Sequencing STRs: Variation and Nomenclature

Forensics@NIST
December 3, 2014

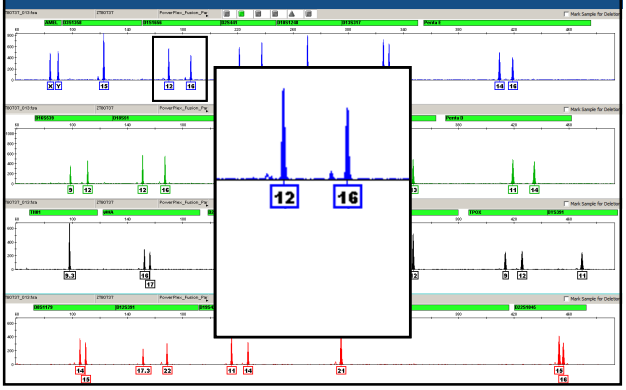
Katherine Butler Gettings, Ph.D.
Applied Genetics Group
Biomolecular Measurement Division

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

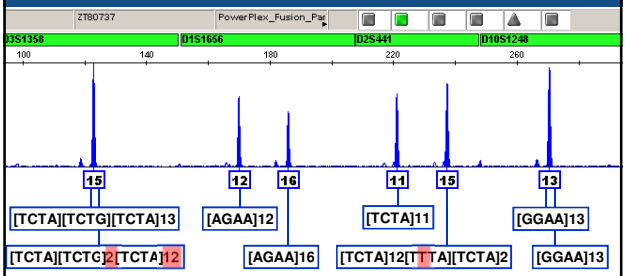
Disclaimer

Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Commerce. 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 endorsement by NIST, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose.

Forensic STR Sequence Diversity

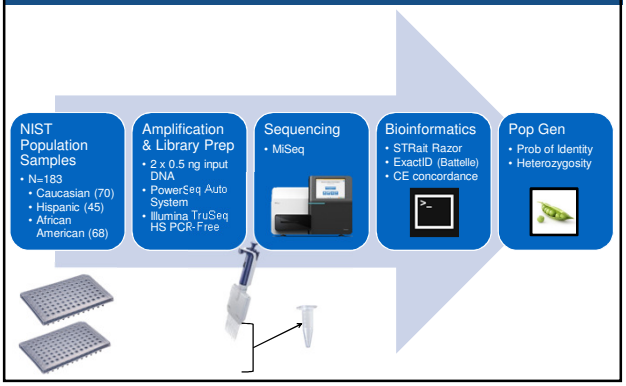


Forensic STR Sequence Diversity

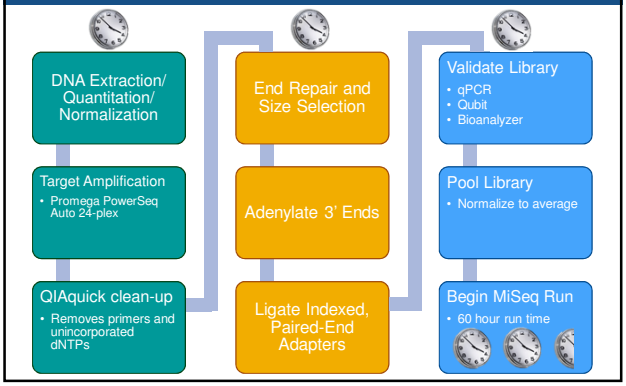


Sequence-Based Heterozygote: A locus that appears homozygous in length-based measurements (such as CE), but is heterozygous by sequence

Forensic STR Sequence Diversity



Forensic STR Sequence Diversity



Forensic STR Sequence Diversity

Plate 1 Samples with Lowest and Highest Coverage

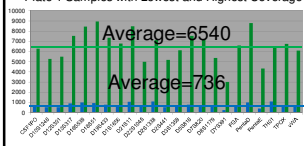
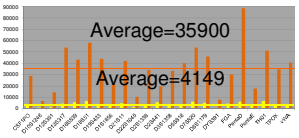
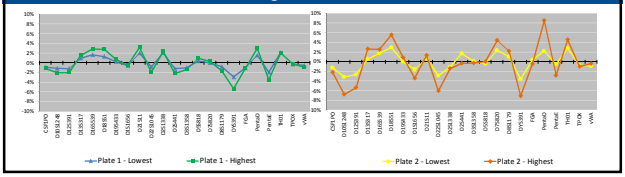


Plate 2 Samples with Lowest and Highest Coverage



10 fold difference in coverage shows similar inter-locus balance



Forensic STR Sequence Diversity

The role of bioinformatics

This is one page of sequence data

The lowest coverage sample has 2,838 pages

All loci and F/R strands are mixed together

Forensic STR Sequence Diversity

Forensic STR Sequence Diversity

Forensic STR Sequence Diversity

CE Concordance Check Results:

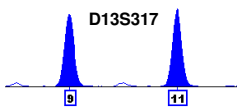
24 loci x 183 samples = 4392 loci evaluated

ExactID and STRait Razor
> 99% concordance with CE data

	Discordant Loci in CE compare		
	Exact ID Only	ExactID+ STRait Razor	STRait Razor Only
D13S317		5	
D7S820		1	
Penta D	15		
D18S51	3		
D19S433		1	2
D12S391			3

Forensic STR Sequence Diversity

	Discordant Loci in CE compare		
	Exact ID Only	ExactID+ STRait Razor	STRait Razor Only
D13S317		5	
D7S820		1	
Penta D	15		
D18S51	3		
D19S433		1	2
D12S391			3



NGS 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

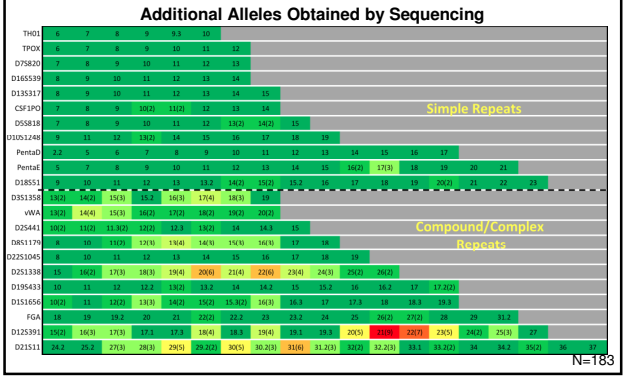
Additional Alleles by Sequence

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

8 alleles by length → 10 alleles by sequence

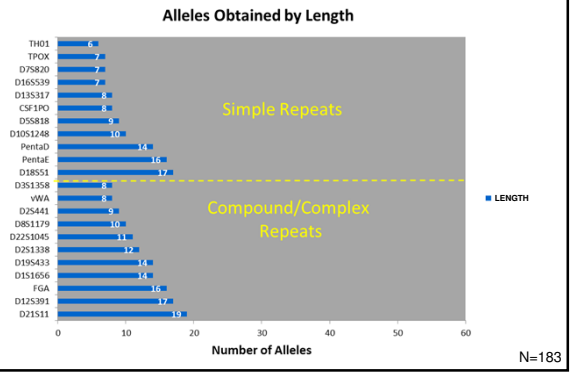
N=183

Forensic STR Sequence Diversity



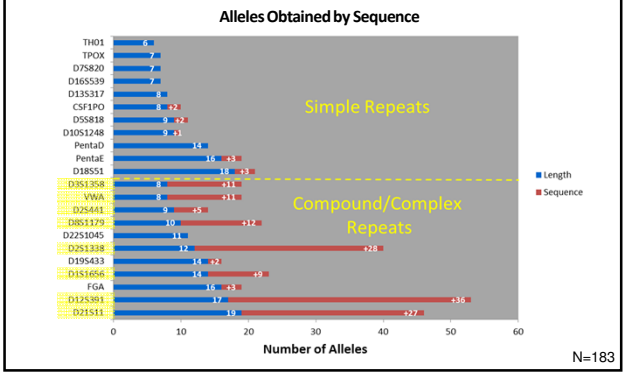
N=183

Forensic STR Sequence Diversity



N=183

Forensic STR Sequence Diversity



N=183

Forensic STR Sequence Diversity

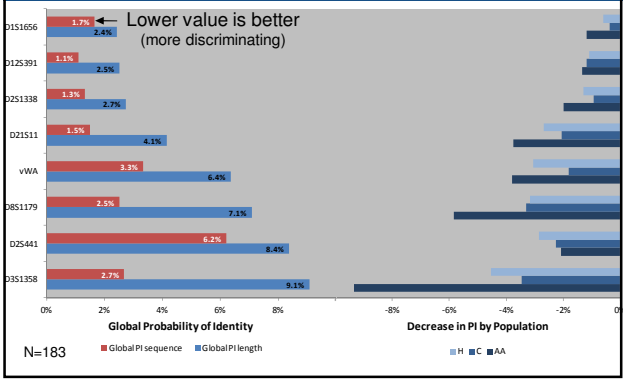
Probability of Identity

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



N=183

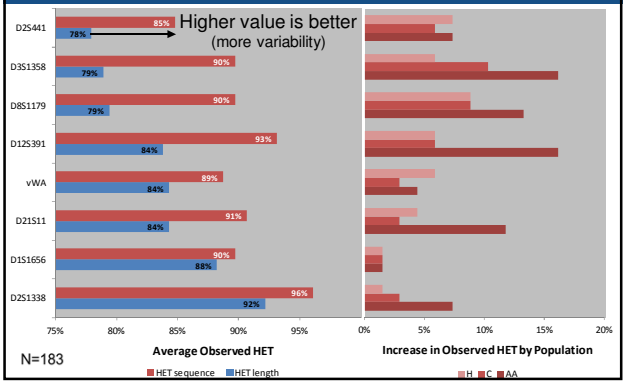
Forensic STR Sequence Diversity

Heterozygosity

$$\frac{\text{\# heterozygotes observed}}{\text{\# of loci tested}}$$

Indicates genetic variability at a locus

Forensic STR Sequence Diversity



Forensic STR Sequence Diversity

Conclusions

Sequencing forensic STR loci in a HTP manner is possible (automation is needed)

Bioinformatic tools are in their infancy, testing across platforms and pipelines is important

At some loci, sequencing will offer significant gains ("core set" for mixture analysis)

Extending analysis to the flanking regions will increase effective number of alleles

Infrastructure such as nomenclature guidelines and allele frequency databases are needed prior to implementation

Forensic STR Sequence Nomenclature

Options for representing sequence data, and possible applications:

1. **Complete Sequence String** entire string of generated sequence

TGACTATGGAGTTATTTTAAGGTTAATATATATAAAGGGTATGATAGAACACTTGTCATAGTTTAGAACGAACTAAC
GATAGATAGATAGATAGATAGATAGATAGATAGACAGATTGATAGTTTTTTTTTATCTCACTAAATAGTCTATAGTA
AACATTTAATTACCAATATTTGGTGCAATTCTGTC

2. **Bracketed sequence**
   - repetitive elements enclosed in brackets and a numeric representation of the repeat length

     $[AATG]_6$ A-TG $[AATG]_3$ = TH01 9.3 allele
   - polymorphisms (SNPs or InDels) in flanking regions identified by "rs" number
3. **Unique Identifier**
   - 13d rs206432C where 13 = repeat length, d = sequence version, rs number = flank polymorphism
   - @|*5 = computer-generated code applied to each unique sequence string within a defined region

Forensic STR Sequence Nomenclature

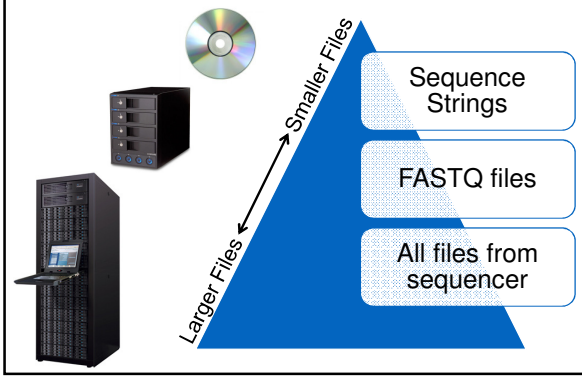
Reporting/Manual Comparisons

- Meaningful unique identifier (e.g. 13d) may be helpful for quick comparisons
- Bracketed sequence is intuitive and may help in explaining results to investigator
- Complete sequence could be appended to report

Database Searching

- Database searching must be unambigous and computationally inexpensive (i.e. fast)
- Two most likely possibilities are unique identifier and complete sequence string

Sequence Data Storage



Acknowledgements

NIST
Peter Vallone
Erica Butts
Mike Coble
David Duewer
Jo Lynne Harenza
Becky Hill
Kevin Kiesler
Margaret Kline
Nate Olson
Harish Swaminathan

Promega
Doug Storts
Jay Patel

Battelle
Seth Faith (NCSU)
Rich Guerrieri
Brian Young

Funding
FBI: DNA as a Biometric

Contact Information
katherine.gettings@nist.gov