Sequencing STRs in forensic mixtures:
Current perspective on the benefits and challenges

Green Mountain DNA Conference
Burlington, VT
August 2, 2016

Katherine Gettings PhD
Applied Genetics Group

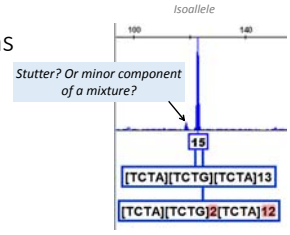
Dr. Pete Vallone
Lisa Borsuk
Kevin Kiesler
Becky (Hill) Steffen

Sequencing STRs in forensic mixtures:
Current perspective on the benefits and challenges

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.

Information presented does not necessarily represent the official position of the National Institute of Standards and Technology or the U.S. Department of Justice.

STR Sequence Applications



- One to one matching?
  - new core loci = higher CE stats
  - Partial profiles
  - Kinship
- Degraded samples?
  - Smaller PCR amplicons relative to CE
- Mixtures
  - Resolve alleles identical by length, but differ by sequence
  - Separate stutter from low level contributors (based on sequence)

Sampling of forensic NGS kits

Run on the MiSeq/FGx

- Illumina ForenSeq (FGx)
  - 27 aSTR + 24 YSTR + 7 XSTR + Amel
  - 94 HID-SNPs + 56 ancestry SNPs + 22 pheno SNPs
  - Multiplex B – all markers
  - Multiplex A – STRs and HID SNPs

- Promega PowerSeq kits (Auto, Y, Mito)
  - 22 aSTR + DYS391 + Amel
  - 22 aSTR +21 YSTR

Run on the PGM or S5

- Thermo Fisher Precision ID (STR, SNP, Mixtures & Mito)
  - Mixture panel: STR, MH, IISNP
  - 29 aSTR + DYS391 + Amel
  - 124 Identity SNPs
  - 165 Ancestry SNPs
  - Mito control region or whole genome

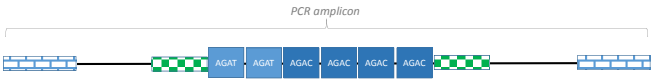
STR Sequence Gains

- PowerSeq Auto (MiSeq)
- 183 samples (3 US pops)
- STRait Razor (& ExactID)

- ForenSeq (FGx)
- Currently working sequencing NIST population samples (>1000)
- Sequencing nearly complete
- Bioinformatic checks
- CE data backfilling
- The goal is to provide the allele frequencies of the sequenced STRs

A little bit of informatics

Recognition Site-Based Informatics for STRs



Software returns:
The length between the recognition sequences (= 24)
A reference table returns a "6" allele **and the sequence between the recognition sites**

PCR primers
Recognition site (≈10 nt)
STR repeat region

[1] STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. Warshauer et al., Forensic Sci Int Genet. 2013 (7):409-17
[2] STRait Razor v2.0: the improved STR Allele Identification Tool--Razor. Warshauer et al., Forensic Sci Int Genet. 2015 (14):182-6
[3] *http://battelleexactid.org/*

Recognition Site-Based Informatics for STRs

Moving the recognition sites out further:
Captures the flanking region SNPs and Indels
Still returns the allele length and **more sequence between the recognition sites**

PCR primers
Recognition site (≈10 nt)
STR repeat region

STR sequence bracketing (custom scripts)

- Makes it easier *for humans* to 'read' the STR sequence
- Check known list of sequence orientation
  - reverse complement sequence if needed
- Select known motif(s) specific for locus
  - D2S1338 – GGAA, GGCA
- 'Move' through sequence identifying repeats
  - GGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGGCAGGCA
  - [GGAA]11 GGCAGGCAGGCAGGCAGGCA
  - [GGAA]11 [GGCA]6
- Additional trimming for sequences with known flanking sequence present
- Additional processing for loci which require specific handling – special rules and exceptions

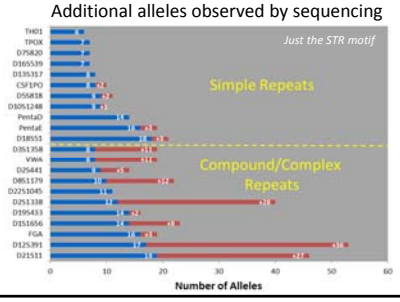
Data analysis

- FASTQ files were analyzed by STRaitRazor
- STRaitRazor data was parsed with custom scripts
- Allele calling: majority coverage and/or het balance (0.4)
  - Data compared to CE calls; discordant allele calls addressed
- Allele sequences were counted for each locus

- Allele not present in lookup table
- SNP or InDel caused a 'bioinformatic null'
- Presence of an InDel resulted in a length difference from CE

Forensic STR Sequence Diversity

Additional alleles observed by sequencing

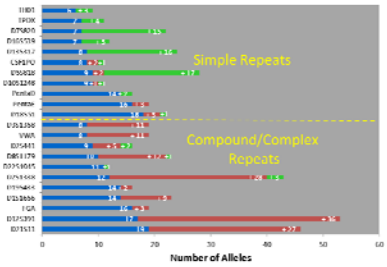


Length
Sequence

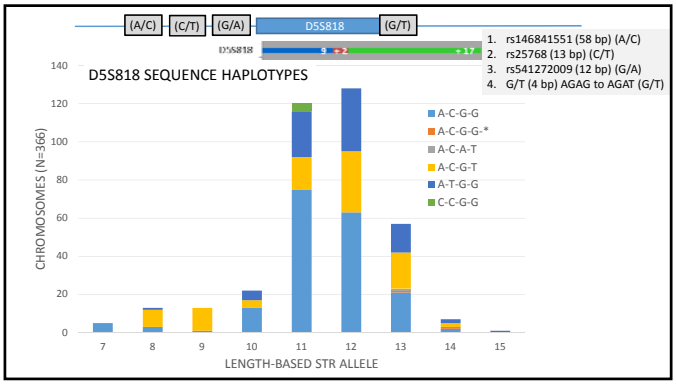
STR sequences provided in paper

Forensic STR Sequence Diversity

Additional alleles observed by sequencing



Length
Sequence
Sequence + Flank



1. rs146841551 (58 bp) (A/C)
2. rs25768 (13 bp) (C/T)
3. rs541272009 (12 bp) (G/A)
4. G/T (4 bp) AGAG to AGAT (G/T)

Category	Locus	rs Number	Minor Allele Frequency		
			African American	Caucasian	Hispanic
Population / Allele Specific SNPs	TPOX	rs13422969	0.135	-	0.011
		rs115644759	0.022	-	-
		rs149212737	-	0.007	-

TPOX: rs13422969 and two additional rare SNPs

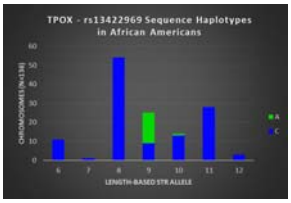
- Well distributed across Africa, rare elsewhere

Category	Locus	rs Number	Minor Allele Frequency		
			African American	Caucasian	Hispanic
Population / Allele Specific SNPs	TPOX	rs13422969	0.135	-	0.011
		rs115644759	0.022	-	-
		rs149212737	-	0.007	-

TPOX: rs13422969 and two additional rare SNPs

- Well distributed across Africa, rare elsewhere

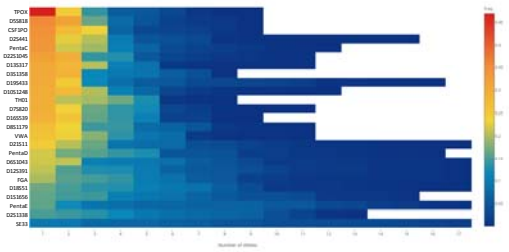
Associated with “9” allele



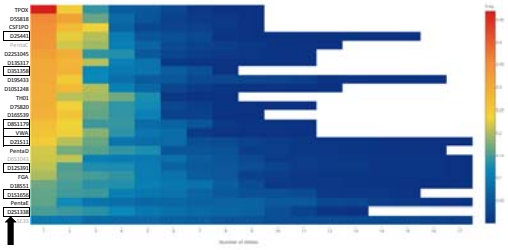
Flanking region SNP categories from paper

1. **Multiple polymorphisms in haplotype (D5S818 example)**
2. “Old” single polymorphisms
3. **Population or allele specific polymorphisms (TPOX example)**
4. Polymorphisms associated with STR sequence variants
5. Rare polymorphisms (< 5%)
6. No polymorphisms

Allele Distribution by Length

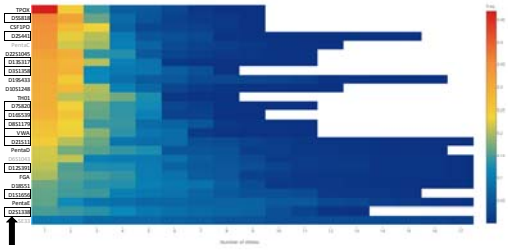


Allele Distribution by Length



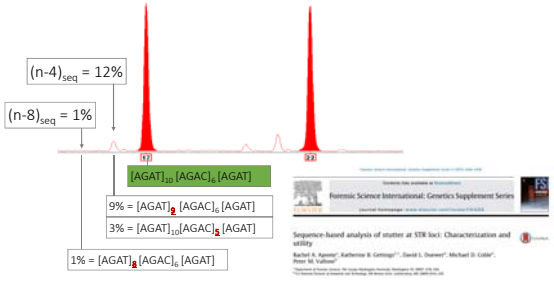
Loci which gain the most from sequencing the repeat region

Allele Distribution by Length

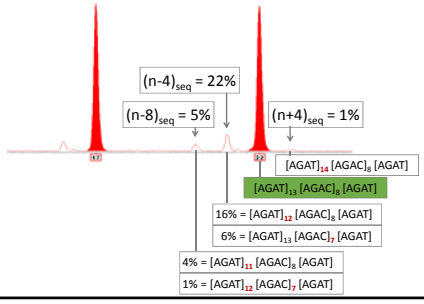


Loci which gain the most from sequencing the repeat region & flanking region (PowerSeq)

D12S391 stutter by sequence

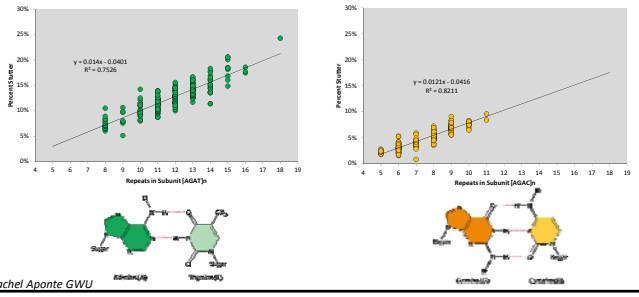


D12S391 stutter by sequence $[AGAT]_n[AGAC]_n$



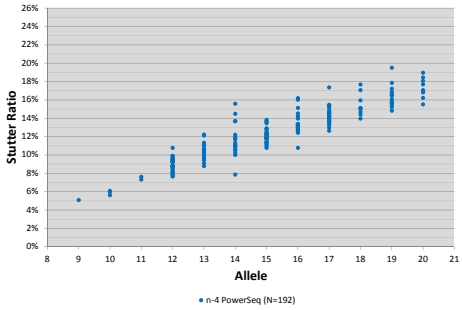
D12S391 stutter by sequence

[AGAT]n trends ~3% higher stutter than [AGA**C**]n

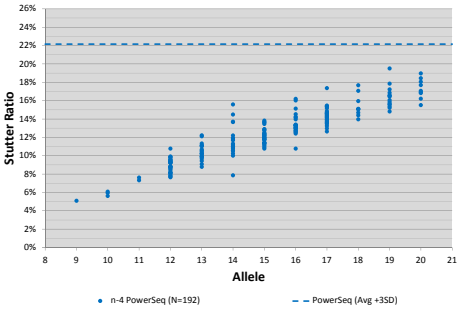


Rachel Aponte GWU

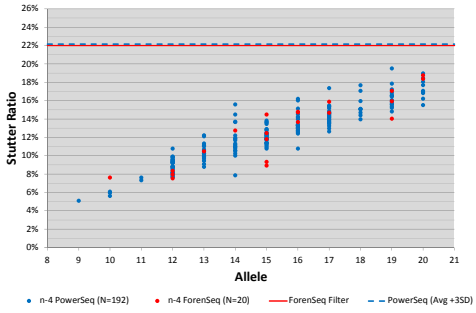
D18S51 n-4 Stutter by Kit



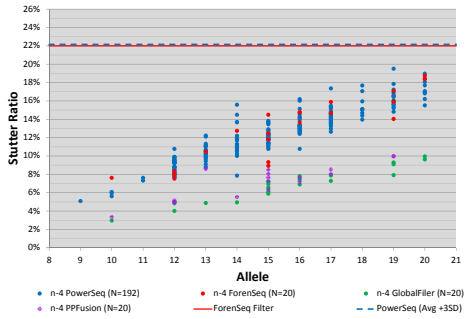
D18S51 n-4 Stutter by Kit



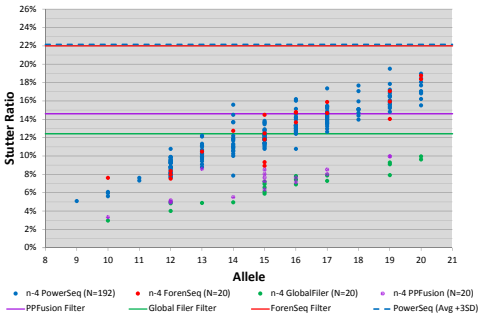
D18S51 n-4 Stutter by Kit



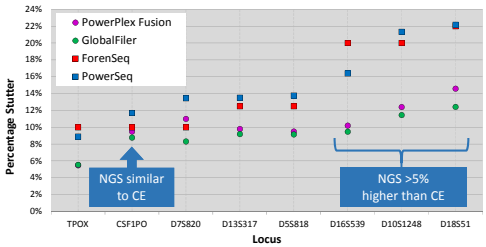
D18S51 n-4 Stutter by Kit



D18S51 n-4 Stutter by Kit



n-4 Stutter Rates by Kit at 8 Simple Repeat Loci



Stutter rates obtained from:

PowerPlex Fusion: K. Oostdik et al. / Forensic Science International: Genetics 12 (2014) 69-76

GlobalFiler: ThermoFisher Scientific GlobalFiler™ PCR Amplification Kit User Guide

ForenSeq: ForenSeq™ Universal Analysis Software Guide, Part # 15053876 Rev. C, June 2015

PowerSeq: Analysis of NIST population samples (N=192), Average + 3 Standard Deviations

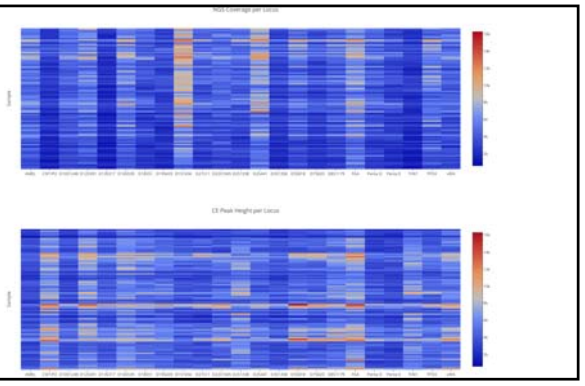
NGS Mixture Study

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

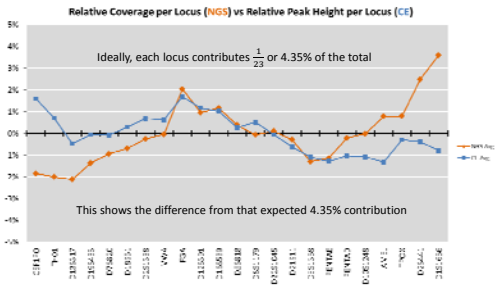
	CE		NGS
Loci	PowerPlex Fusion + PowerPlex Y23		PowerSeq Auto + Y
Input DNA	0.5 ng each		0.5 ng total
Amp Parameters	30 cycles		30 cycles, same as PPF
Everything Else	3500xL		TruSeq PCR free Library Prep, MiSeq v3

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	13:1 A/B/M	13:1 A/B/M	13:1 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



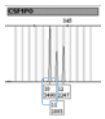
NGS Mixture Study

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

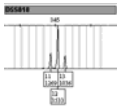
$$\frac{3\ alleles\ "10"}{6\ alleles\ total} = 0.5$$



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

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

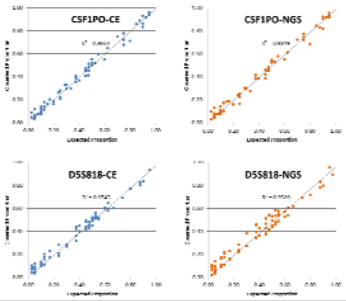
	A	B	M
DSS818	11,12	12,13	12,12



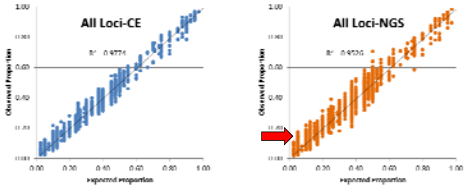
DSS818	Expected	Rep 1	Rep 2	Rep 3	Average
11	0.167	0.217	0.178	0.285	0.227
12	0.667	0.605	0.659	0.533	0.599
13	0.167	0.177	0.163	0.182	0.174

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

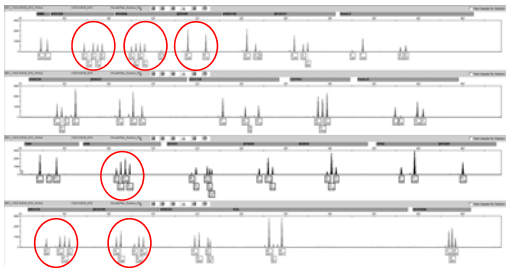
Expected vs Observed Mixture Contributions



Expected vs Observed Mixture Contributions

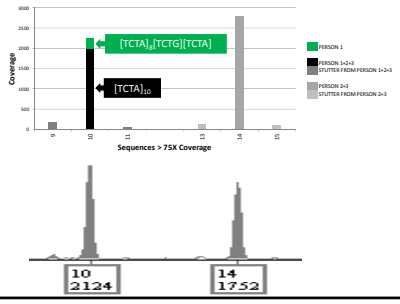


NGS Mixture Study 1:9:9



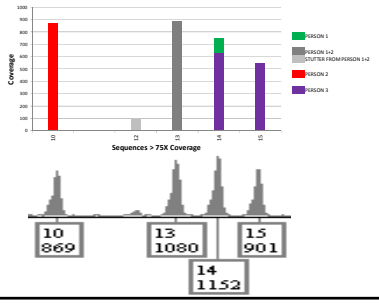
D2S441 1:9:9

1 additional allele



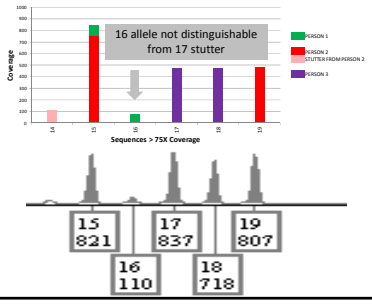
D8S1179 1:9:9

1 additional allele



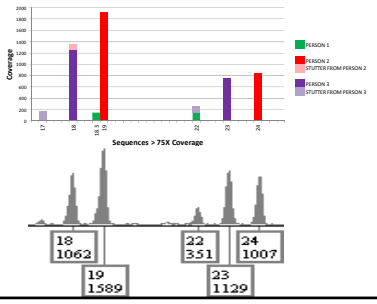
D3S1358 1:9:9

1 additional allele



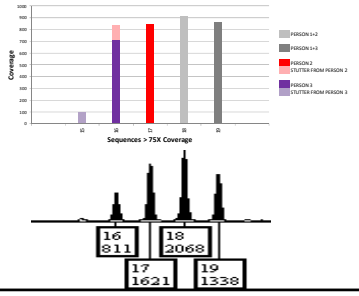
D12S391 1:9:9

1 additional allele
2 alleles help with stutter



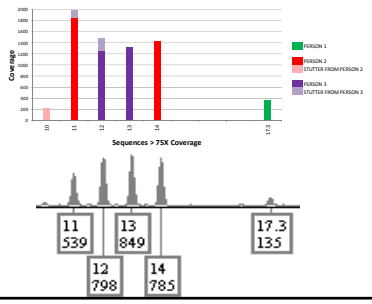
vWA 1:9:9

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

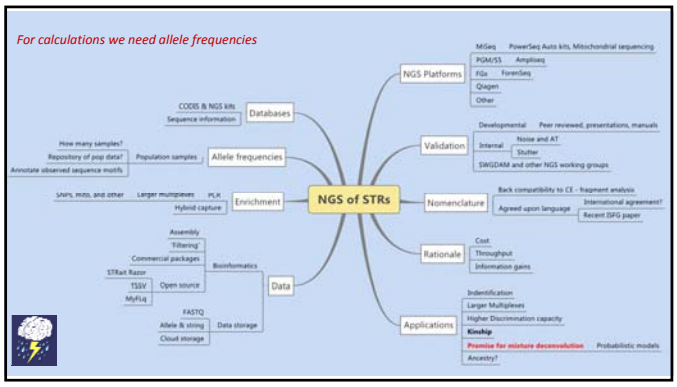
Conclusions

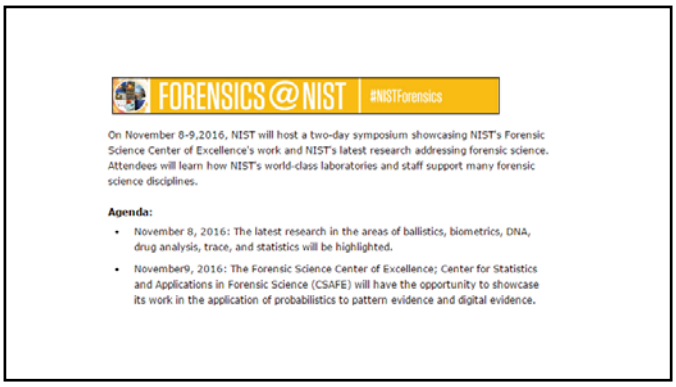
Prior to implementation:

- Sequence-based allele frequency databases
- Characterization of any NGS-specific effects on
  - peak height ratios
  - minor components/stutter

assay and locus specific

- Probabilistic genotyping software amenable to sequence data (and sequence-based stutter!)





Acknowledgments

- Rachel Aponte (GWU – now at Bode)
- Funding
  - NIST - Forensic DNA
  - FBI - DNA as a Biometric

Dr. Pete Vallone
Lisa Borsuk
Kevin Kiesler
Becky (Hill) Steffen
Dr. Mike Coble

katherine.gettings@nist.gov