Forensic Genetic NGS: Sequence Diversity of STRs

Peter M. Vallone, Ph.D.
Leader, Applied Genetics Group

NGS/MPS Workshop
26th Congress of the International Society for Forensic Genetics
August 31, 2015
Krakow, Poland

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

Disclaimer

We will mention commercial products and information, but we are in no way attempting to endorse any specific products.

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.

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

Our group receives or has received funding from the FBI Laboratory and the National Institute of Justice.

Outline for Today

- STR sequence diversity
- Investigating STR sequence diversity at NIST
  - Informatics
  - Sequence and genotype diversity
  - Flanking region diversity
  - Stutter
- NIST resources for NGS researchers

NGS Activities at NIST - Overview

Samples	Markers	Platforms	Assays
SRM 2391c	STR	MiSeq	PowerSeq Auto System
NIST Population Samples	mtDNA	PGM	ForenSeq Panel
SRM 2392/2392I (mtDNA)	SNP		AmpliSeq Identity SNP Panel
GWU Samples			AmpliSeq Ancestry SNP Panel

NGS Activities at NIST - Overview

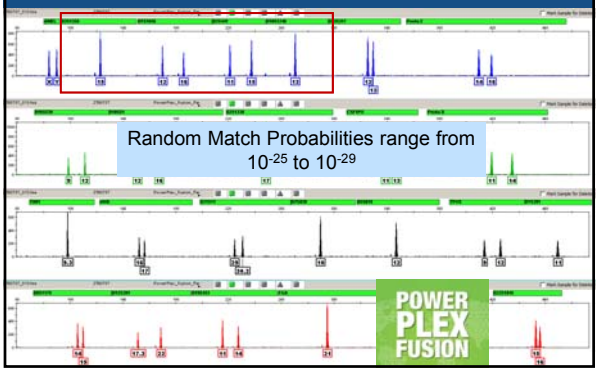
Samples	Markers	Platforms	Assays
SRM 2391c	STR	MiSeq	PowerSeq Auto System
NIST Population Samples	mtDNA	PGM	ForenSeq Panel
SRM 2392/2392I (mtDNA)	SNP		AmpliSeq Identity SNP Panel
GWU Samples			AmpliSeq Ancestry SNP Panel

NGS has potential for finer resolution of STR amplicons not detectable by CE-length based methods

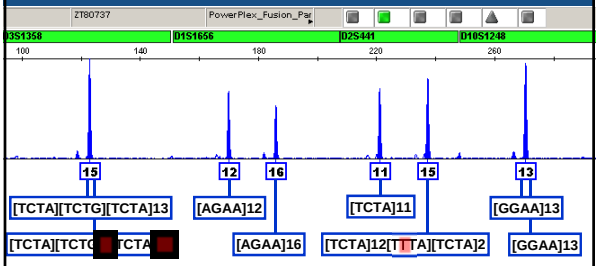
- Additional STR alleles
- Flanking region SNPs and InDels

- *Resolve homozygous by length peaks*
- *Resolve minor contributor peaks from stutter*

Forensic STR Sequence Diversity

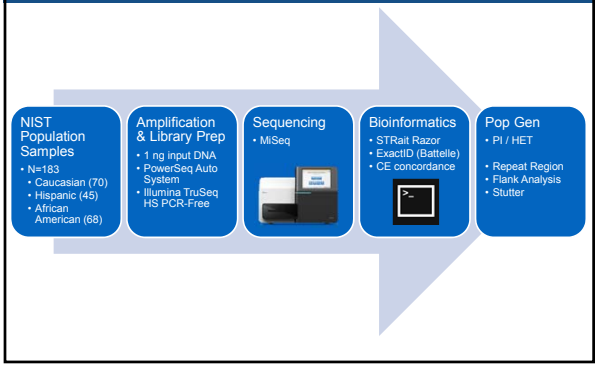


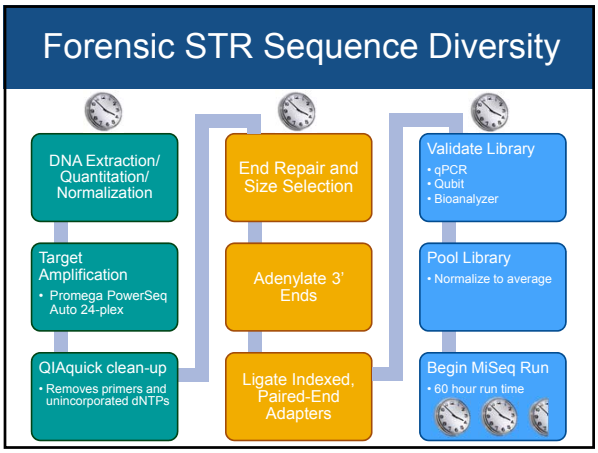
Forensic STR Sequence Diversity

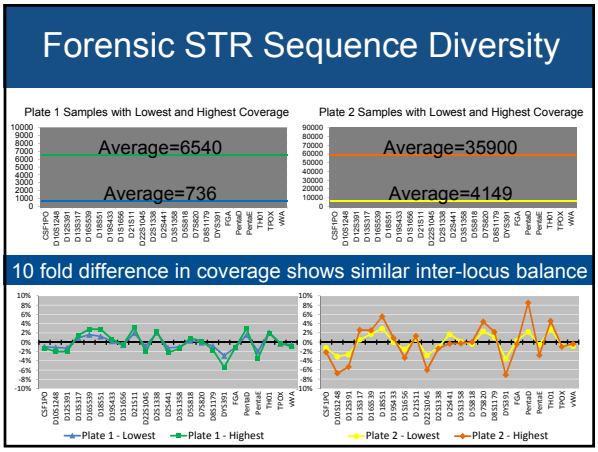


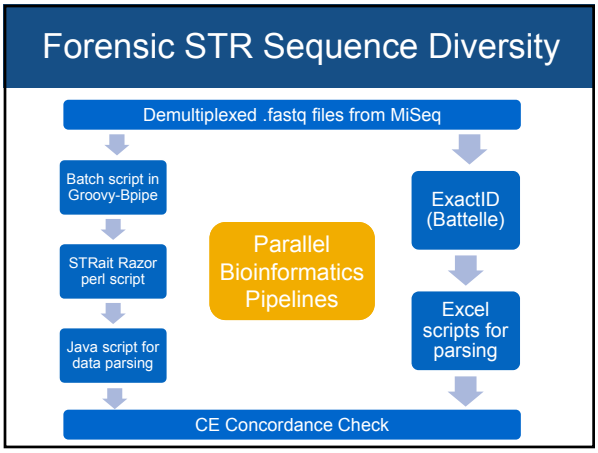
Sequence-Based Heterozygote or Isoallele: A locus that appears homozygous in length-based measurements (such as CE), but is heterozygous by sequence

Forensic STR Sequence Diversity




Forensic STR Sequence Diversity
DNA Extraction/ Quantitation/ Normalization
Target Amplification
• Promega PowerSeq Auto 24-plex
QIAquick clean-up
• Removes primers and unincorporated dNTPs
End Repair and Size Selection
Adenylate 3' Ends
Ligate Indexed, Paired-End Adapters
Validate Library
• qPCR
• Qubit
• Bioanalyzer
Pool Library
• Normalize to average
Begin MiSeq Run
• 60 hour run time


Forensic STR Sequence Diversity
Plate 1 Samples with Lowest and Highest Coverage
Average=6540
Average=736
Plate 2 Samples with Lowest and Highest Coverage
Average=35900
Average=4149
10 fold difference in coverage shows similar inter-locus balance
Plate 1 - Lowest
Plate 1 - Highest
Plate 2 - Lowest
Plate 2 - Highest


Forensic STR Sequence Diversity
Demultiplexed .fastq files from MiSeq
Batch script in Groovy-Bpipe
STRait Razor perl script
Java script for data parsing
Parallel Bioinformatics Pipelines
ExactID (Battelle)
Excel scripts for parsing
CE Concordance Check

Recognition Site-Based Informatics for STRs

Computer Returns:
The length between the recognition sequences = 36 bp
Reference table in software returns a "9" allele
The sequence between the recognition sites

PCR Primers
STR Repeat Region
Recognition site (~10 nt)

[1] *http://battelleexactid.org/*
[2] 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
[3] STRait Razor v2.0: the improved STR Allele Identification Tool--Razor. Warshauer et al., Forensic Sci Int Genet. 2015 (14):182-6

Forensic STR Sequence Diversity

CE (length-based genotype) concordance check results

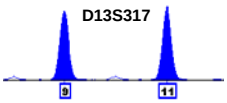
24 loci x 183 samples = 4392 loci evaluated

- 99% concordance with CE data

Why were some discordances observed between the CE and NGS measurements…informatics

Forensic STR Sequence Diversity

Obs 5/183



Sequence $[TATC]_{10}$,$[TATC]_{11}$

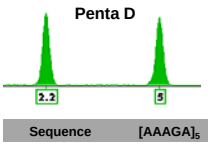
← Repeat Region NGS Recognition Region 4 bp Deletion CE Primer Binding Site→

```
TATC TATC TATC AATCAATCATCTATCTATCTTTCTGTC----TTTTTGGGCTGCCTATGGCTCAA
TATC TATC TATC AATCAATCATCTATCTATCTTTCTGTCTGTCTTTTTGGGCTGCCTATGGCTCAA
```

Flanking region InDel: Bioinformatic pipelines may reduce the region used for genotyping, resulting in deletions not being "counted" as they would via CE

Forensic STR Sequence Diversity

Obs 15/183



Sequence	$[AAAGA]_5$

13 bp deletion in Recognition Region

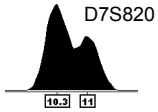
←CE Primer Binding Site — Repeat Region →

```
TAGGTTACAGAGCAAGACACCATCTCAAG------------ AAAGA AAAGA AAAGA AAAGA AAA
TAGGTGACAGAGCAAGACACCATCTCAAGAAAGAAAAAAAG AAAGA AAAGA AAAGA AAAGA AAA
```

Bioinformatic Null Allele: A true allele that is present within the raw sequence data but is not detected by the bioinformatic pipeline

Forensic STR Sequence Diversity

Obs 1/183



NGS Sequence	$[GATA]_{11}$
	$[GATA]_{11}$

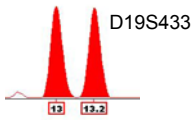
REPEAT REGION — RECOGNITION SITE — 1 bp DELETION

```
GATA GATA GATA GATA GATA GACAGATTGATAGTTTTTTTT-ATCTCACTAAATAGTCTATAGT
GATA GATA GATA GATA GATA GACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGT
```

Flanking region InDel: Bioinformatic pipelines may reduce the region used for genotyping, resulting in deletions not being "counted" as they would via CE

Forensic STR Sequence Diversity

Obs 3/183



NGS Sequence	$[AAGG]_{13}$

2 bp deletion in

REPEAT REGION — RECOGNITION REGION

```
AAGG AAGG AAGG AAGG AAGG AAGG AAGG AGAGAGGAAGAAAGAGAGAAGATTTTTATTCGGGT
AAGG AAGG AAGG AAGG AAGG AAGG AAGG AGAG--GAAGAAAGAGAGAAGATTTTTATTCGGGT
```

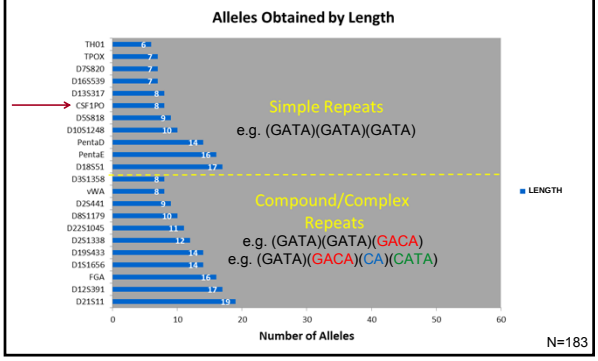
Bioinformatic Null Allele: A true allele that is present within the raw sequence data but is not detected by the bioinformatic pipeline

Comments

- This is not a criticism of the informatics tools/methods
  - After discovery the listed discordance issues are easily addressed
- No issues with sequencing the amplicons
- How do we want to capture STR allele information?
  - Defined by recognition sites adjacent to the flanking region?
  - Defined by PCR primers?
  - Some set of core recognition sites?

Maintain back compatibility with existing technology (CE) enable comparisons between labs

Forensic STR Sequence Diversity



N=183

Forensic STR Sequence Diversity

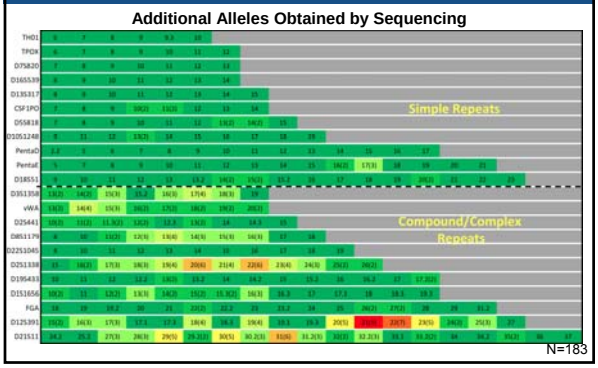
Additional Alleles by Sequence

		CSF1PO
7	[AGAT]7	AGAT AGAT AGAT AGAT AGAT AGAT AGAT
8	[AGAT]8	AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT
9	[AGAT]9	AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT
10	[AGAT]10	AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT
10	[AGGT][AGAT]9	AGGT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT
11	[AGAT]11	AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT
11	[AGAT]3 AGGT [AGAT]7	AGAT AGAT AGAT AGGT AGAT AGAT AGAT AGAT AGAT AGAT AGAT
12	[AGAT]12	AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT
13	[AGAT]13	AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT
14	[AGAT]14	AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT

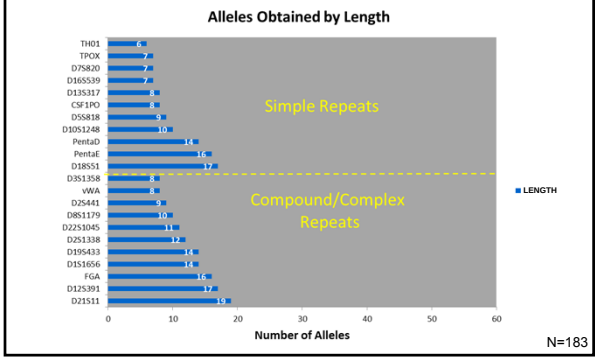
8 alleles by length → 10 alleles by sequence

N=183

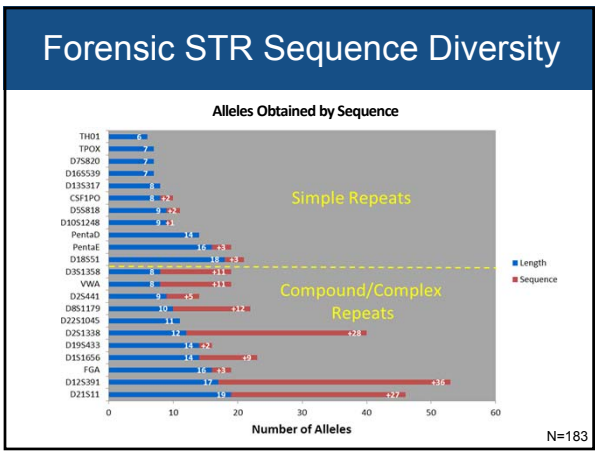
Forensic STR Sequence Diversity



Forensic STR Sequence Diversity



Forensic STR Sequence Diversity



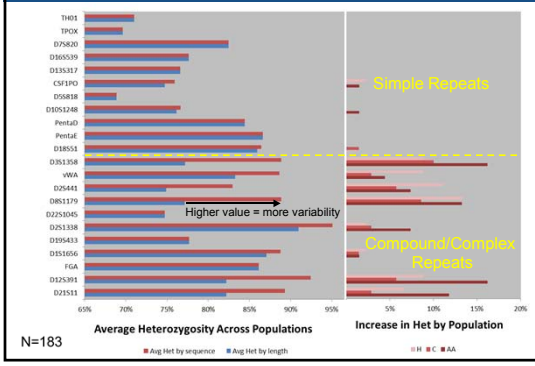
Forensic STR Sequence Diversity

Heterozygosity

heterozygotes observed
of loci tested

Indicates genetic variability at a locus

Forensic STR Sequence Diversity



Forensic STR Sequence Diversity

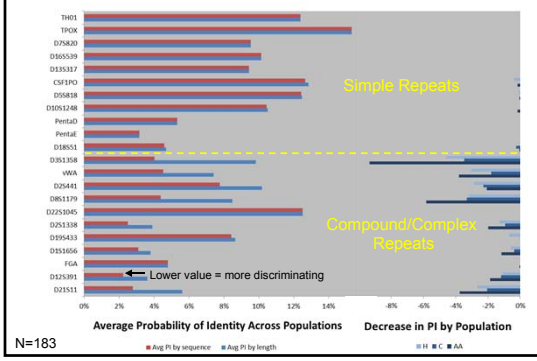
Probability of Identity

Sum of each genotype frequency^2 at each locus

$$= \sum_{i=1}^{n} xi^2$$

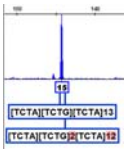
Probability that two unrelated individuals
selected at random
will have the same genotype at a locus

Forensic STR Sequence Diversity



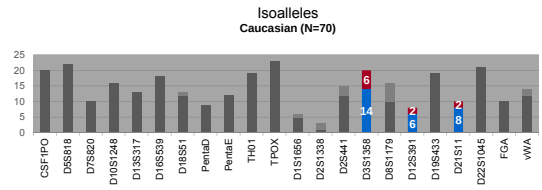
Summary of isoalleles

- 22 loci x 183 samples = 4026 genotypes
  - 828 are homozygous by length
  - 103 instances of Isoalleles (12.4%)



- 79/183 samples contained one or more isoalleles (43.7%)
  - 58 samples (1)
  - 18 samples (2)
  - 3 samples (3)

Forensic STR Sequence Diversity

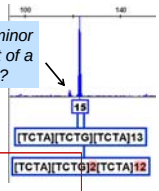


Frequency of Sequence-Based Heterozygotes		
Locus	Gelardi (N=197)	NIST (N=70)
D3S1358	23%	30%
D12S391	33%	25%
D21S11	41%	20%

Isoalleles

- Within an individual
  - Sequence based heterozygote
  - Increased heterozygosity ($p^2 \rightarrow 2pq$)
  - Marginal benefits for one-to-one matching?
- Between individuals (DNA mixtures)
  - Resolve overlapping alleles
    - 15 → '15 from person A' and '15 from person B'
  - Resolve stutter artifacts versus minor components

Stutter or minor component of a mixture?



Manuscript submitted to FSIG

Title:

Sequence variation of 22 autosomal STR loci detected by next generation sequencing

Authors and Affiliations:

Katherine Butler Gettings[1*], Kevin M. Kiesler[1], Seth A. Faith[2], Elizabeth Montano[3], Christine H. Baker[3], Brian A. Young[3], Richard A. Guerrieri[3], and Peter M. Vallone[1]

Flanking Region Variation

Recognition Site-Based Informatics for STRs

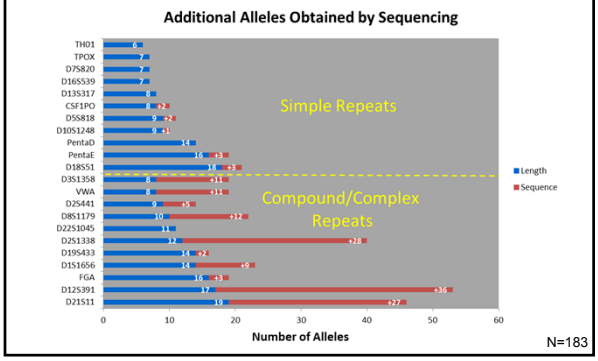
Moving recognition sites out will capture information within the flanking regions

PCR Primers

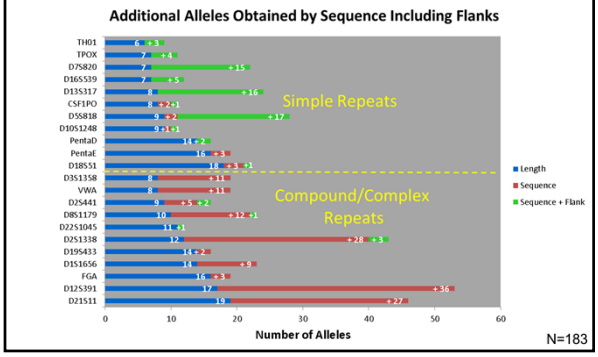
STR Repeat Region

Recognition site (~10 nt)

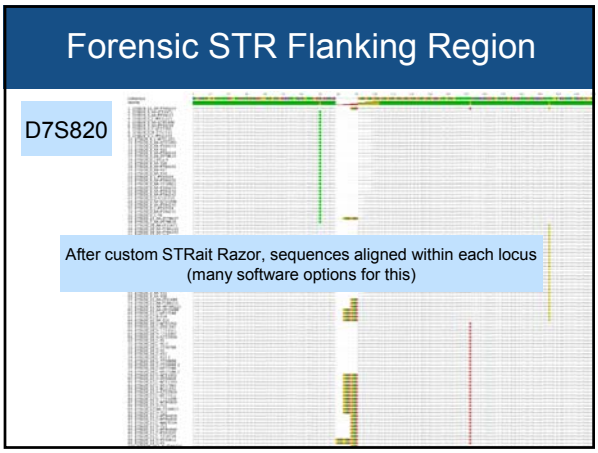
Forensic STR Sequence Diversity



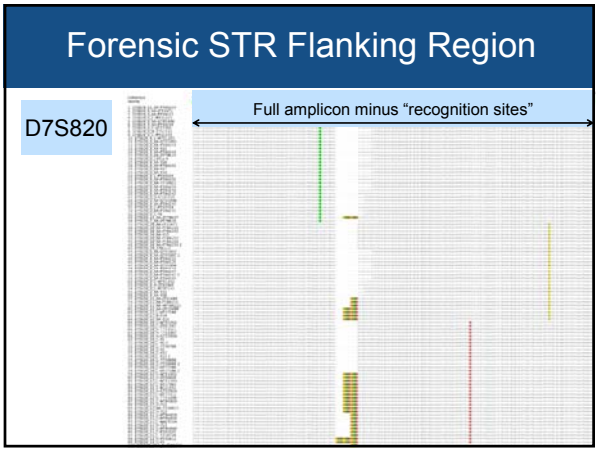
Forensic STR Sequence Diversity



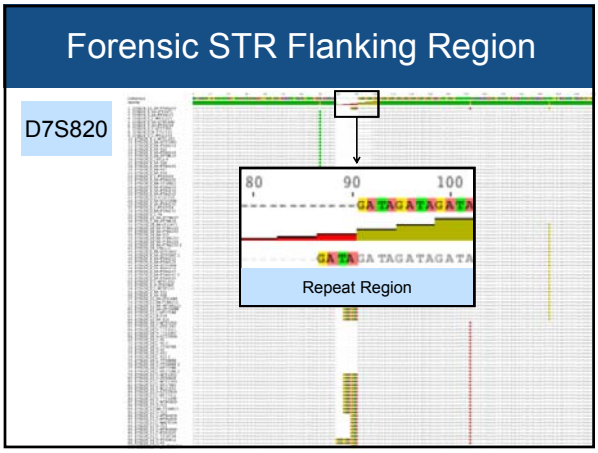
Forensic STR Flanking Region



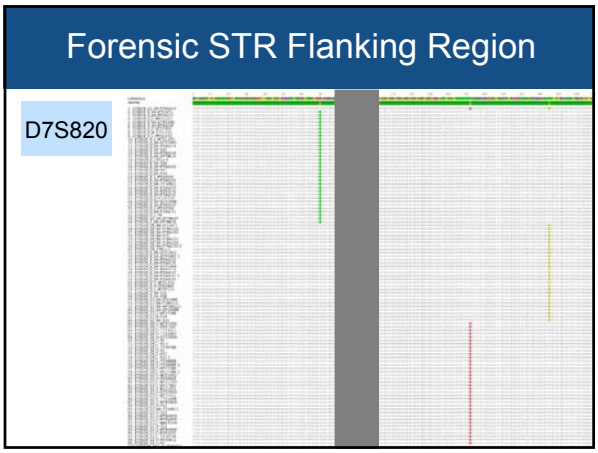
Forensic STR Flanking Region



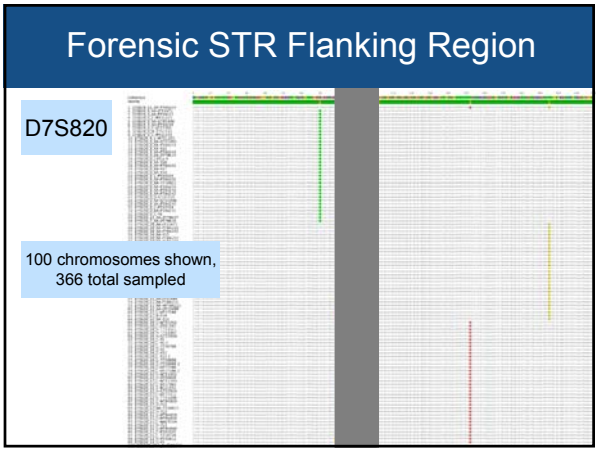
Forensic STR Flanking Region



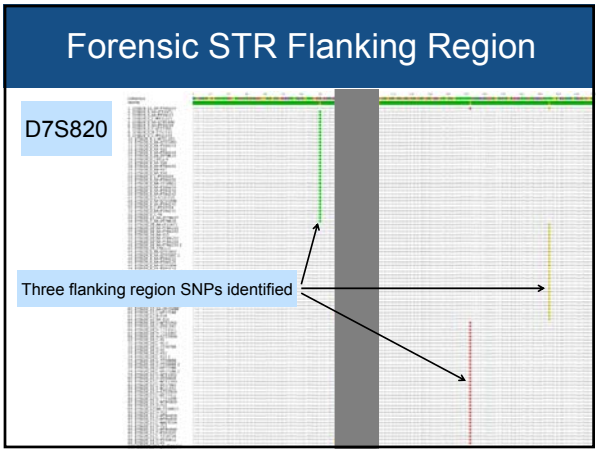
Forensic STR Flanking Region



Forensic STR Flanking Region



Forensic STR Flanking Region

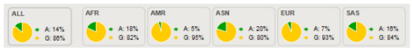


Forensic STR Flanking Region

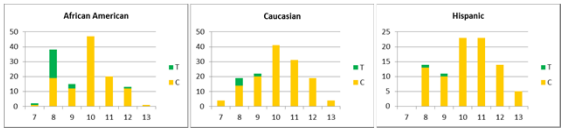
D7S820

Distribution of *rs16887642* by Population in 1000 Genomes Project

rs16887642 (note- opposite strand A/G reported):



Distribution of *rs16887642* by Population / Allele in N=183



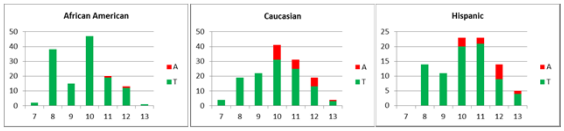
Forensic STR Flanking Region

D7S820

Distribution of *rs7789995* by Population in 1000 Genomes Project

rs7789995 (note- opposite strand A/T reported):

Distribution of *rs7789995* by Population / Allele in N=183



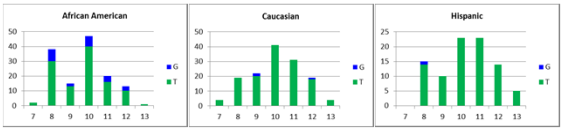
Forensic STR Flanking Region

D7S820

Distribution of *rs7786079* by Population in 1000 Genomes Project

rs7786079 (note- opposite strand A/C reported):

Distribution of *rs7786079* by Population / Allele in N=183



Forensic STR Flanking Region

D7S820

Allele	Sequence	rs16887642-rs7789995-rs7786079
7	[GATA]7	C-T-T
8	[GATA]8	C-T-T
9	[GATA]9	C-T-G
10	[GATA]10	C-T-T
11	[GATA]11	T-T-T
12	[GATA]12	C-T-G
13	[GATA]13	C-T-T
9	[GATA]9	T-T-T
10	[GATA]10	C-T-G
10	[GATA]10	C-T-T
10	[GATA]10	C-A-T
10.3/11	[GATA]11	C-_-T (deletion)
11	[GATA]11	C-T-G
11	[GATA]11	C-T-T
11	[GATA]11	C-A-G
11	[GATA]11	C-A-T
12	[GATA]12	C-T-G
12	[GATA]12	C-T-T
12	[GATA]12	T-T-T
12	[GATA]12	C-A-T
13	[GATA]13	C-T-T
13	[GATA]13	C-A-T

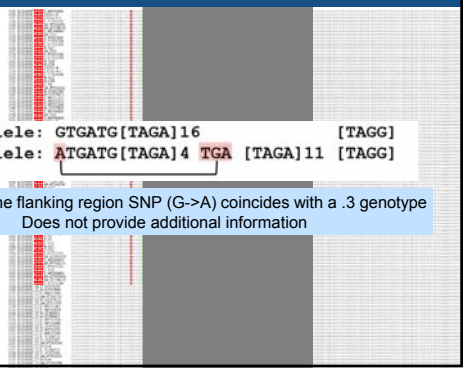
15 Additional Alleles

Forensic STR Flanking Region

D1S1656

Forensic STR Flanking Region

D1S1656



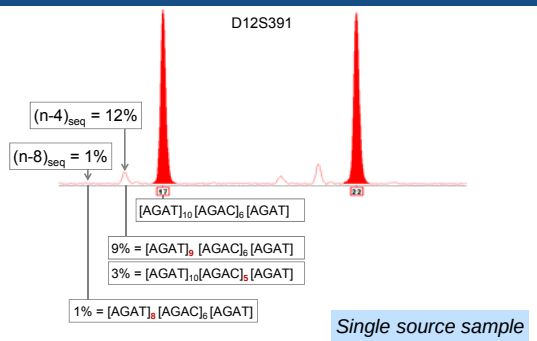
In this case the flanking region SNP (G->A) coincides with a .3 genotype
Does not provide additional information

Flanking region manuscript in preparation & ISFG oral presentation

Wednesday

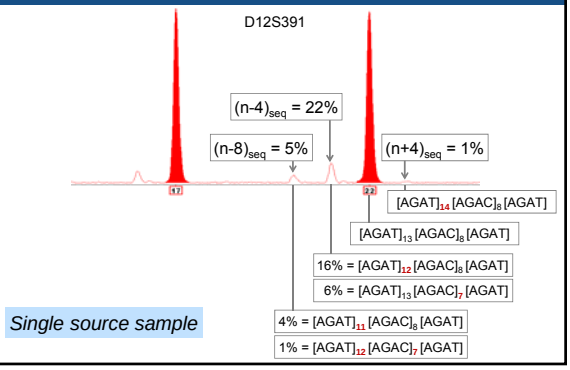
11.00-11.30	Coffee & Posters
11.30-11.45	David Ballard Complex mixture interpretation using massively parallel sequencing
11.45-12.00	Kristiaan van der Gaag Massive parallel sequencing of autosomal short tandem repeats and SNPs, the next level of forensic mixture analysis
* 12.00-12.15	Katherine Gettings The next dimension in STR sequencing: polymorphisms in flanking regions and their allelic associations
12.15-13.00	Panel discussion Walther Parson, Chris Phillips, Katherine Gettings, Peter de Knijff, David Ballard, Bruce Budowle, Niels Morling **Sequencing based STR nomenclature**
13.00-14.30	Lunch

Stutter by Sequence



Single source sample

Stutter by Sequence



Single source sample

Signal and Noise

- Sources of noise
  - Biological and sequencing chemistry
- Even the 'majority peaks' (alleles) will contain some sequence variants
- Interpretable for single source samples
- BUT we need to understand S/N in order to use NGS to resolve mixtures
  - Is the 'other' sequence from another contributor or just noise?

Poster #100

SEQUENCE-BASED ANALYSIS OF STUTTER AT STR LOCI: CHARACTERIZATION AND UTILITY (P)

R.A. Aponte[1], K.B. Gettings[2], D.L. Duewer[2], M.D. Coble[2], and P. M. Vallone[2]

[1] Department of Forensic Sciences, The George Washington University, Washington, DC, U.S.
[2] National Institute of Standards and Technology, Gaithersburg, MD, U.S.
katherine.gettings@nist.gov

Comments on mixtures

What is needed to understand the true benefits of sequencing to mixtures?

- Validate analytical thresholds for NGS work
- Assess sensitivity of NGS methods to detect minor components
  - Still PCR front end – will we observe better than 10:1?
- In practice - how often will *resolvable* overlapping alleles be observed in a mixture?
- Need the allele frequencies for 'new' alleles
  - What size population databases are needed? Greater than 200?
- Incorporate NGS sequence data into probabilistic genotyping software (STRMix, True Allele, etc)
  - What are the gains in stats (Log LR)? Improved contributor ratio estimates?

More loci – better loci?

NIST Support for Forensic NGS Research
FSI Genetics review article


Contents lists available at ScienceDirect

Forensic Science International: Genetics

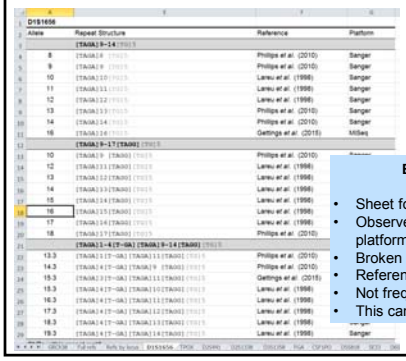
journal homepage: www.elsevier.com/locate/fsig

STR allele sequence variation: Current knowledge and future issues

Katherine Butler Gettings[a,*], Rachel A. Aponte[b], Peter M. Vallone[a], John M. Butler[c]

Updating observed STR sequence variations for 24 autosomal loci on STRBase

NIST Support for NGS Research


D1S1656

Allele	Repeat Structure	Reference	Platform
	[TAGA]9-14		
8	[TAGA]8	Phillips et al. (2010)	Sanger
9	[TAGA]9	Phillips et al. (2010)	Sanger
10	[TAGA]10	Lareu et al. (1998)	Sanger
11	[TAGA]11	Lareu et al. (1998)	Sanger
12	[TAGA]12	Lareu et al. (1998)	Sanger
13	[TAGA]13	Phillips et al. (2010)	Sanger
14	[TAGA]14	Phillips et al. (2010)	Sanger
16	[TAGA]16	Gettings et al. (2015)	MiSeq
	[TAGA]9-17[TAGG]		
10	[TAGA]9 [TAGG]	Phillips et al. (2010)	
12	[TAGA]11[TAGG]	Lareu et al. (1998)	
13	[TAGA]12[TAGG]	Lareu et al. (1998)	
14	[TAGA]13[TAGG]	Lareu et al. (1998)	
15	[TAGA]14[TAGG]	Lareu et al. (1998)	
16	[TAGA]15[TAGG]	Lareu et al. (1998)	
17	[TAGA]14[TAGG]	Lareu et al. (1998)	
18	[TAGA]17[TAGG]	Phillips et al. (2010)	
	[TAGA]1-4[T-GA][TAGA]9-14[TAGG]		
13.3	[TAGA]1[T-GA][TAGA]11[TAGG]	Phillips et al. (2010)	
14.3	[TAGA]4[T-GA][TAGA]9 [TAGG]	Phillips et al. (2010)	
15.3	[TAGA]3[T-GA][TAGA]11[TAGG]	Gettings et al. (2015)	
15.3	[TAGA]4[T-GA][TAGA]10[TAGG]	Lareu et al. (1998)	
16.3	[TAGA]4[T-GA][TAGA]11[TAGG]	Lareu et al. (1998)	
17.3	[TAGA]4[T-GA][TAGA]12[TAGG]	Lareu et al. (1998)	
18.3	[TAGA]4[T-GA][TAGA]13[TAGG]	Lareu et al. (1998)	
19.3	[TAGA]4[T-GA][TAGA]14[TAGG]	Lareu et al. (1998)	Sanger

Excel Workbook

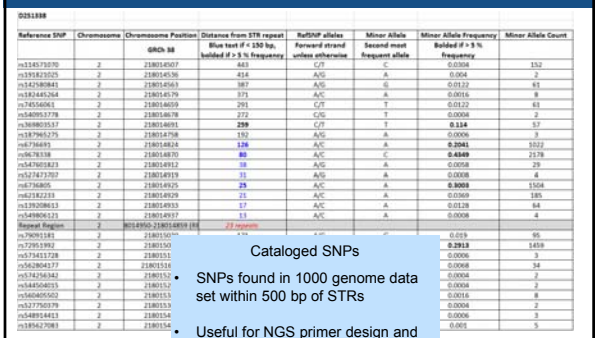
- Sheet for each STR locus
- Observed alleles, repeat structure, platform
- Broken out by sub-motif
- References
- Not frequency data
- This can be updated as needed

NIST Support for NGS Research

Annotations

- GRCh38 genome with STR repeat region and flanking SNPs identified
- A file for each locus can be downloaded

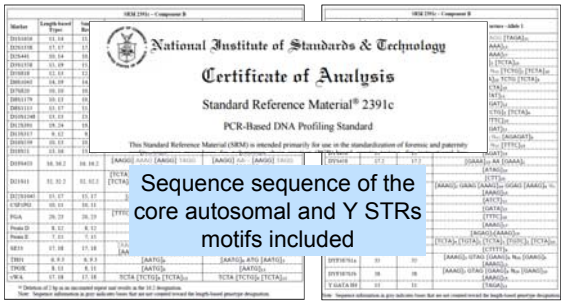
NIST Support for NGS Research



Cataloged SNPs

- SNPs found in 1000 genome data set within 500 bp of STRs
- Useful for NGS primer design and bioinformatics

NIST Support for NGS Research SRM 2391c



Sequence sequence of the core autosomal and Y STRs motifs included

https://www-s.nist.gov/srmors/view_cert.cfm?srm=2391C

Acknowledgements

NIST
Katherine Gettings
Kevin Kiesler
Nate Olson
Jo Lynne Harenza
Mike Coble
Becky Hill
Margaret Kline

Student Interns
Rachel Aponte (GWU)
Harish Swaminathan (Rutgers)
Anna Blendermann (MC)

Battelle
Seth Faith (now @NCSU)
Rich Guerrieri
Brian Young

Promega
Doug Storts
Jay Patel

Contact Information
peter.vallone@nist.gov