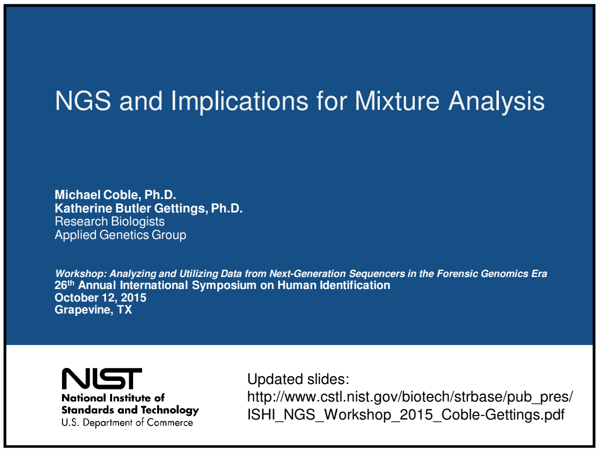


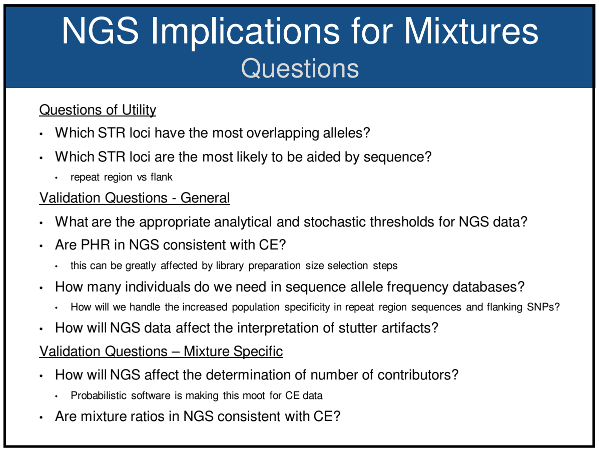
NGS and Implications for Mixture Analysis

Michael Coble, Ph.D.
Katherine Butler Gettings, Ph.D.
Research Biologists
Applied Genetics Group

Workshop: Analyzing and Utilizing Data from Next-Generation Sequencers in the Forensic Genomics Era
26th Annual International Symposium on Human Identification
October 12, 2015
Grapevine, TX

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

Updated slides:
http://www.cstl.nist.gov/biotech/strbase/pub_pres/ISHI_NGS_Workshop_2015_Coble-Gettings.pdf



NGS Implications for Mixtures
Questions

Questions of Utility

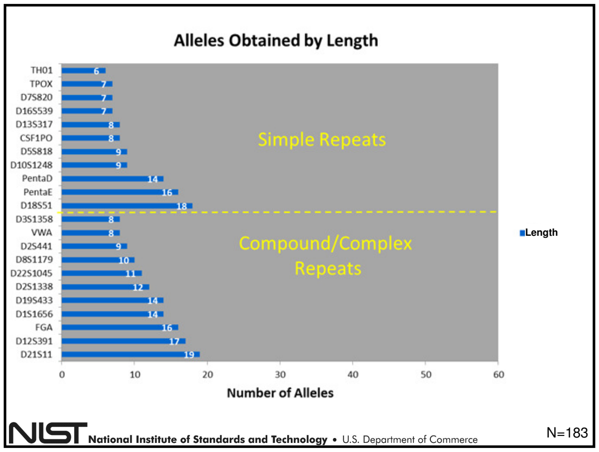
- Which STR loci have the most overlapping alleles?
- Which STR loci are the most likely to be aided by sequence?
  - repeat region vs flank

Validation Questions - General

- What are the appropriate analytical and stochastic thresholds for NGS data?
- Are PHR in NGS consistent with CE?
  - this can be greatly affected by library preparation size selection steps
- How many individuals do we need in sequence allele frequency databases?
  - How will we handle the increased population specificity in repeat region sequences and flanking SNPs?
- How will NGS data affect the interpretation of stutter artifacts?

Validation Questions – Mixture Specific

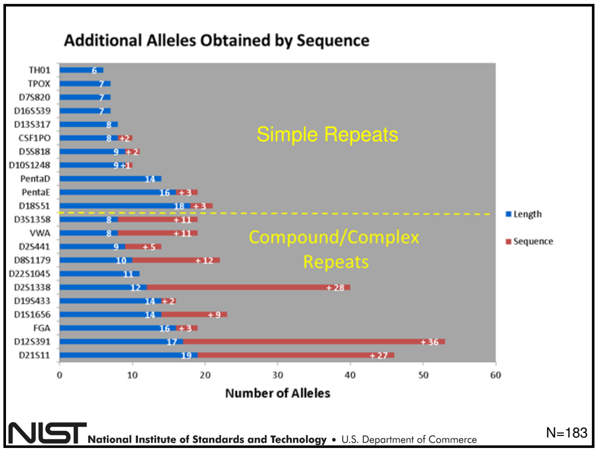
- How will NGS affect the determination of number of contributors?
  - Probabilistic software is making this moot for CE data
- Are mixture ratios in NGS consistent with CE?



Alleles Obtained by Length

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

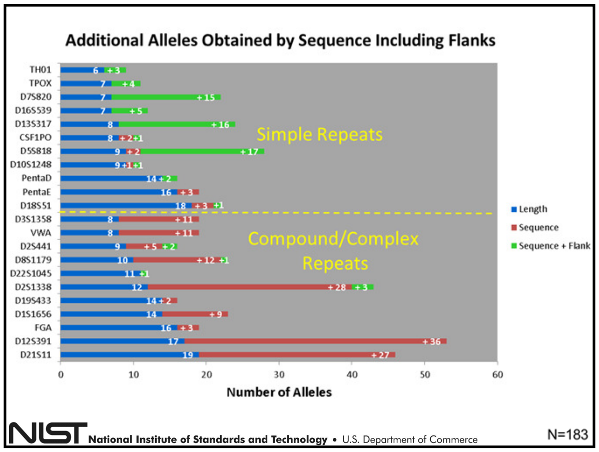
N=183



Additional Alleles Obtained by Sequence

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

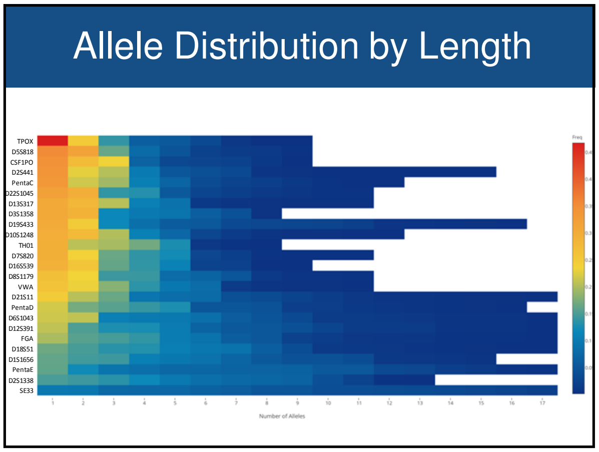
N=183



Additional Alleles Obtained by Sequence Including Flanks

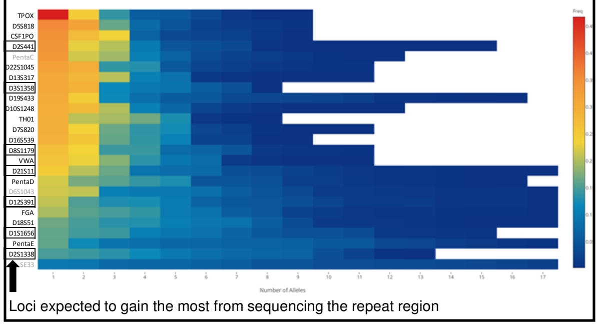
National Institute of Standards and Technology • U.S. Department of Commerce

N=183



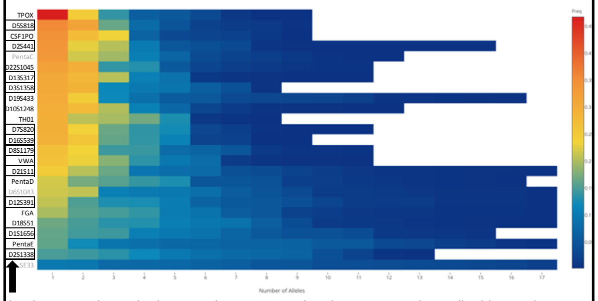
Allele Distribution by Length

Allele Distribution by Length



Loci expected to gain the most from sequencing the repeat region

Allele Distribution by Length



Loci expected to gain the most from sequencing the repeat region & flanking region

NGS of STR Mixtures – Proof of Concept

Created 10 difficult 2-person mixtures by CE

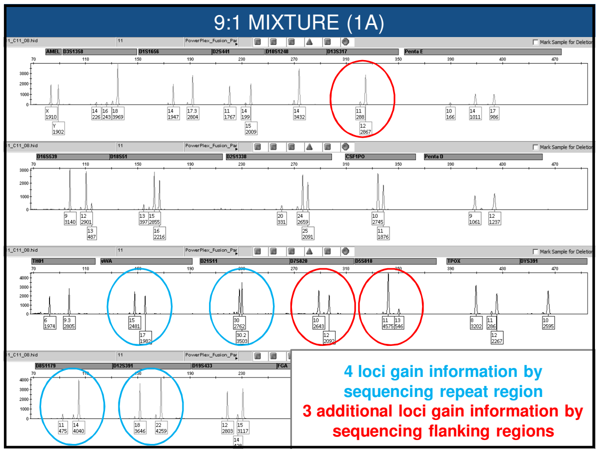
- 9:1 ratio
- Maximal overlapping alleles
- Very few loci have four distinguishable alleles by CE

Inferred sequences from NGS data

"Best case" help from NGS in 2-person mixtures

Does not include help with stutter

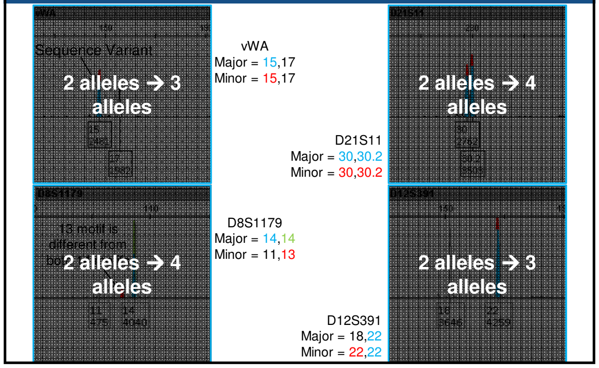
9:1 MIXTURE (1A)



4 loci gain information by sequencing repeat region

3 additional loci gain information by sequencing flanking regions

9:1 MIXTURE
Repeat Region Gains by Sequencing



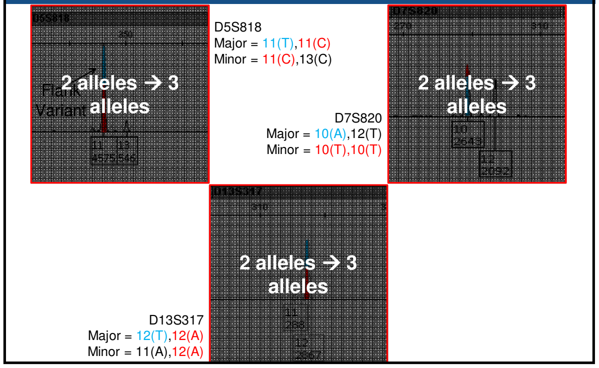
vWA
Major = 15,17
Minor = 15,17

D21S11
Major = 30,30.2
Minor = 30,30.2

D8S1179
Major = 14,14
Minor = 11,13

D12S391
Major = 18,22
Minor = 22,22

9:1 MIXTURE
Flanking Region Gains by Sequencing



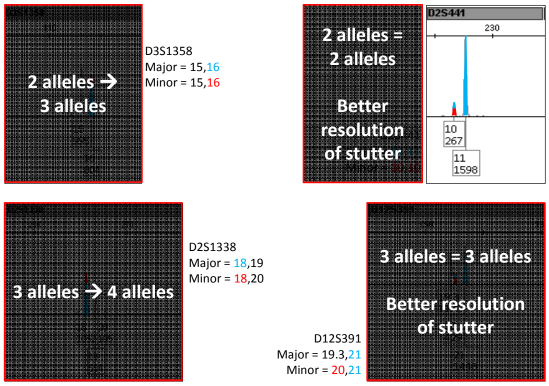
D5S818
Major = 11(T),11(C)
Minor = 11(C),13(C)

D7S820
Major = 10(A),12(T)
Minor = 10(T),10(T)

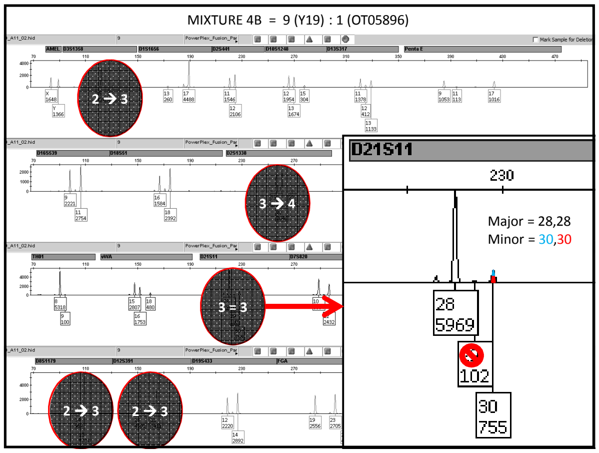
D13S317
Major = 12(T),12(A)
Minor = 11(A),12(A)

http://www.cstl.nist.gov/biotech/strbase/pub_pres/ISHI_NGS_Workshop_2015_Coble-Gettings.pdf

MIXTURE 2B = 9 (UT57318) : 1 (JT52076)



MIXTURE 4B = 9 (Y19) : 1 (OT05896)



NGS Applications to STR Mixtures

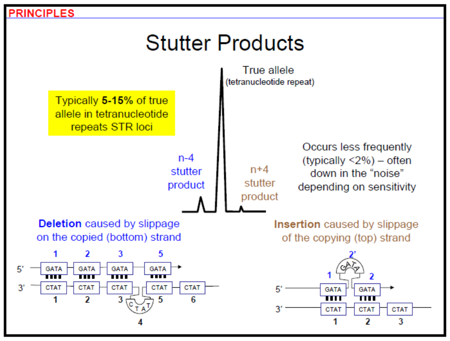
summary of 9:1 mixtures

	# loci gain- repeat region	# loci gain- flanking region	% total gain
1	4	3	32%
2	4	3	32%
3	3	1	18%
4	4	1	23%
5	7	3	45%
6	7	3	45%
7	5	1	27%
8	5	1	27%
9	5	1	27%
10	5	1	27%

On average, 30% of loci gained information by sequence

(*remember this is best case!*)

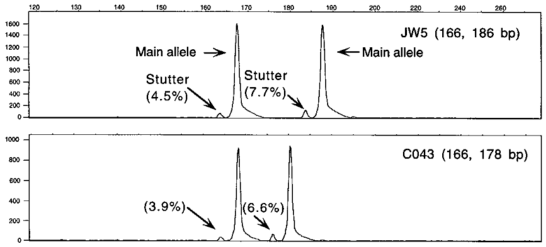
Stutter by Sequence



PRINCIPLES
Stutter Products
True allele
(tetranucleotide repeat)
Typically **5-15%** of true allele in tetranucleotide repeats STR loci
n-4 stutter product
n+4 stutter product
Occurs less frequently (typically <2%) – often down in the “noise” depending on sensitivity
Deletion caused by slippage on the copied (bottom) strand
Insertion caused by slippage of the copying (top) strand
5'
3'
GATA
CTAT
1
2
3
4
5
6
2'

Image courtesy of Mike Coble (http://www.cstl.nist.gov/strbase/mixture/4%20-%20Stutter.pdf

Stutter by Sequence

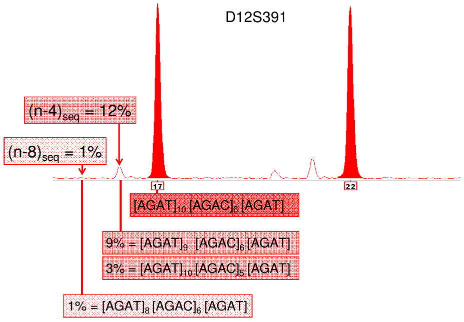


Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA

P. Sean Walsh*,+, Nicola J. Fildes+ and Rebecca Reynolds

Nucleic Acids Research, 1996, Vol. 24, No. 14

Stutter by Sequence



$(n\text{-}4)_{seq}$ = 12%

$(n\text{-}8)_{seq}$ = 1%

$[AGAT]_{10}$ $[AGAC]_6$ [AGAT]

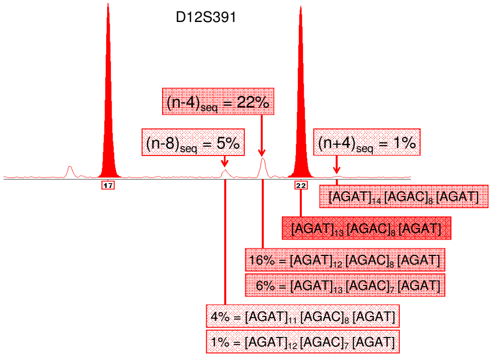
9% = $[AGAT]_9$ $[AGAC]_6$ [AGAT]

3% = $[AGAT]_{10}$ $[AGAC]_5$ [AGAT]

1% = $[AGAT]_8$ $[AGAC]_6$ [AGAT]

http://www.cstl.nist.gov/biotech/strbase/pub_pres/ISHI_NGS_Workshop_2015_Coble-Gettings.pdf

Stutter by Sequence



D12S391

(n-4)$_{seq}$ = 22%

(n-8)$_{seq}$ = 5%

(n+4)$_{seq}$ = 1%

[AGAT]$_{14}$ [AGAC]$_{8}$ [AGAT]

[AGAT]$_{13}$ [AGAC]$_{8}$ [AGAT]

16% = [AGAT]$_{12}$ [AGAC]$_{8}$ [AGAT]

6% = [AGAT]$_{13}$ [AGAC]$_{7}$ [AGAT]

4% = [AGAT]$_{11}$ [AGAC]$_{8}$ [AGAT]

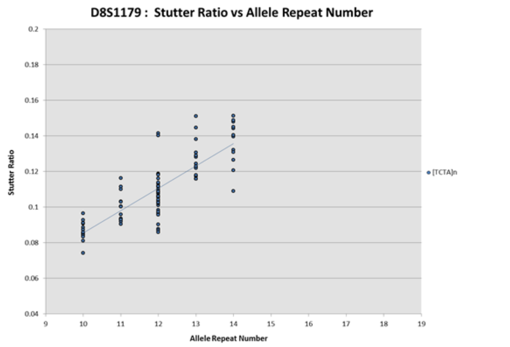
1% = [AGAT]$_{12}$ [AGAC]$_{7}$ [AGAT]

Stutter by Sequence

D8S1179

Allele	Repeat Structure
	[TCTA]10–14
10	[TCTA]10
11	[TCTA]11
12	[TCTA]12
13	[TCTA]13
14	[TCTA]14

Stutter by Sequence

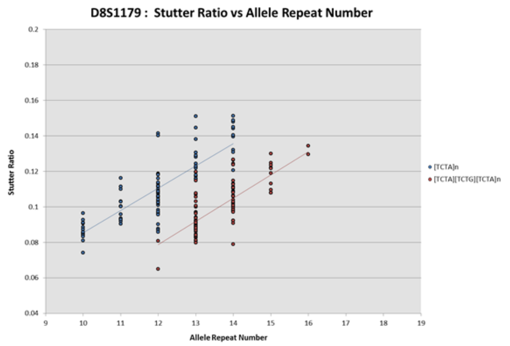


Stutter by Sequence

D8S1179

Allele	Repeat Structure
	[TCTA]10–14
10	[TCTA]10
11	[TCTA]11
12	[TCTA]12
13	[TCTA]13
14	[TCTA]14
	[TCTA][TCTG][TCTA]10–14
12	[TCTA][TCTG][TCTA]10
13	[TCTA][TCTG][TCTA]11
14	[TCTA][TCTG][TCTA]12
16	[TCTA][TCTG][TCTA]14

Stutter by Sequence

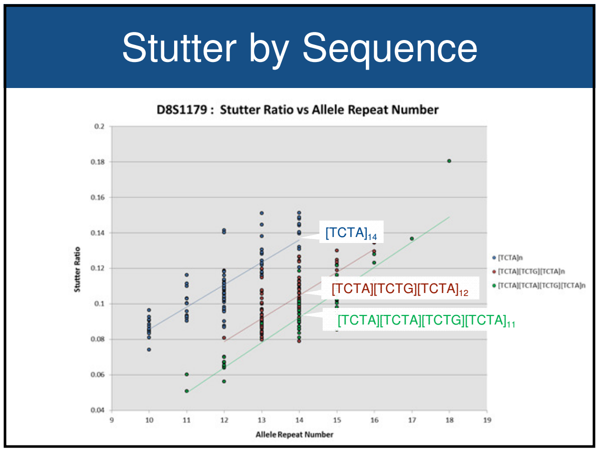


Stutter by Sequence

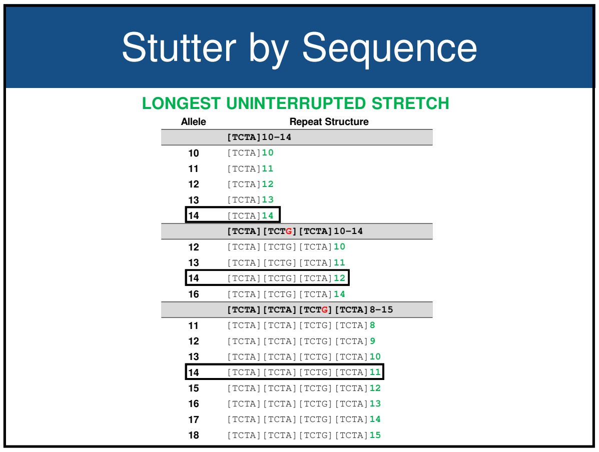
D8S1179

Allele	Repeat Structure
	[TCTA]10–14
10	[TCTA]10
11	[TCTA]11
12	[TCTA]12
13	[TCTA]13
14	[TCTA]14
	[TCTA][TCTG][TCTA]10–14
12	[TCTA][TCTG][TCTA]10
13	[TCTA][TCTG][TCTA]11
14	[TCTA][TCTG][TCTA]12
16	[TCTA][TCTG][TCTA]14
	[TCTA][TCTA][TCTG][TCTA]8–15
11	[TCTA][TCTA][TCTG][TCTA]8
12	[TCTA][TCTA][TCTG][TCTA]9
13	[TCTA][TCTA][TCTG][TCTA]10
14	[TCTA][TCTA][TCTG][TCTA]11
15	[TCTA][TCTA][TCTG][TCTA]12
16	[TCTA][TCTA][TCTG][TCTA]13
17	[TCTA][TCTA][TCTG][TCTA]14
18	[TCTA][TCTA][TCTG][TCTA]15

http://www.cstl.nist.gov/biotech/strbase/pub_pres/ISHI_NGS_Workshop_2015_Coble-Gettings.pdf



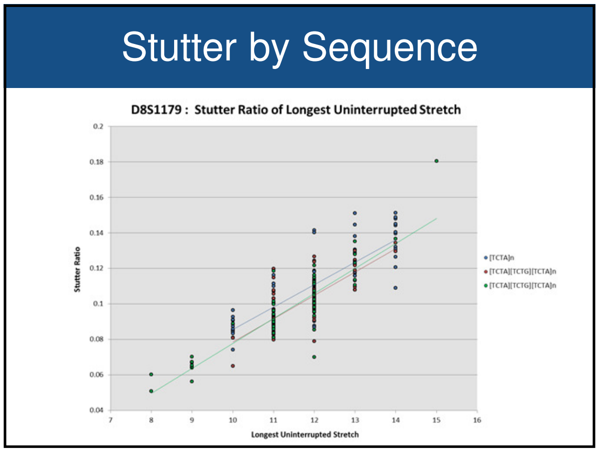
Stutter by Sequence



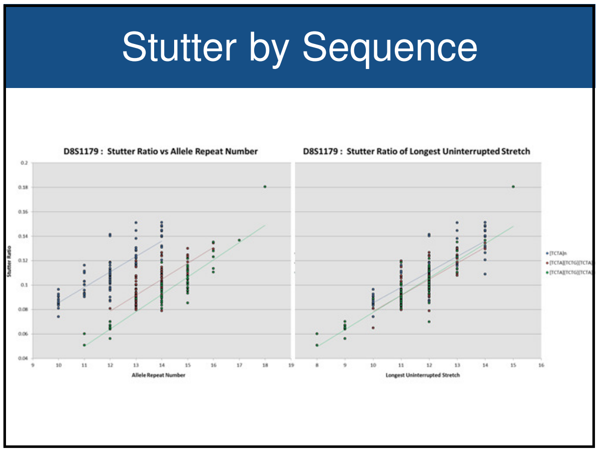
Stutter by Sequence

LONGEST UNINTERRUPTED STRETCH

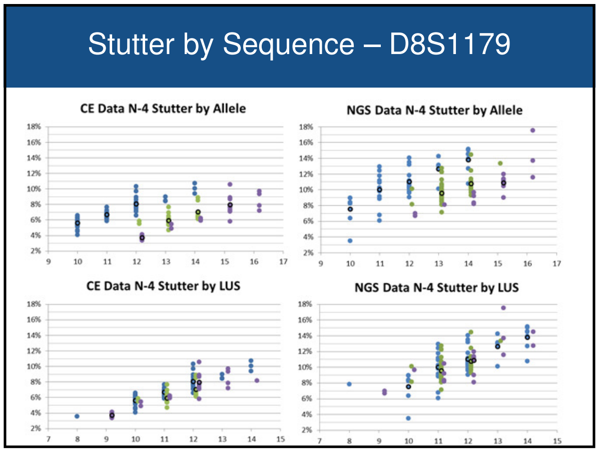
Allele	Repeat Structure
	[TCTA]10-14
10	[TCTA]10
11	[TCTA]11
12	[TCTA]12
13	[TCTA]13
14	[TCTA]14
	[TCTA] [TCTG] [TCTA]10-14
12	[TCTA] [TCTG] [TCTA]10
13	[TCTA] [TCTG] [TCTA]11
14	[TCTA] [TCTG] [TCTA]12
16	[TCTA] [TCTG] [TCTA]14
	[TCTA] [TCTA] [TCTG] [TCTA]8-15
11	[TCTA] [TCTA] [TCTG] [TCTA]8
12	[TCTA] [TCTA] [TCTG] [TCTA]9
13	[TCTA] [TCTA] [TCTG] [TCTA]10
14	[TCTA] [TCTA] [TCTG] [TCTA]11
15	[TCTA] [TCTA] [TCTG] [TCTA]12
16	[TCTA] [TCTA] [TCTG] [TCTA]13
17	[TCTA] [TCTA] [TCTG] [TCTA]14
18	[TCTA] [TCTA] [TCTG] [TCTA]15



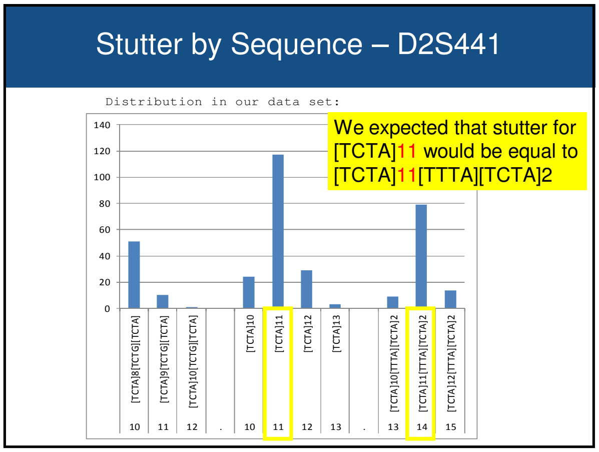
Stutter by Sequence



Stutter by Sequence



Stutter by Sequence – D8S1179

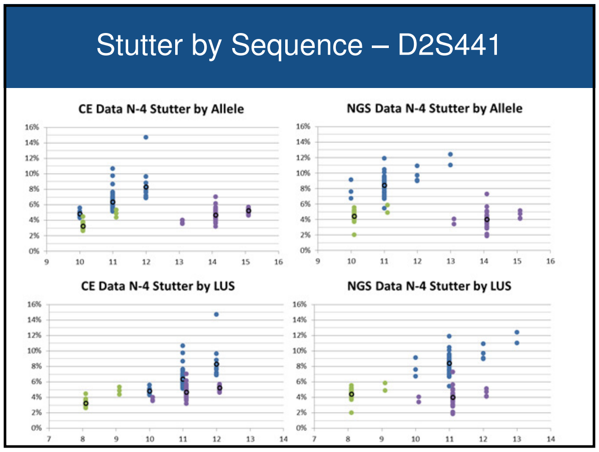


Stutter by Sequence – D2S441

Distribution in our data set:

We expected that stutter for [TCTA]11 would be equal to [TCTA]11[TTTA][TCTA]2

http://www.cstl.nist.gov/biotech/strbase/pub_pres/ISHI_NGS_Workshop_2015_Coble-Gettings.pdf



Stutter by Sequence – D2S441

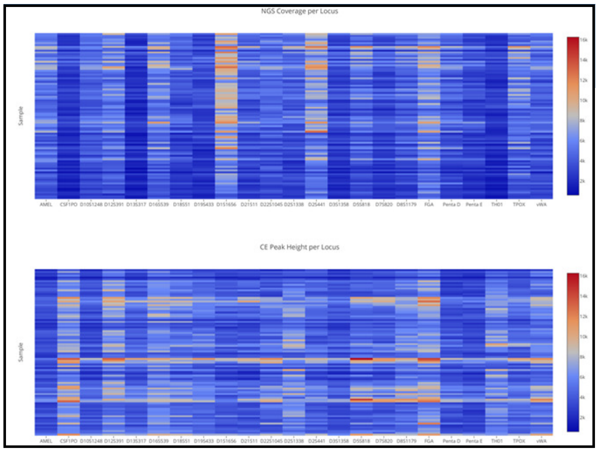
NGS Mixture Study

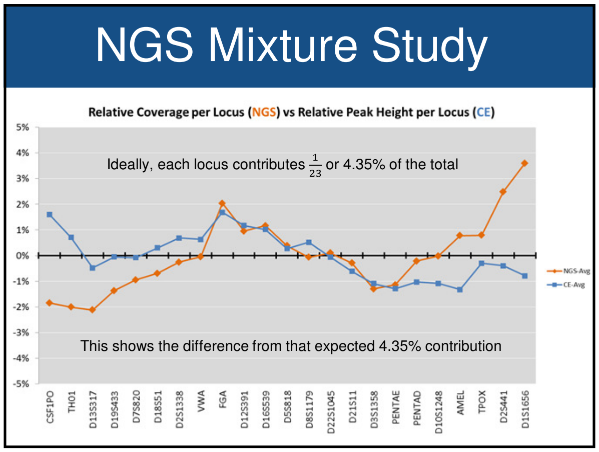
Are mixture ratios by NGS the same as mixture ratios by CE?

	CE		NGS
Loci	PowerPlex Fusion + PowerPlex Y23	🍎 🍎	PowerSeq Auto + Y
Input DNA	0.5 ng each	🍎 🍎	0.5 ng total
Amp Parameters	30 cycles	🍎 🍎	30 cycles, same as PPF
Everything Else	3500xL	👤 👤	TruSeq PCR free Library Prep, MiSeq v3

NGS Mixture Study

	1	2	3	4	5	6	7	8	9	10	11	12
A	A	B	M	9:1 A/B	9:1 A/B	9:1 A/B	19:1 A/M	19:1 A/M	19:1 A/M	8:1:1 A/B/M	8:1:1 A/B/M	8:1:1 A/B/M
B	1:1 A/B	1:1 A/M	1:1 B/M	9:1 B/A	9:1 B/A	9:1 B/A	19:1 M/A	19:1 M/A	19:1 M/A	1:8:1 A/B/M	1:8:1 A/B/M	1:8:1 A/B/M
C	3:1 A/B	3:1 A/B	3:1 A/B	9:1 A/M	9:1 A/M	9:1 A/M	19:1 B/M	19:1 B/M	19:1 B/M	1:1:8 A/B/M	1:1:8 A/B/M	1:1:8 A/B/M
D	3:1 B/A	3:1 B/A	3:1 B/A	9:1 M/A	9:1 M/A	9:1 M/A	19:1 M/B	19:1 M/B	19:1 M/B	18:1:1 A/B/M	18:1:1 A/B/M	18:1:1 A/B/M
E	3:1 A/M	3:1 A/M	3:1 A/M	9:1 B/M	9:1 B/M	9:1 B/M	1:1:1 A/B/M	1:1:1 A/B/M	1:1:1 A/B/M	1:18:1 A/B/M	1:18:1 A/B/M	1:18:1 A/B/M
F	3:1 M/A	3:1 M/A	3:1 M/A	9:1 M/B	9:1 M/B	9:1 M/B	3:1:1 A/B/M	3:1:1 A/B/M	3:1:1 A/B/M	1:1:18 A/B/M	1:1:18 A/B/M	1:1:18 A/B/M
G	3:1 B/M	3:1 B/M	3:1 B/M	19:1 A/B	19:1 A/B	19:1 A/B	1:3:1 A/B/M	1:3:1 A/B/M	1:3:1 A/B/M	9:9:1 A/B/M	9:9:1 A/B/M	9:9:1 A/B/M
H	3:1 M/B	3:1 M/B	3:1 M/B	19:1 B/A	19:1 B/A	19:1 B/A	1:1:3 A/B/M	1:1:3 A/B/M	1:1:3 A/B/M	1:9:9 A/B/M	1:9:9 A/B/M	1:9:9 A/B/M



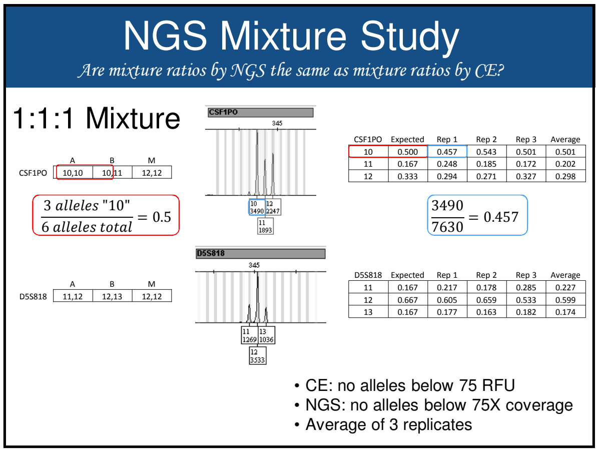


NGS Mixture Study

Relative Coverage per Locus (NGS) vs Relative Peak Height per Locus (CE)

Ideally, each locus contributes $\frac{1}{23}$ or 4.35% of the total

This shows the difference from that expected 4.35% contribution



NGS Mixture Study

Are mixture ratios by NGS the same as mixture ratios by CE?

1:1:1 Mixture

	A	B	M
CSF1PO	10,10	10,11	12,12

$$\frac{3\ alleles\ "10"}{6\ alleles\ total} = 0.5$$

CSF1PO	Expected	Rep 1	Rep 2	Rep 3	Average
10	0.500	0.457	0.543	0.501	0.501
11	0.167	0.248	0.185	0.172	0.202
12	0.333	0.294	0.271	0.327	0.298

$$\frac{3490}{7630} = 0.457$$

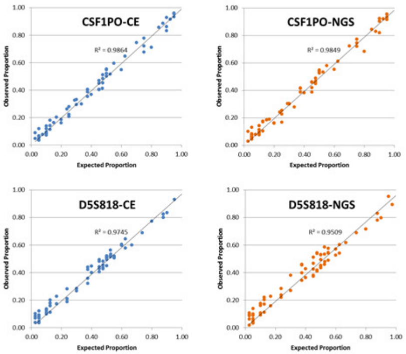
	A	B	M
D5S818	11,12	12,13	12,12

D5S818	Expected	Rep 1	Rep 2	Rep 3	Average
11	0.167	0.217	0.178	0.285	0.227
12	0.667	0.605	0.659	0.533	0.599
13	0.167	0.177	0.163	0.182	0.174

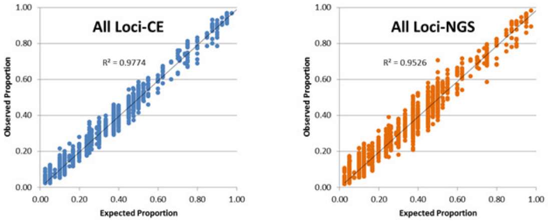
- CE: no alleles below 75 RFU
- NGS: no alleles below 75X coverage
- Average of 3 replicates

http://www.cstl.nist.gov/biotech/strbase/pub_pres/ISHI_NGS_Workshop_2015_Coble-Gettings.pdf

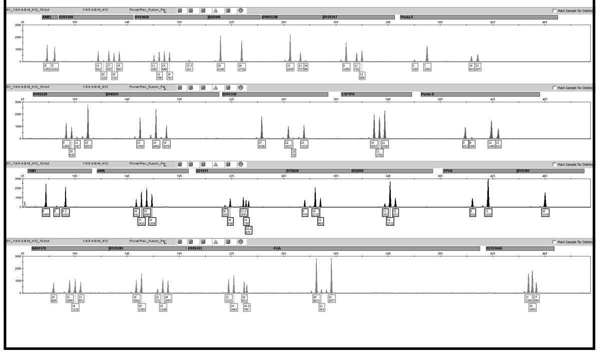
Expected vs Observed Mixture Contributions



Expected vs Observed Mixture Contributions

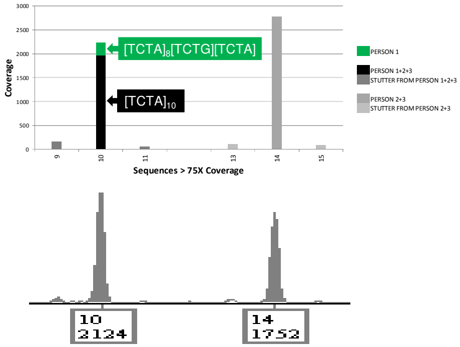


NGS Mixture Study 1:9:9



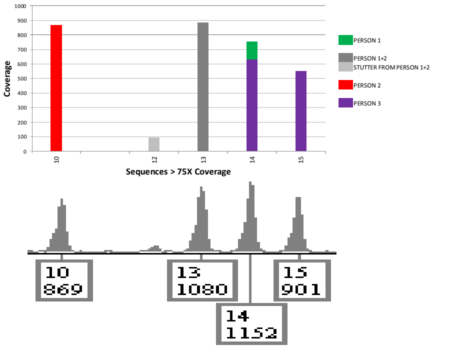
D2S441 1:9:9

1 additional allele



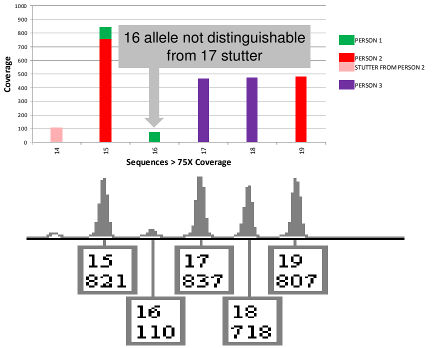
D8S1179 1:9:9

1 additional allele



D3S1358 1:9:9

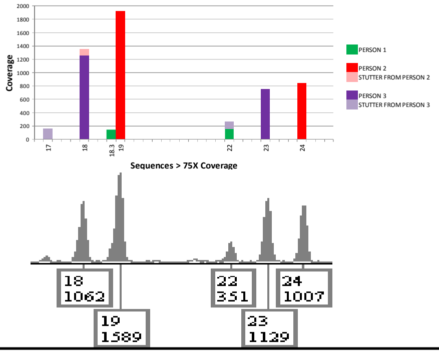
1 additional allele



http://www.cstl.nist.gov/biotech/strbase/pub_pres/ISHI_NGS_Workshop_2015_Coble-Gettings.pdf

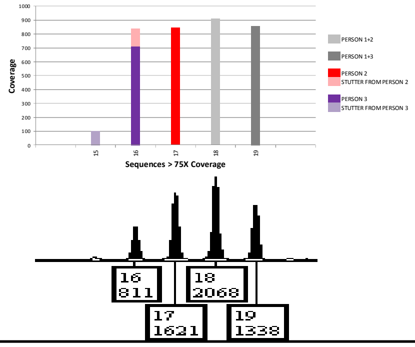
D12S391 1:9:9

1 additional allele
2 alleles
help with stutter



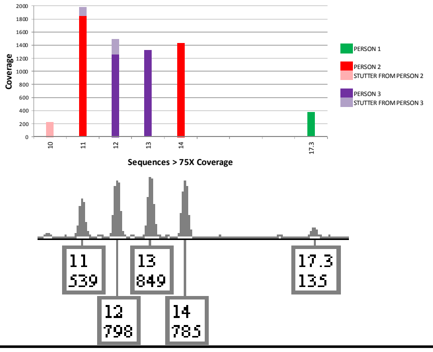
vWA 1:9:9

1 allele
help with stutter



D1S1656 1:9:9

2 alleles
help with stutter



Summary 1:9:9

Locus	Additional Alleles	Help with Stutter
D2S441	1	
D8S1179	1	
D3S1358	1	
D12S391	1	1
vWA		1
D1S1656		2

NGS profile contains four additional alleles and improved stutter attribution for four alleles

NGS Implications for Mixtures

Conclusions

- Sequencing forensic STR loci can uncover underlying sequence variation in the repeat and flanking regions
- This will increase allelic diversity, thus increasing the ability to discriminate among individuals in a mixture
- Additionally, sequence specific stutter ratios may improve mixture models

NGS Implications for Mixtures

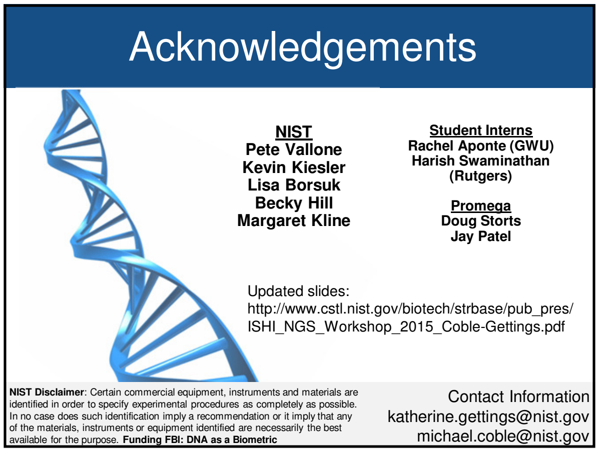
Conclusions

The gain is difficult to quantify

Prior to implementation:

- Sequence-based allele frequency databases
- Characterization of peak height ratios and stutter by NGS (assay and locus specific)
- Probabilistic genotyping software amenable to sequence data (and sequence-based stutter!)

http://www.cstl.nist.gov/biotech/strbase/pub_pres/ISHI_NGS_Workshop_2015_Coble-Gettings.pdf



Acknowledgements

NIST
Pete Vallone
Kevin Kiesler
Lisa Borsuk
Becky Hill
Margaret Kline

Student Interns
Rachel Aponte (GWU)
Harish Swaminathan (Rutgers)

Promega
Doug Storts
Jay Patel

Updated slides:
http://www.cstl.nist.gov/biotech/strbase/pub_pres/ISHI_NGS_Workshop_2015_Coble-Gettings.pdf

NIST Disclaimer: Certain commercial 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 it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose. **Funding FBI: DNA as a Biometric**

Contact Information
katherine.gettings@nist.gov
michael.coble@nist.gov