how it's normally used

The Sample

- The goal was to find a "neutral" sample provided by a lab that has neither STRmix nor TrueAllele
- Some requirements:
  - 3130/fsa data
  - Identifiler
  - Reasonable mixture with some drop-out
- We failed to find a 3rd party sample – so the mixture is an old USACIL validation sample
- Slight disclaimer – We have STRmix validated for Identifiler+™, and this sample is Identifiler™

Analysis of the .fsa in the STRmix Lab

- Step 1 – Analyze the .fsa files just like always
- Use the validated analytical thresholds (AT)
- Clean up artifacts if any, and evaluate the overall quality of the profile
- Interpret the sample
  - Number of contributors
  - Inclusions/exclusions
  - USACIL only uses STRmix if there is a probative inclusion

Analysis of the .fsa in the STRmix Lab

- USACIL uses channel specific AT
  - Blue – 33
  - Green – 42
  - Yellow – 72
  - Red – 77

Caveat: Drop-out is directly related to AT; if you use a lower AT, the LR of this sample – or any other – will go up. We know we have more drop out in yellow and red, but those channels have more noise

- STRmix models stutter as allele specific, so USACIL built software (ArmedXpert™) that uses allele specific stutter filters on the egram when looking at profiles
  - y=mx+b
  - m is locus specific slope
  - x is the allele
  - b is y intercept plus 2SD

Analysis of the .fsa in the STRmix Lab

- All samples analyzed in OSIRIS v2.6
- ArmedXpert v3.8.0.2 used for visualizing egrams, artifact determination/edits, etc.
- This is done for all controls, blanks, Q's and K's just like pre-STRmix days
- We make a table for STRmix to use
  - It's just a simple .txt file
  - Q samples have stutter peaks in the table
  - K samples are only the alleles

STRmix Considerations

- Step 2 – Run STRmix
- We have to tell STRmix the number of contributors
- We load the Q sample and the appropriate K references
- We have to decide the H1 (numerator) and H2 (denominator) propositions (or Hp and Hd)
- There may be multiple choices –
  - Each set of propositions results in a different likelihood ratio (LR)
  - Each LR means something different

STRmix LR

- Bayes' Theorem

$$\underbrace{\Pr(H_1|E)}_{\text{Posterior odds}} = \underbrace{\Pr(E|H_1)}_{\text{Likelihood}} \times \underbrace{\Pr(H_1)}_{\text{Prior odds}}$$

H1 or Hp LR Numerator

$$\Pr(H_2|E) = \Pr(E|H_2) \times \Pr(H_2)$$

H2 or Hd LR Denominator

- So you end up with

$$\underbrace{\frac{\cancel{\Pr(H_1|E)}}{\cancel{\Pr(H_2|E)}}}_{\text{Posterior is for the Jury to decide – Guilt or Innocence}} = \underbrace{\frac{\Pr(E|H_1)}{\Pr(E|H_2)}}_{\text{STRmix does this LR for DNA evidence}} \times \underbrace{\frac{\cancel{\Pr(H_1)}}{\cancel{\Pr(H_2)}}}_{\text{Prior is the other evidence in the case - LP, Witnesses, etc}}$$

Setting up the LR

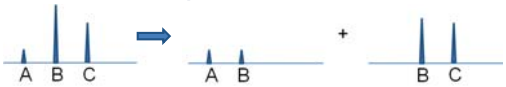
- We need some relevant information:

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

- Is it reasonable to assume someone? Both?
  - If our evidence came from V vag swab:
    - LR = V+S / V+U
  - If our evidence came from S underwear:
    - LR = S+V / S+U
  - If our evidence came from neutral crime scene
    - LR = S+V / U+U
  - Under most circumstances we would never do
    - LR = S+U / U+U or LR = V+U / U+U

Quick View of How STRmix Works

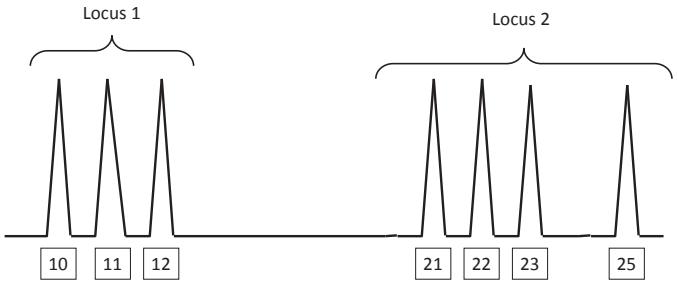
- We human experts are used to looking at the peak pattern, and then determining which genotypes could make that pattern:



- STRmix does the opposite:
  - We tell it how many contributors
  - It knows detected peaks from **observed** egram
  - It starts with genotypes, and applies biology
  - It comes up with an **expected** egram based on the values it guesses for various biological parameters

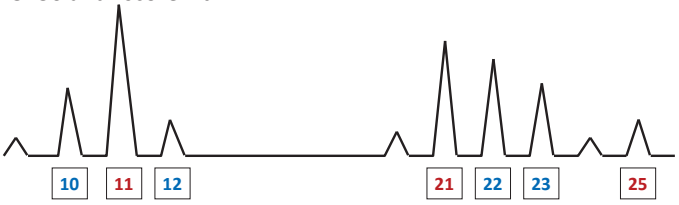
Quick View of How STRmix Works

-Here is a simple 2 locus/2 person mixture
-We tell STRmix there are 2 contributors, Locus 1 has alleles 10, 11, and 12 and Locus 2 has alleles 21, 22, 23, and 24
-Now STRmix starts "guessing" – Contributor 1 is 10,12/22,23 and Contributor 2 is 11,11/21,25



Quick View of How STRmix Works

-STRmix™ started with 10,12/22,23 donor and an 11,11/21,25 donor
-After guessing Major/minor, degradation, stutter, etc it has this **Expected Profile**
-Now it will now compare it to the **Observed Profile** from the 3130 and "score" it



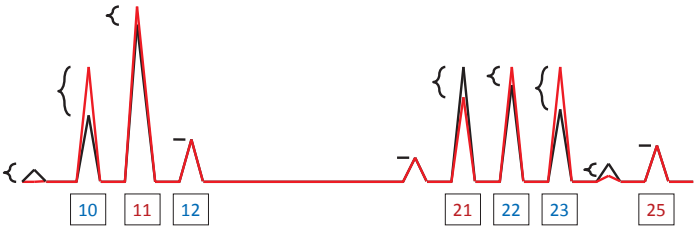
Think of it as a giant "hot/cold" game – genotypes of warmer guesses get points, when all the points handed out, they're normalized, best fits get most weight (%)

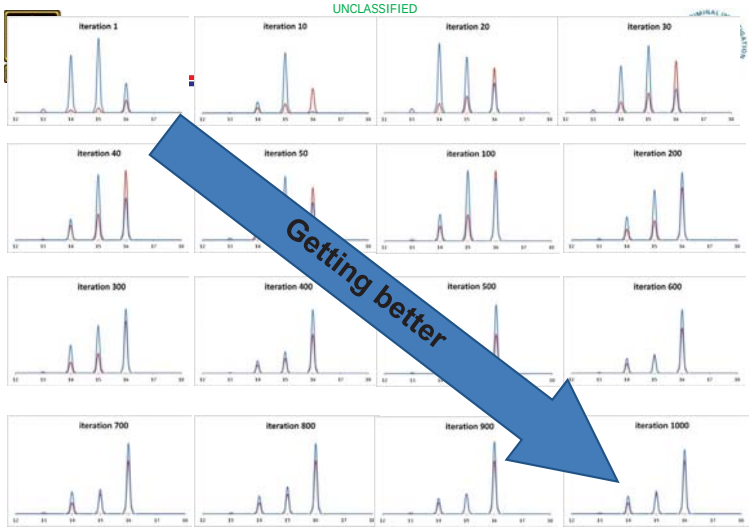
Quick view of How STRmix Works

Expected **Observed**

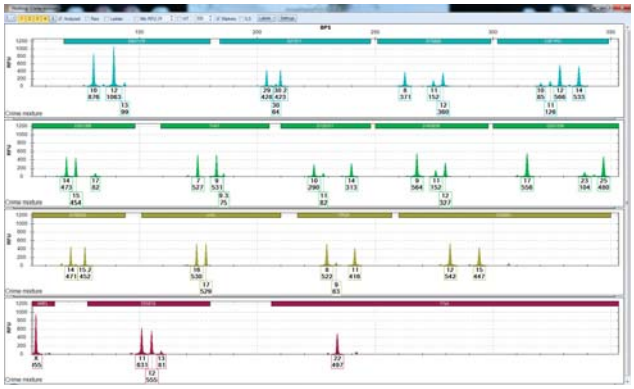
You can see from the example that there are a number of peak height differences between the expected and observed

- A likelihood will be calculated on this guess/iteration and compared to the previous guess to determine if it's a good guess or bad guess
- The genotypes in the better guess between the two get a point





The Mixture Egram For This Talk



Egram Interpretation

- We interpret by viewing the egram
  - Fix artifacts
  - Elevated stutter
  - Quality/number of contributors
- We then compare the references
- ArmedXpert gives nice visual summaries – but these aren't used to make inclusion/exclusions
- We only run STRmix if we make an inclusion
  - We do not run if exclusion or inconclusive
  - We are the experts, and use STRmix for weights

Egram Interpretation

- All alleles of V are found

	D851179	D21511	D75820	CSF1PO	D351358	TH01	D135317	D165539	D251338	D195433	vWA	TPOX	D18551	Amel	D55818	FGA		Ex.	Inc.	Ex.	Inc.
Victim [1 Reference]	10, 12	29, 30.2	8, 12	12, 14	14, 15	7, 9	10, 14	9, 12	17, 25	14, 15.2	16, 17	8, 11	12, 15	X	11, 12	22					
Crime mixture	10, 12, 13	29, 30, 30.2	8, 11, 12	10, 11, 12, 14	14, 15, 17	7, 9, 9.3	10, 11, 14	9, 11, 12	17, 23, 25	14, 15.2	16, 17	8, 9, 11	12, 15	X	11, 12, 13	22		16	5	11	
Suspect	12, 13	29, 30	11	10, 11	15, 17	6, 9.3	11	11	23	13, 14	17	8, 9	15, 18	X, Y	11, 13	24, 26		1	0	1	
Victim	10, 12	29, 30.2	8, 12	12, 14	14, 15	7, 9	10, 14	9, 12	17, 25	14, 15.2	16, 17	8, 11	12, 15	X	11, 12	22		16	16	0	
Exact match																					
Included																					
Find Where Reference Included																					

- Some alleles of S are missing

	D851179	D21511	D75820	CSF1PO	D351358	TH01	D135317	D165539	D251338	D195433	vWA	TPOX	D18551	Amel	D55818	FGA		Ex.	Inc.	Ex.	Inc.
Suspect [1 Reference]	12, 13	29, 30	11	10, 11	15, 17	6, 9.3	11	11	23	13, 14	17	8, 9	15, 18	X, Y	11, 13	24, 26					
Crime mixture	10, 12, 13	29, 30, 30.2	8, 11, 12	10, 11, 12, 14	14, 15, 17	7, 9, 9.3	10, 11, 14	9, 11, 12	17, 23, 25	14, 15.2	16, 17	8, 9, 11	12, 15	X	11, 12, 13	22		11	0	11	
Suspect	12, 13	29, 30	11	10, 11	15, 17	6, 9.3	11	11	23	13, 14	17	8, 9	15, 18	X, Y	11, 13	24, 26		16	16	0	
Victim	10, 12	29, 30.2	8, 12	12, 14	14, 15	7, 9	10, 14	9, 12	17, 25	14, 15.2	16, 17	8, 11	12, 15	X	11, 12	22		1	0	1	
Exact match						6				13			18		24,26						
Included																					
Find Where Reference Included																					

- However, it is my opinion as the expert that Suspect cannot be excluded/is included as minor

STRmix Results

- USACIL uses the NIST population database and reports the lowest LR for Caucasian, Hispanic and Black population groups
- STRmix can give four classes of answers:
  - LR > 1 means Hp is favored (more likely if Mr. POI)
  - LR < 1 means Hd is favored (more likely if Unknown)
  - LR = 1 means neutral (equally likely for Mr. POI and U)
  - LR = 0 means exclusion (Hp fails to fit the POI)
  - USACIL allows for inconclusive - "uninformative LR"
    - Between 1 and 10 is "uninformative"
    - Numbers larger than 10 can also be "uninformative" under certain conditions

STRmix Results

- Assuming our evidence is V swab (vaginal)
  - LR = V+S / V+U = 3.21×10^{7} or 32 million
- Assuming our evidence is S related (his boxers)
  - LR = S+V / S+U = 6.96×10^{18} or 6.9 quintillion
- Assuming our evidence is neutral (cloth from park)
  - LR = V+S / U+U = 1.25×10^{27} or 1.2 octillion
- Notice the trend; 10^{7} with 10^{18} gives ≈ 10^{27}
- (These runs took from 10 to 20 seconds)

- This evidence is X times more likely if it came from (Hp proposition) than if it came from (Hd proposition)

STRmix Results

- STRmix uses layers of conservativeness
- Consider V+S / V+U for Caucasian
- Product rule point source LR (Theta = zero)
  - 1.57×10^{8} (150 million)
- NRC 4.2 and point source
  - 8.72×10^{7} (87 million)
- NRC 4.2 and HPD (allele frequencies only)
  - 3.30×10^{7} (33 million)
- NRC 4.2 and full HPD (allele freqs and MCMC)
  - 3.21×10^{7} (32 million) → USACIL uses this LR

$$Het\ \frac{2(\theta+(1-\theta)p)(\theta+(1-\theta)q)}{(1+\theta)(1+2\theta)}$$

$$Hom\ \frac{(2\theta+(1-\theta)p)(3\theta+(1-\theta)p)}{(1+\theta)(1+2\theta)}$$

STRmix Output

- STRmix output has all kinds of useful information
  - Your propositions

```
Number of known contributors under Hp: 2
 Victim.csv, Suspect.csv
Number of known contributors under Hd: 1
 Victim.csv
```

  - Contributor order and proportions

```
---Order---                    Mixture Proportions (%)
Victim.csv - Contributor 1     Contributor 1 - 90.0
Suspect.csv - Contributor 2    Contributor 2 - 10.0
```

STRmix Output

- STRmix output has all kinds of useful information
  - Individual degradation curves

```
Degradation (rfu/bp), starting at 119.0bp
Contributor 1 - 0.7491114559272556
Contributor 2 - 0.08000170151952672
```

  - How long did the run take and how many rounds of guessing to hit the 500K "good" guesses

```
Pre-Burnin time taken: 0.016 secs
Burnin time taken: 5.116 secs
post Burnin time taken: 2.652 secs
Total iterations
849964.0
```


STRmix Output

- The two most important parts of the output
  - Locus LRs and total LR

```
Locus 12(TPOX): Pr(E|Hp) = 0.38305, Pr(E|Hd) = 0.07084, LR = 5.40708
Locus 13(D18S51): Pr(E|Hp) = 0.08953, Pr(E|Hd) = 0.09934, LR = 0.90124
Locus 14(D5S818): Pr(E|Hp) = 0.38792, Pr(E|Hd) = 0.08013, LR = 4.84101
Locus 15(FGA): Pr(E|Hp) = 0.41593, Pr(E|Hd) = 0.38265, LR = 1.08697
LR total = 1.5701092782308355E8
99.0% 1-sided lower HPD interval = 5.547800187855319E7
```

  - Genotype combinations and weights

```
D21S11
[29,30.2]   [29,30]     0.3756747344593418
[29,30.2]   [30,30.2]   0.3656048962073688
[29,30.2]   [30,30]     0.20895943828209196
[29,30.2]   [-1,30]     0.049760931051197464
D7S820
```


The Overall Results (V+S / V+U) (product rule)

```
Locus 1(D8S1179): Pr(E|Hp) = 0.76275,  Pr(E|Hd) = 0.10205,  LR = 7.47435
Locus 2(D21S11):  Pr(E|Hp) = 0.37567,  Pr(E|Hd) = 0.07917,  LR = 4.74526
Locus 3(D7S820):  Pr(E|Hp) = 0.7439,   Pr(E|Hd) = 0.0508,   LR = 14.64244
Locus 4(CSF1PO):  Pr(E|Hp) = 0.96527,  Pr(E|Hd) = 0.13382,  LR = 7.21303
Locus 5(D3S1358): Pr(E|Hp) = 0.28923,  Pr(E|Hd) = 0.07293,  LR = 3.96565
Locus 6(TH01):    Pr(E|Hp) = 0.05432,  Pr(E|Hd) = 0.11496,  LR = 0.4725
Locus 7(D13S317): Pr(E|Hp) = 0.54059,  Pr(E|Hd) = 0.10586,  LR = 5.10669
Locus 8(D16S539): Pr(E|Hp) = 0.59187,  Pr(E|Hd) = 0.09388,  LR = 6.30467
Locus 9(D2S1338): Pr(E|Hp) = 0.40147,  Pr(E|Hd) = 0.02584,  LR = 15.5388
Locus 10(D19S433):Pr(E|Hp) = 0.15827,  Pr(E|Hd) = 0.14448,  LR = 1.09544
Locus 11(vWA):    Pr(E|Hp) = 0.23571,  Pr(E|Hd) = 0.14808,  LR = 1.59179
Locus 12(TPOX):   Pr(E|Hp) = 0.38305,  Pr(E|Hd) = 0.07084,  LR = 5.40708
Locus 13(D18S51): Pr(E|Hp) = 0.08953,  Pr(E|Hd) = 0.09934,  LR = 0.90124
Locus 14(D5S818): Pr(E|Hp) = 0.38792,  Pr(E|Hd) = 0.08013,  LR = 4.84101
Locus 15(FGA):    Pr(E|Hp) = 0.41593,  Pr(E|Hd) = 0.38265,  LR = 1.08697
LR total = 1.5701092782308355E8
```

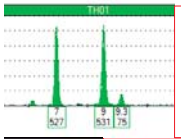
This is the product rule output so I can show you how to calculate the LR manually without having to use 4.2 – I'd make a mistake!

LR > 1 favors Hp at 10 loci
LR < 1 favors Hd at 1 locus
LR ≈ 1 is neutral at ~4 loci (vWA?)
~~LR =~~ 0 -nothing exclusionary

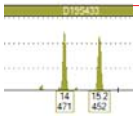
The Overall Results (V+S / V+U) (product rule)

```
Locus 6(TH01):    LR = 0.4725
Locus 10(D19S433):LR = 1.09544
Locus 11(vWA):    LR = 1.59179
Locus 13(D18S51): LR = 0.90124
Locus 15(FGA):    LR = 1.08697
```

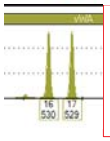
I have to look at these LRs from the Hp perspective – Why are they not >1? If I can't agree, *maybe* there's a problem with the inclusion?



THO1 is missing a 6, the 6 is in stutter position and the minor is well represented in green, so Hp is penalized



D19 is missing a 13, but not as much info in that channel (one minor allele), so it's neutral



vWA has minor masked by major, so it's mostly neutral – maybe there, maybe not

D18 and FGA both have drop-out, but since they are large loci, it's neutral – you expect drop-out with this level of minor

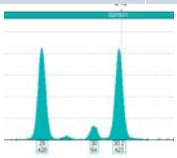
Building the LR for V+S / V+U

- Make a list genotype combinations that matter
  - Hint: This is what STRmix does for us

Genotypes				
[29,30.2] [29,30]				
[29,30.2] [30,30.2]				
[29,30.2] [30,30]				
[29,30.2] [30,Q]				

V = 29,30.2

Allele frequencies
29 = 0.2020
30 = 0.2822
30.2 = 0.0291
Q = 0.4866



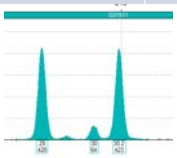
Building the LR for V+S / V+U

- Calculate the genotype frequencies
  - Easy with product rule, you've been doing it for years

Genotypes	P1 Genotype Frequency	P2 Genotype Frequency		
[29,30.2] [29,30]	1	2x0.2020x0.2822		
[29,30.2] [30,30.2]	1	2x0.2822x0.0291		
[29,30.2] [30,30]	1	0.2822^2		
[29,30.2] [30,Q]	1	2x0.2822x0.4866		

V = 29,30.2

Allele frequencies
29 = 0.2020
30 = 0.2822
30.2 = 0.0291
Q = 0.4866



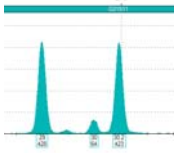
Building the LR for V+S / V+U

- Determine the weights
  - Not so easy – thanks STRmix – but should make sense

Genotypes	P1 Genotype Frequency	P2 Genotype Frequency	Weight	
[29,30.2] [29,30]	1	0.114009	0.37567	
[29,30.2] [30,30.2]	1	0.016424	0.36560	
[29,30.2] [30,30]	1	0.079637	0.20895	
[29,30.2] [30,Q]	1	0.274637	0.04976	

V = 29,30.2

Allele frequencies
29 = 0.2020
30 = 0.2822
30.2 = 0.0291
Q = 0.4866



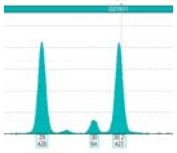
Building the LR for V+S / V+U

- Get the weighted genotype values
  - Modify the genotype frequencies - multiply across rows

Genotypes	P1 Genotype Frequency	P2 Genotype Frequency	Weight	Weighted Values
[29,30.2] [29,30]	1	0.114009	0.37567	0.042830
[29,30.2] [30,30.2]	1	0.016424	0.36560	0.006005
[29,30.2] [30,30]	1	0.079637	0.20895	0.016641
[29,30.2] [30,Q]	1	0.274637	0.04976	0.013666

V = 29,30.2

Allele frequencies
29 = 0.2020
30 = 0.2822
30.2 = 0.0291
Q = 0.4866



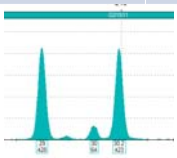
Building the LR for V+S / V+U

- Add down
  - You just solved for H2 – the denominator of the LR

Genotypes	P1 Genotype Frequency	P2 Genotype Frequency	Weight	Weighted Frequency
[29,30.2] [29,30]	1	0.114009	0.37567	0.042830
[29,30.2] [30,30.2]	1	0.016424	0.36560	0.006005
[29,30.2] [30,30]	1	0.079637	0.20895	0.016641
[29,30.2] [30,Q]	1	0.274637	0.04976	0.013666
				0.07914

V = 29,30.2

Allele frequencies
29 = 0.2020
30 = 0.2822
30.2 = 0.0291
Q = 0.4866



This is our denominator.

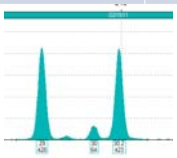
Building the LR for V+S / V+U

- Solve for H1
  - Only one type matters for H1 - S = 29,30

Genotypes	P1 Genotype Frequency	P2 Genotype Frequency	Weight	Weighted Frequency
[29,30.2] [29,30]	1	0.114009	0.37567	0.042830
~~[29,30.2] [30,30.2]~~	~~1~~	~~0.016424~~	~~0.36560~~	~~0.006005~~
~~[29,30.2] [30,30]~~	~~1~~	~~0.079637~~	~~0.20895~~	~~0.016641~~
~~[29,30.2] [30,Q]~~	~~1~~	~~0.274637~~	~~0.04976~~	~~0.013666~~
				~~**0.079142**~~

V = 29,30.2

Allele frequencies
29 = 0.2020
30 = 0.2822
30.2 = 0.0291
Q = 0.4866



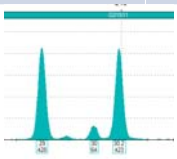
Building the LR for V+S / V+U

- Solve for H1
  - And Prosecution knows both genotypes – so it's just the weight

Genotypes	P1 Genotype Frequency	P2 Genotype Frequency	Weight	Weighted Frequency
[29,30.2] [29,30]	1	1	0.37567	0.375675
~~[29,30.2] [30,30.2]~~	~~1~~	~~0.016424~~	~~0.36560~~	~~0.006005~~
~~[29,30.2] [30,30]~~	~~1~~	~~0.079637~~	~~0.20895~~	~~0.016641~~
~~[29,30.2] [30,Q]~~	~~1~~	~~0.274637~~	~~0.04976~~	~~0.013666~~
				0.375675

V = 29,30.2

Allele frequencies
29 = 0.2020
30 = 0.2822
30.2 = 0.0291
Q = 0.4866



This is our numerator. Notice anything?

Building the LR

- From the D21 LR portion of STRmix output:

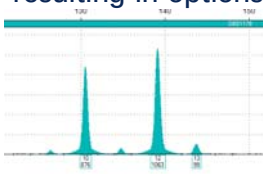
```
Locus 2(D21S11): Pr(E|Hp) = 0.37567, Pr(E|Hd) = 0.07917, LR = 4.74526
```

- We plug-in our Hp = 0.37567
- And our Hd that we did earlier = 0.07914
- Then we can divide those and solve for our LR = 4.746

- Caveats if you try to match this on your own:
  - I rounded numbers to fit cells in the table and the table on the slide – using all decimals gives exactly the same Hd and LR
  - STRmix modifies allele frequency values from the database by (x + 1/k)/N+1

There's lots more…

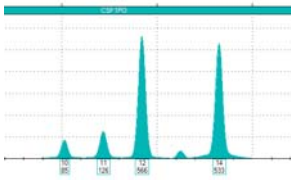
- STRmix considers peaks as stutters and alleles, resulting in options we maybe wouldn't think about



D8S1179

[10,12]	[12,13]	0.7627452456809936
[10,12]	[10,13]	0.1897821554795262
[10,12]	[13,13]	0.043358306940058634
[10,12]	[11,13]	0.0028836515428888754
[10,12]	[-1,13]	0.0011588726110752926
[10,12]	[9,13]	7.176774545745467E-5

Considers 9 and 11 stutters as alleles



CSF1PO

[12,14]	[10,11]	0.9652667642394266
[12,14]	[10,12]	0.014308841315632191
[12,14]	[10,14]	0.012452292097077053
[12,14]	[10,10]	0.006483804019934962
[12,14]	[10,13]	8.835668334188271E-4
[12,14]	[-1,10]	6.047314945103557E-4

Considers the 11 as stutter (and 13 stutter as allele)

There's lots more…

- But I know I'm out of time…..

- FYI –
  - USACIL has been reporting with STRmix LRs since November 2014
  - Army, Navy, Air Force and Marine courts martial have all heard testimony with STRmix LRs

Acknowledgement

DFSC/USACIL Team:

Joel Sutton (Technical Leader)
Lindsey Smith
Curt Schuerman

ESR/FSSA has provided lots of training and slides:

John Buckleton
Jo Bright
Duncan Taylor

Tim Kalafut
Forensic Biologist, USACIL

timothy.s.kalafut.civ@mail.mill
404-469-4469