Underlying sequence variation within STRs: considerations for nomenclature, storage, searching, and reporting

Peter M. Vallone, Ph.D.
Leader, 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

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

Outline for Today

- STR sequence diversity
- Investigating STR sequence diversity
  - Sequencing of 183 samples for 22 autosomal STRs
  - Informatics
  - Sequence diversity (within the STR motif)
  - Flanking region sequence diversity
- NIST resources for NGS researchers
- Nomenclature: thoughts and discussion

NGS Activities at NIST - Overview

Samples	Markers	Platforms	Assays
SRM 2391c	STR	MiSeq/FGx	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

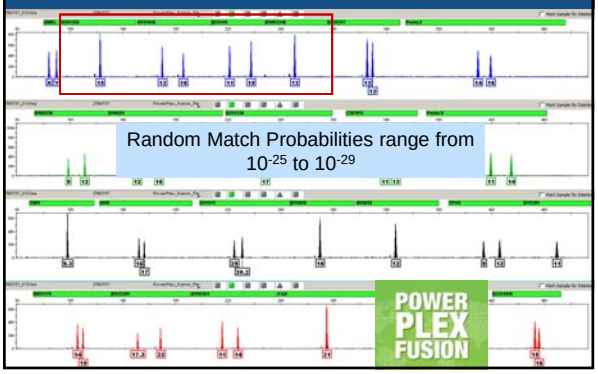
NGS Activities at NIST - Overview

Samples	Markers	Platforms	Assays
SRM 2391c	STR	MiSeq/FGx	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

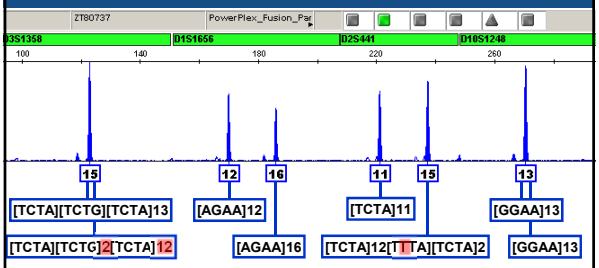
NGS has potential for finer resolution of STR amplicons not detectable by CE-length based methods

- Additional STR alleles
- Flanking region SNPs and InDels
- *Resolve minor contributor peaks from stutter*
- *Resolve homozygous by length peaks*

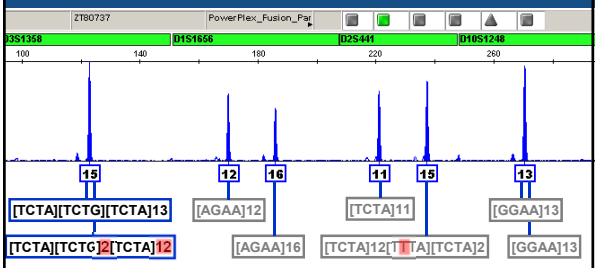
Forensic STR Sequence Diversity



Forensic STR Sequence Diversity

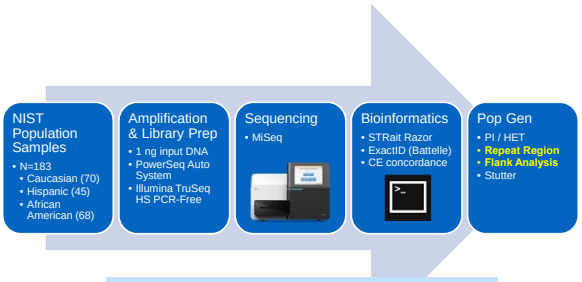


Forensic STR Sequence Diversity



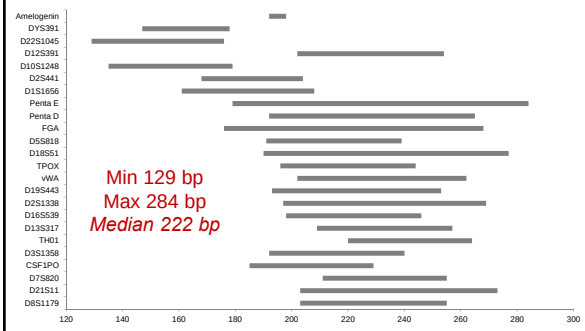
Sequence-Based Heterozygote or Isoallele: A locus that appears homozygous in length-based measurements (such as CE), but is heterozygous by sequence

Forensic STR Sequence Diversity



600 cycle (v3 2 x 300)
Plate 1 avg. coverage ≈ 3,962x (6.7 Gb total generated)
Plate 2 avg. coverage ≈ 10,752x (16 Gb total generated)

PowerSeq Auto System



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

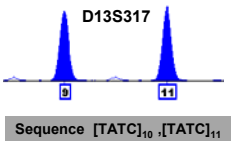
24 loci x 183 samples = 4392 loci evaluated

➢ 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 $[TATC]_{10}$,$[TATC]_{11}$

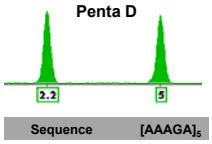
← Repeat Region | NGS Recognition Region | 4 bp Deletion | CE Primer Binding Site→

```
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

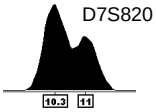
←CE Primer Binding Site | Repeat Region →

```
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 $[GATA]_{11}$ $[GATA]_{11}$

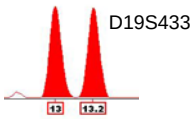
REPEAT REGION | RECOGNITION SITE | 1 bp DELETION

```
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 RECOGNITION REGION

REPEAT REGION | RECOGNITION REGION

```
AAGG AAGG AAGG AAGG AAGG AAGG AAGG AGAGAGGAAGAAAAGAGAGAAGATTTTTATTCGGGT
AAGG AAGG AAGG AAGG AAGG AAGG AAGG AGAG--GAAGAAAAGAGAGAAGATTTTTATTCGGGT
```

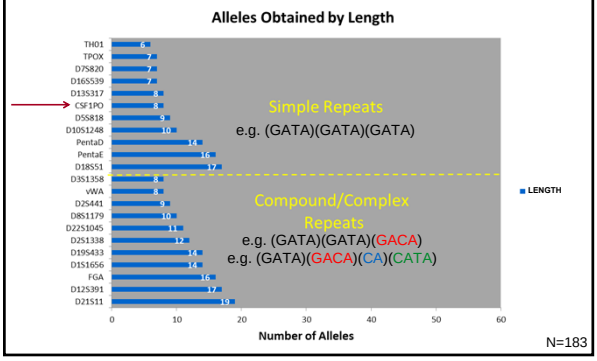
Bioinformatic Null Allele: A true allele that is present within the raw sequence data but is not detected by the bioinformatic pipeline

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 the STR allele?
  - Defined by recognition sites adjacent to the flanking region?
  - Defined by PCR primers?
  - Some set of core recognition sites?

These are all to maintain back compatibility with existing technology (CE)

Forensic STR Sequence Diversity



N=183

Forensic STR Sequence Diversity

Additional Alleles by Sequence

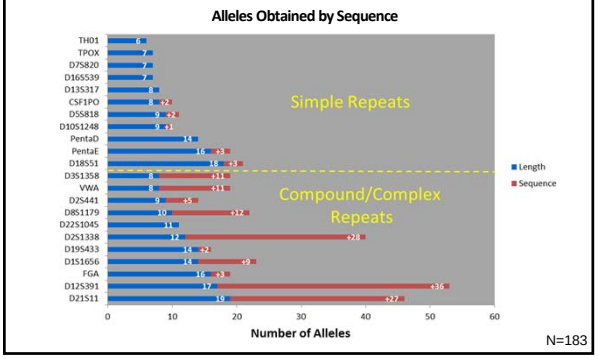
CSF1PO

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



N=183

Manuscript submitted to FSIG

Title:

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

Authors and Affiliatic

Katherine Butler Getti ano[3], Christine H. Baker[3], Brian A. You

[1]U.S. National Institu

D2S441	10	[TCTA]10
	10	[TCTA]8[TCTG][TCTA]
	11	[TCTA]11
	11	[TCTA]9[TCTG][TCTA]
	11.3	[TCTA]3[TCA][TCTA]8
	11.3	[TCTA]4[TCA][TCTA]7
	12	[TCTA]10[TCTG][TCTA]
	12	[TCTA]12
	12.3	[TCTA]4[TCA][TCTA]8
	13	[TCTA]10[TTTA][TCTA]2
	13	[TCTA]13
	14	[TCTA]11[TTTA][TCTA]2
	14.3	[TCTA]3[TCA][TCTA]11
	15	[TCTA]12[TTTA][TCTA]2

- Repeat region sequences

Flanking Region Variation

Moving out past the repeat regions…

Recognition Site-Based Informatics for STRs

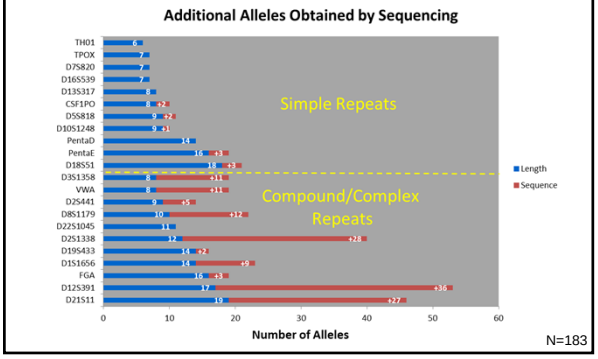
Moving recognition sites out will capture information within the flanking regions

PCR Primers

STR Repeat Region

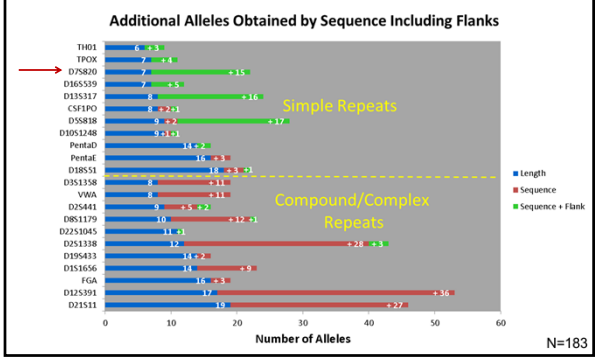
Recognition site (~10 nt)

Forensic STR Sequence Diversity

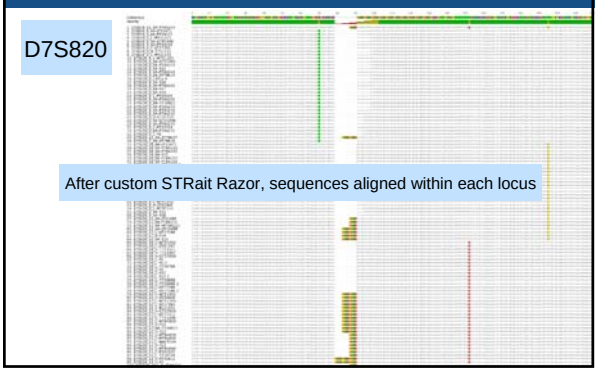


N=183

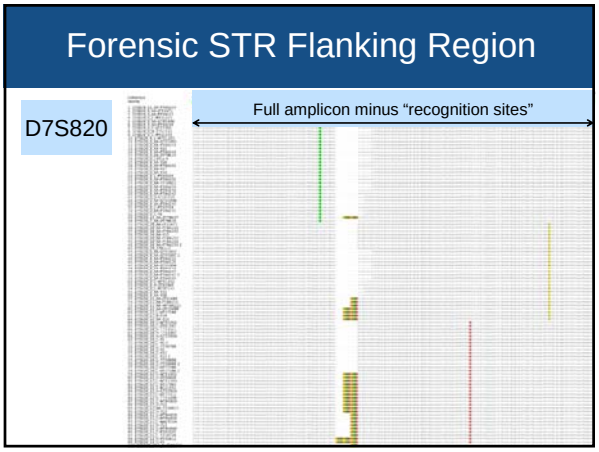
Forensic STR Sequence Diversity



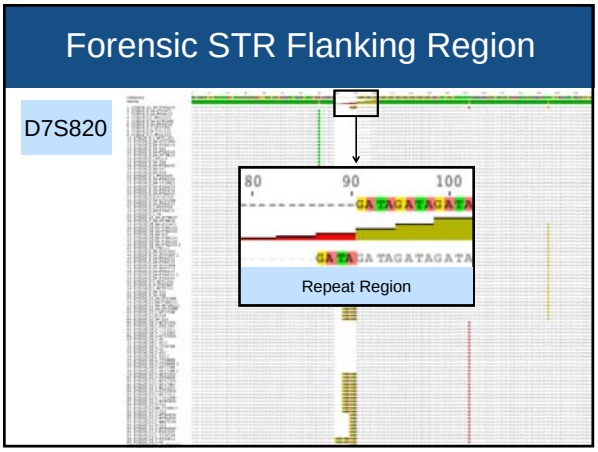
Forensic STR Flanking Region



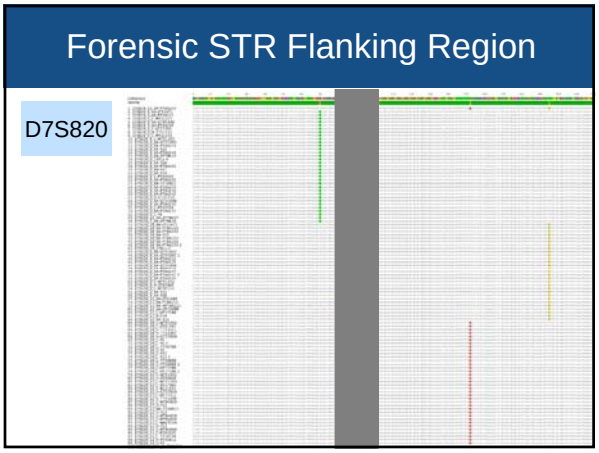
Forensic STR Flanking Region



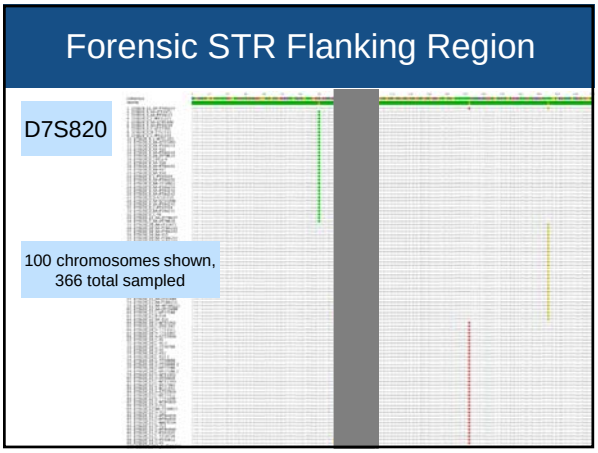
Forensic STR Flanking Region



Forensic STR Flanking Region

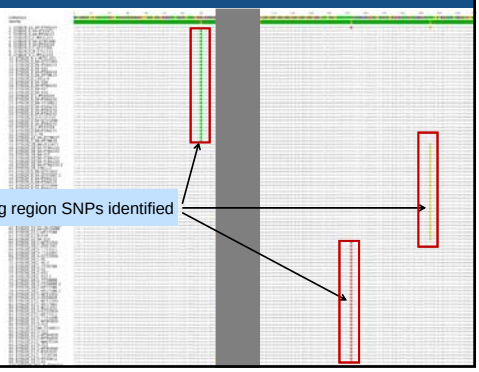


Forensic STR Flanking Region



Forensic STR Flanking Region

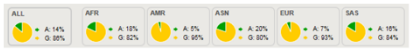
D7S820



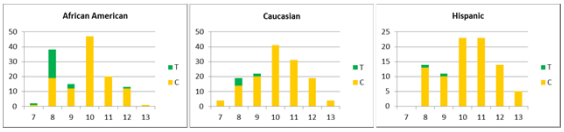
Forensic STR Flanking Region

D7S820

Distribution of *rs16887642* by Population in 1000 Genomes Project



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

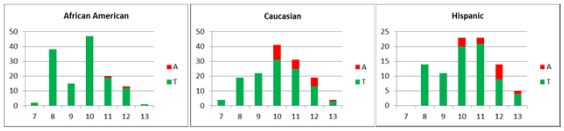


Forensic STR Flanking Region

D7S820

Distribution of *rs7789995* by Population in 1000 Genomes Project

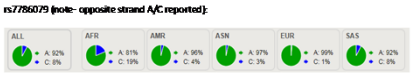
Distribution of *rs7789995* by Population / Allele in N=183



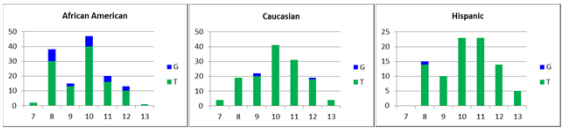
Forensic STR Flanking Region

D7S820

Distribution of *rs7786079* by Population in 1000 Genomes Project



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



Forensic STR Flanking Region

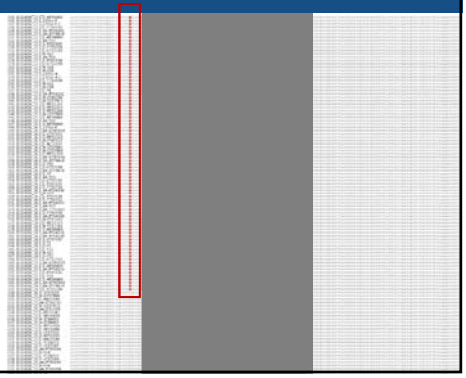
D7S820

Allele	Sequence	rs16887642-rs7789995-rs7786079
7	[GATA]7	C-T-T
8	[GATA]8	C-T-T
8	[GATA]8	C-T-G
8.0	[GATA]8.0	C-T-T
8.1	[GATA]8.1	C-T-T
9.2	[GATA]9.2	C-T-G
9.3	[GATA]9.3	C-T-T
9	[GATA]9	T-T-T
10	[GATA]10	C-T-N
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

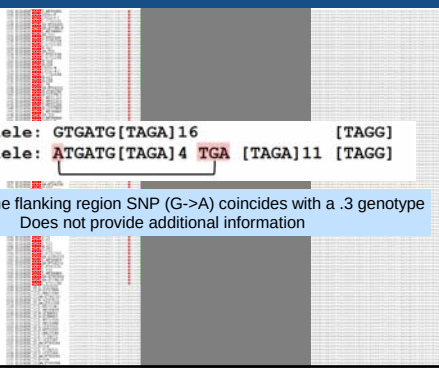
Forensic STR Flanking Region

D1S1656



Forensic STR Flanking Region

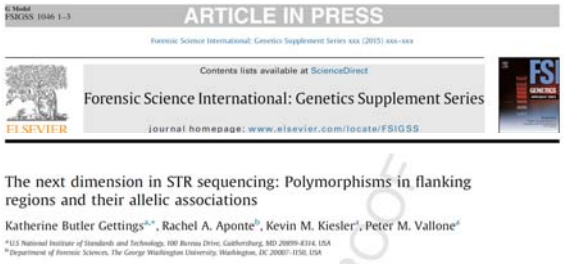
D1S1656



```
17   Allele: GTGATG[TAGA]16          [TAGG]
16.3 Allele: ATGATG[TAGA]4 TGA [TAGA]11 [TAGG]
```

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

Forensic STR Flanking Region



ARTICLE IN PRESS

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

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

The next dimension in STR sequencing: Polymorphisms in flanking regions and their allelic associations

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

Forensic STR Sequence Diversity
Comments

- 22 autosomal forensic STR loci were sequenced for a set of 183 samples (U.S. sample groups)
  - Sequence variation within the STR region and surrounding flanking regions were characterized
- Loci with compound and complex repeat motifs contained the majority of 'additional' information by NGS
  - Sequencing repeat region offers significant gains for some loci – (complex/compound)
  - Extending analysis to the flanking regions offers additional gains – (simple repeats)

NIST Support for NGS Research
FSI Genetics review article

Contents lists available at ScienceDirect

Forensic Science International: Genetics

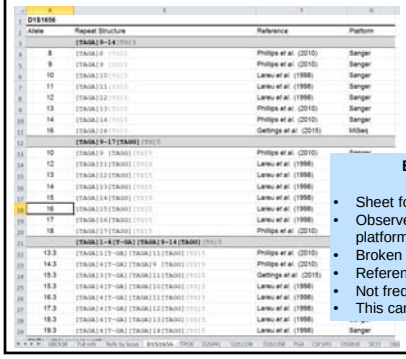
journal homepage: www.elsevier.com/locate/fsig

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



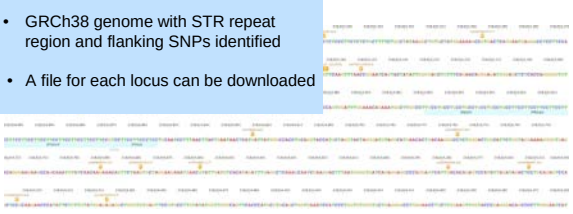
Excel Workbook

- 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

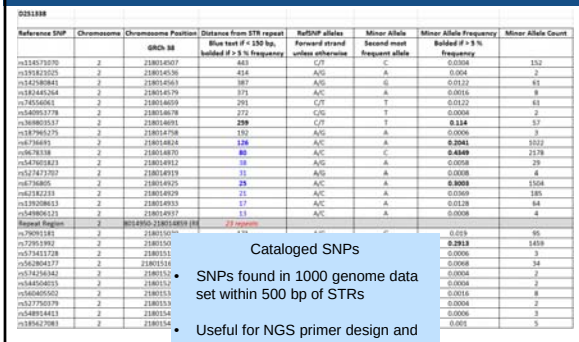
NIST Support for NGS Research

Annotations

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



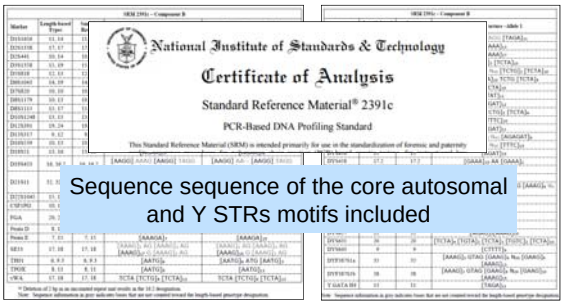
NIST Support for NGS Research



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

Considerations for nomenclature, storage, searching, and reporting

How do we communicate and categorize this additional information?

- General thoughts
- Recent paper by Gelardi et al.
- ISFG nomenclature panel discussion

Adopting sequence-based genotype information
What are our concerns?

- Take advantage of the additional information provided by sequencing the STR allele (and flanking regions)
- Maintain back compatibility to existing databases
  - Genotyping by length-based CE methods
- Report and represent data in a consistent manner
- Application of sequence data
  - **Storage, searching, and reporting**

Nomenclature

- Important so that all labs 'speak the same language'
  - The established nomenclature rules for the core STR loci can be expanded upon if needed
- Designate sequenced alleles in an internationally accepted format
- How will bracket notation fit into our plans?
- Do we need to define amplicon boundaries?
  - Core recognition sites (pros/cons)
- How will we designate sequence that flanks the STR repeat? (SNPs, indels)

D8S1179	
Allele	Repeat Structure
	[TCTA]7-14
7	[TCTA]7
8	[TCTA]8
9	[TCTA]9
10	[TCTA]10
11	[TCTA]11
12	[TCTA]12
13	[TCTA]13
14	[TCTA]14
	[TCTA]1[TCTG]1[TCTA]10-14
12	[TCTA]1[TCTG]1[TCTA]10
13	[TCTA]1[TCTG]1[TCTA]11
14	[TCTA]1[TCTG]1[TCTA]12
16	[TCTA]1[TCTG]1[TCTA]14
	[TCTA]2[TCTG]1[TCTA]9-15
11	[TCTA]2[TCTG]1[TCTA]8
12	[TCTA]2[TCTG]1[TCTA]9
13	[TCTA]2[TCTG]1[TCTA]10
14	[TCTA]2[TCTG]1[TCTA]11
15	[TCTA]2[TCTG]1[TCTA]12
16	[TCTA]2[TCTG]1[TCTA]13
17	[TCTA]2[TCTG]1[TCTA]14
18	[TCTA]2[TCTG]1[TCTA]15
	[TCTA]2[TCTG]2[TCTA]11-15
15	[TCTA]2[TCTG]2[TCTA]11
16	[TCTA]2[TCTG]2[TCTA]12
17	[TCTA]2[TCTG]2[TCTA]13
19	[TCTA]2[TCTG]2[TCTA]15

D8S1179 Example

Four 'flavors' of the motif are observed (so far…)
Tempting to just call them A, B, C, D, etc

14 [TCTA]14 (A)

14 [TCTA]1 [TCTG]1 [TCTA]12 (B)

14 [TCTA]2 [TCTG]1[TCTA]11 (C)

What if/when a new motif and/or SNP is detected within one of the motifs?

D2S1338	
Allele	Repeat Structure
	[TGCC]4-8[TTCC]6-19
10	[TGCC]4[TTCC]6
12	[TGCC]4[TTCC]8
12	[TGCC]6[TTCC]6
13	[TGCC]6[TTCC]7
15	[TGCC]5[TTCC]10
16	[TGCC]4[TTCC]12
16	[TGCC]6[TTCC]10
17	[TGCC]4[TTCC]13
17	[TGCC]5[TTCC]12
17	[TGCC]6[TTCC]11
18	[TGCC]3[TTCC]15
18	[TGCC]6[TTCC]12
18	[TGCC]7[TTCC]11
19	[TGCC]6[TTCC]13
19	[TGCC]7[TTCC]12
19	[TGCC]8[TTCC]11
20	[TGCC]6[TTCC]14
20	[TGCC]7[TTCC]13
20	[TGCC]8[TTCC]12
21	[TGCC]7[TTCC]14
21	[TGCC]8[TTCC]13
22	[TGCC]6[TTCC]16
22	[TGCC]7[TTCC]15
22	[TGCC]8[TTCC]14
22	[TGCC]9[TTCC]13
23	[TGCC]7[TTCC]16
23	[TGCC]9[TTCC]14
26	[TGCC]7[TTCC]19
	[TGCC]6-8[TTCC]10-17([GTCC][TTCC]2
19	[TGCC]6[TTCC]10([GTCC][TTCC]2
20	[TGCC]7[TTCC]10([GTCC][TTCC]2
21	[TGCC]6[TTCC]12([GTCC][TTCC]2
21	[TGCC]7[TTCC]11([GTCC][TTCC]2
22	[TGCC]6[TTCC]13([GTCC][TTCC]2
22	[TGCC]7[TTCC]12([GTCC][TTCC]2
23	[TGCC]6[TTCC]14([GTCC][TTCC]2
23	[TGCC]7[TTCC]13([GTCC][TTCC]2
24	[TGCC]6[TTCC]15([GTCC][TTCC]2

A

B

The A B designation for D2S1338 would be problematic due to multiple alleles with the same motif

17 [TGCC]4[TTCC]13

17 [TGCC]5[TTCC]12

17 [TGCC]6[TTCC]11

Recent Paper

Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles

Chiara Gelardi, Eszter Rockenbauer, Sigrun Dalsgaard, Claus Børsting, Niels Morling

Gelardi nomenclature format:
Locus | Length | Sequence | Flank Variants

Recent Paper

Table 1
Novel D3S1358, D12S391 and D21S11 alleles.

CE allele name	SGS allele name	Count	Frequency
14	D3S1358[14]TCTA[1]TCTG[3]TCTA[10]	1	0.0025
14	D3S1358[14]TCTA[1]TCTG[1]TCTA[12]	1	0.0025
20	D3S1358[20]TCTA[1]TCTG[4]TCTA[15]	1	0.0025
18	D12S391[18]AGAT[11]AGAC[7]	3	0.0076
18	D12S391[18]AGAT[12]AGAC[5]AGAT[1]	4	0.0102
19	D12S391[19]AGAT[11]AGAC[7]AGAT[1]	1	0.0025
19	D12S391[19]AGAT[11]AGAC[8]	6	0.0152
20	D12S391[20]AGGT[1]AGAT[10]AGAC[9]	2	0.0051
21	D12S391[21]AGAT[13]GGAC[1]AGAC[7]	1	0.0025
21	D12S391[21]AGGT[1]AGAT[11]AGAC[9]	1	0.0025

Recent Paper

Table 1
Novel D3S1358, D12S391 and D21S11 alleles.

Gelardi nomenclature example
Locus | length | sequence | flank variants

D3S1358	[14]	TCTA[1]TCTG[3]TCTA[10]
D12S391	[20]	AGGT[1]AGAT[10]AGAC[9]
D21S11	[32.2]	TCTA[6]TCTG[6]itvsTCTA[11]TA[1]TCTA[1]

itvs: TCTA[3]TA[1]TCTA[3]TCA[1]TCTA[2]TCCA[1]TA[1] (43 bp)

21	D12S391[21]AGAT[13]GGAC[1]AGAC[7]	1	0.0025
21	D12S391[21]AGGT[1]AGAT[11]AGAC[9]	1	0.0025

Reporting

- Right now we report: D2S1338 [20,22]
- Is the full sequence string too much in a report?
  - Just the STR motif or include flanking regions (length limits?)
- Should the sequence string be condensed down to something easier to understand in a report?
  - Bracket notation? (example next slide)
- An unique code might be fine for storage and searching, but not intuitive for a report (lookup key)
  - Example: gatagatagatagatagatagatagata = 'TK421'

Reporting

D2S1338 [20,22] example

String

20 TGCCTGCCTGCCTGCCTGCCTGCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCGTCCTTCCTTCC

22 TGCCTGCCTGCCTGCCTGCCTGCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCGTCCTTCCTTCC

Bracket

20 $[TGCC]_7[TTCC]_{10}[GTCC][TTCC]_2$

22 $[TGCC]_7[TTCC]_{12}[GTCC][TTCC]_2$

Will need to follow some set of rules for string to bracket conversion performed by computer

Searching

- Sequence string for each allele? Assign a unique code?
  - Don't think bracket notation will be useful for searching
- Encoding each string for ease of searching
  - Example: gatagatagatagatagatagatagata = 'TK421'
- Do we need agreed upon boundaries for one-to one searching (core amplicon length)
  - The issue of the same allele, but varying amplicon lengths
- Search by length genotype then sequence string?
  - 11 to 11 match; then compare sequence data?

Issues of searching algorithm speed and data storage

Storage

- Probably do not want Mb or Gb of data for each sample
  - FASTQ -> trimmed data -> coverage per locus -> final sequence
- The reviewed, interpreted, and finalized genotype file does not have to be that big
  - Final genotypes can be represented an agreed upon reporting format
- What will be our needs to retain raw data?
  - .fsa/.HID versus FASTQ; ability to reanalyze raw data

Forensic STR Sequence Nomenclature

- Continue to generate more STR sequence data
  - Catalog "new" alleles and frequencies
  - Discuss optimal way to categorize sequence variation
    - e.g. systematic or first reported
- SWGDAM NGS Working Group
- **ISFG committee for STR nomenclature**

Panel discussion
Walther Parson, Chris Phillips, Katherine Gettings, Peter de Knijff, David Ballard, Bruce Budowle, Niels Morling
Sequencing based STR nomenclature

26th ISFG Krakow Poland

September 2, 2015

Panel Discussion

Christophe Van Neste, Ghent University
FLAD: Forensic Loci Allele Database

- A minimal nomenclature
  - Not reporting the full sequence
  - An identifier can be referenced to the full sequence
  - **FL1A12 ↔ AGATAGATAGATAGATAGATAGAT**
  - https://forensic.ugent.be/FLAD
- A custom solution
  - Avoiding GenBank refseqs
  - dbSNP rs#
- Identifier stored in FLAD
  - Need QC on data

Christophe Van Neste, Ghent University
FLAD: Forensic Loci Allele Database

- A minimal nomenclature
  - Not reporting the full sequence
  - Article in Press
  - Forensic Loci Allele Database (FLAD): automatically generated, permanent identifiers for sequenced forensic alleles
  - Christophe Van Neste
  - Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium
  - dbSNP rs#
- Identifier stored in FLAD
  - Need QC on data

Christophe Van Neste, Ghent University
FLAD: Forensic Loci Allele Database

FA[HEX] → FA1A, FA23 (FA: Forensic Allele + random hex number)

↓ Adding locus information

FL[D]A[HEX] → FL3A1A, FL1A23 (FLA: Forensic Locus … Allele …)

↓ Adding primerset information

FL[D[.D]]A[HEX] → FL3.1A1A, FL3.2A3C

↓ Adding allele annotation

FL[D][X|A|P][HEX] → FL3X1A, FL1P23

(X = unvalidated allele, P = validated, but rare allele, A = common population allele)

Christophe Van Neste, Ghent University
FLAD: Forensic Loci Allele Database

FA[HEX] → F

↓

FL[D]A[HEX]

↓

FL[D[.D]]A[HE

↓

FL[D][X|A|P]

Loci	Genotype
Amelogenin [FL1]	X[A01], Y[A03]
D3S1358 [FL3]	16[P0A]
THO1 [FL2]	6[A3B], 9.3[A20]
D21S11 [FL11]	30[A05], 30.2[AFF]
D18S51 [FL13]	15[A10]
D8S1179 [FL6]	14[A5D]
TPOX [FL5]	8[A10], 12[A02]
FGA [FL8]	23[A41]

(X = unvalidated allele, P = validated, but rare allele, A = common population allele)

Kristiaan van der Gaag, Rick de Leeuw & Peter de Knijff, Department of Human Genetics Leiden University Medical Center

Fully descriptive and relative to a reference genome

D13S317[CE12]-Chr13-GRCh37-g.82.722.160:82.722.223-TATC[13]AATC[1]ATCT[3]-g.x.136G>A

full name

D13S317[CE12]-Chr13-GRCh37-g.82.722.160:82.722.223-TATC[13]AATC[1]ATCT[3]-g.x.136G>A

Locus and CE allele

D13S317[CE12]-**Chr13-GRCh37**-g.82.722.160:82.722.223-TATC[13]AATC[1]ATCT[3]-g.x.136G>A

CHR and human genome reference version

D13S317[CE12]-Chr13-GRCh37-**g.82.722.160:82.722.223**-TATC[13]AATC[1]ATCT[3]-g.x.136G>A

STR repeat region coordinates of reference allele

D13S317[CE12]-Chr13-GRCh37-g.82.722.160:82.722.223-**TATC[13]AATC[1]ATCT[3]**-g.x.136G>A

Full description of STR motif

D13S317[CE12]-Chr13-GRCh37-g.82.722.160:82.722.223-TATC[13]AATC[1]ATCT[3]-**g.x.136G>A**

Genome coordinate of derived SNP

Sascha Willuweit

Department of Forensic Genetics, Institute of Legal Medicine and Forensic Sciences, Charite´ – Universitatsmedizin, Berlin, Germany

- Reporting full sequences is not feasible (yet)
- Long lists of sequences are correct, but not practical
- Need legacy compatibility (with CE data)
- A Nomenclature Authority (NA) would assign a code/type to a new allele
  - Centralized server
  - Reliable, secure
  - ISFG organized/maintained

Sascha Willuweit

Department of Forensic Genetics, Institute of Legal Medicine and Forensic Sciences, Charite´ – Universitatsmedizin, Berlin, Germany

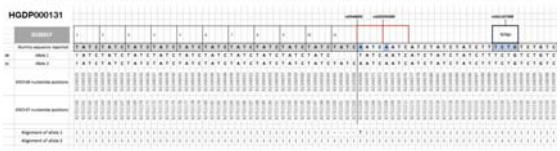
- Submit sequenced allele to NA
  - If already observed (at least twice) a fixed value is returned
  - If not, a temporary identifier is assigned until confirmed
- Example, submit a sequence of a 14 allele
  - Already observed, returned the value '14*f'
  - New allele, assigned '14*68'
  - When observed again – this is converted into '14*g' (fixed value)

Chris Phillips

Univ of Santiago de Compostela

Is a 'synthesis' possible and can we simplify sequences still further?

- Define a sequence only by its differences from the reference genome
- Upload raw sequences as, say, Excel lists which are then processed by the NA



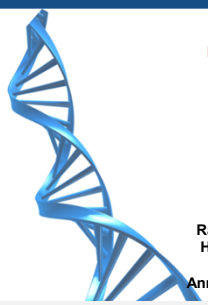
D13S317 CE 10,11 (RR only)	D13S317 CE 10,11 (amplicon)	D13S317 CE 11,13 (extra repeats)
A1	A1	A1
13:82722200-82722203 –	13:82722136 **G**	A2
13:82722204 **T**	13:82722200-82722203 –	13:82722203 **+1-8 TATC**
A2	13:82722204 **T**	13:82722203 **+2 G**
	A2	

Thoughts

- Pros/cons of using a code versus full descriptor
  - How difficult is it to deal with full sequence strings?
- References to genome builds can/will change
  - How to deal with this issue?
- Commercial kits/software will have to be informed
- QC of sequence submitted to nomenclature assigning websites (if they exist) will be important
  - See EMPOP

Lots of aspects to think about---stay tuned

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

ISFG Nomenclature Panel Members
Peter de Knijff
Sascha Willuweit
Christophe Van Neste
Chris Phillips

Contact Information
peter.vallone@nist.gov