STR Nomenclature Meeting

April 11-12, 2019 London

5’ to 3’:
Walther Parson, Lisa Borsuk, Peter Schneider, Brian Young, Rebecca Just, Jodi Irwin, David Ballard, Sascha Willuweit, Cydne Holt, Chris Phillips, Jonathan King, Tunde Huszar, Peter Gill, Christian Sell, Kris Van der Gaag, Laurence Devesse, Claus Borsting, Doug Hares, Katherine Gettings, Rob Lagace, Jerry Hoogenboom, Martin Bodner, Peter deKnijff, Sebastian Ganschow, Pedro Barrio, Teresa Gross

STRAND *working group*
align | name | define

D.Ballard:KCL,UK | M.Bodner&W.Parson:GMI,Austria | L.Borsuk&K.Gettings:NIST,US | J.King:UNT,US | C.Phillips:USC,Spain

STRAND working group
align | name | define

Agenda: April 11, 2019

8:30am	-	8:45am	Arrival
8:45am	-	9:30am	Welcome and Opening Remarks - STRAND WG
9:30am	-	10:00am	Tunde Huszar
10:00am	-	10:15am	Coffee Break
10:15am	-	10:45am	Pedro Barrio
10:45am	-	11:15am	Claus Borsting
11:15am	-	11:45am	Brian Young
11:45am	-	12:00pm	Discussion
12:00pm	-	1:30pm	Lunch on your own
1:30pm	-	2:00pm	Peter deKnijff
2:00pm	-	2:30pm	Kris Van der Gaag / Jerry Hoogenboom
2:30pm	-	3:00pm	Sascha Willuweit
3:00pm	-	3:15pm	Coffee Break
3:15pm	-	3:45pm	Rebecca Just
3:45pm	-	4:15pm	Peter Gill
4:15pm	-	4:45pm	Sebastian Ganschow
4:45pm	-	5:00pm	Discussion and Day 1 Closing Remarks

Agenda: April 12, 2019

8:30am	-	8:45am	Arrival
8:45am	-	9:00am	Welcome and Opening Remarks
9:00am	-	11:00am	STRAND WG Facilitated Discussion

- Reference genomes, existing databases
- Quality control
- Bioinformatics
- Implementation

11:00am	-	11:15am	Coffee Break
11:15am	-	12:00pm	Summary, Path Forward, and Closing Remarks

D.Ballard:KCL,UK | M.Bodner&W.Parson:GMI,Austria | L.Borsuk&K.Gettings:NIST,US | J.King:UNT,US | C.Phillips:USC,Spain

STRAND working group
align | name | define

STR sequence nomenclature

NGS of STRs: Nomenclature Panel at ISFG Conference Krakow 2015

NGS of STRs: Considerations of the ISFG (2016)

Forensic Science International: Genetics 22 (2016) 54–63

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements

Walther Parson[a,b,*], David Ballard[c], Bruce Budowle[d,e], John M. Butler[f], Katherine B. Gettings[f], Peter Gill[g,h], Leonor Gusmão[i,j,k], Douglas R. Hares[l], Jodi A. Irwin[l], Jonathan L. King[d], Peter de Knijff[m], Niels Morling[n], Mechthild Prinz[o], Peter M. Schneider[p], Christophe Van Neste[q], Sascha Willuweit[r], Christopher Phillips[s]

the full sequence (sequence string),
the alignment of sequences relative to a reference sequence
the annotation of alleles

3.1. Most Cited Articles, 2018 (Published IF Window 2016-2017)

Citations	Citations (lifetime)	Article Title	Authors	Publication Year	Document Type
32	66	Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements	Parson W.,Ballard D.,Budowle B.,Butler J.M.,Gettings K.B.,Gill P.,Gusmao L.,Hares D.R.,Irwin J.A.,King J.L.,Knijff P.D.,Morling N.,Prinz M.,Schneider P.M.,Neste C.V.,Willuweit S.,Phillips C.	2016	Article
29	81	Evaluation of the Illumina® Beta Version ForenSeq™ DNA Signature Prep Kit for use in genetic profiling	Churchill J.D.,Schmedes S.E.,King J.L.,Budowle B.	2016	Article
27	55	Sequence variation of 22 autosomal STR loci detected by next generation sequencing	Gettings K.B.,Kiesler K.M.,Faith S.A.,Montano E.,Baker C.H.,Young B.A.,Guerrieri R.A.,Vallone P.M.	2016	Article
26	39	Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories	Jager A.C.,Alvarez M.L.,Davis C.P.,Guzman E.,Han Y.,Way L.,Walichiewicz P.,Silva D.,Pham N.,Caves G.,Bruand J.,Schlesinger F.,Pond S.J.K.,Varlaro J.,Stephens K.M.,Holt C.L.	2017	Article
23	35	Massively parallel sequencing of short tandem repeats - Population data and mixture analysis results for the PowerSeq™ system	Van Der Gaag K.J.,De Leeuw R.H.,Hoogenboom J.,Patel J.,Storts D.R.,Laros J.F.J.,De Knijff P.	2016	Article

Quality Control Concept for Forensic Autosomal STRs (STRidER, 2016)

Forensic Science International: Genetics 24 (2016) 97–102

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER)

Martin Bodner[a], Ingo Bastisch[b], John M. Butler[c], Rolf Fimmers[d], Peter Gill[e,f], Leonor Gusmão[g,h,i], Niels Morling[j], Christopher Phillips[k], Mechthild Prinz[l], Peter M. Schneider[m], Walther Parson[a,n,*]

Forensic Science International: Genetics 31 (2017) 111–117

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci

Katherine Butler Gettings[a,*], Lisa A. Borsuk[a], David Ballard[b], Martin Bodner[c], Bruce Budowle[d,e], Laurence Devesse[b], Jonathan King[d], Walther Parson[c,f], Christopher Phillips[g], Peter M. Vallone[a]

NCBI BioProject—STRseq and **STRidER**

Collaboration in QC and exchange of data

“The devil’s in the detail”: Release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide (2018)

C. Phillips[a,*], K. Butler Gettings[b], J.L. King[c], D. Ballard[d], M. Bodner[e], L. Borsuk[b], W. Parson[e,f]

[a] *Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain*
[b] *National Institute of Standards and Technology, Biomolecular Measurement Division, Gaithersburg, MD, USA*
[c] *Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, Fort Worth, TX, USA*
[d] *King’s Forensics, King’s College London, Franklin-Wilkins Building, London, UK*
[e] *Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria*
[f] *Forensic Science Program, The Pennsylvania State University, University Park, PA, USA, USA*

+ **revised** STR Sequence Guide as **dynamic** document at STRidER

HOME | QUERY | BATCH QUERY | ABOUT | FREQUENCIES | FORMULAE | QUALITY CONTROL | STR SEQ NOMENCLATURE

STR Sequence Nomenclature

The **‘Forensic STR Sequence Structure’ file** is an updated set of forensic STR sequences that was originally published as *Supplementary File S1* in the article:

The most recent version of this permanently curated and updated Forensic STR sequence structure file containing updated information is available for download here. The updates since the last version are reported in a change log contained in the file. To receive information on new releases of the Forensic STR sequence structure file and to stay updated about STRidER, register here for the STRidER newsletter.

STRAND working group

align | name | define

Goals

Continue collection of STR sequence information to understand variation

Update STRSeq @NCBI and STR sequence guide @STRidER

Harmonize efforts to develop a common STR nomenclature system

Propose a common STR nomenclature system to the community through the ISFG

This is a nomenclature panel

This is **NOT** a panel to focus on technical and analytical problems

STRAND *working group*

align | name | define

Katherine Gettings - NIST Applied Genetics Group

- NIST population sample sequencing
- Reference materials for STR sequencing

STRAND *working group*

align | name | define

- NIST population sample sequencing

Forensic Science International: Genetics 37 (2018) 106–

Contents lists available at ScienceDirect

Forensic Science International: G

journal homepage: www.elsevier.com/locate/

Research paper

Sequence-based U.S. population data for 27 autosomal S

Katherine Butler Gettings*, Lisa A. Borsuk, Carolyn R. Steffen, Kevin M. K

U.S. National Institute of Standards and Technology, Biomolecular Measurement Division, 100 Bureau Drive, Gaithersburg, MD

ARTICLE INFO

Keywords:
STR
Allele frequency
Sequence

ABSTRACT

This manuscript reports Short Tandem Repeat (STR) se
27 autosomal STR loci: D1S1656, TPOX, D2S441, D
D6S1043, D7S820, D8S1179, D9S1122, D10S1248, T
D17S1301, D18S51, D19S433, D20S482, D21S11, Penta
bioinformatic pipelines and all samples have been evalu
analysis at all loci. Each reported sequence includes hi
according to the most recent guidance of the Internatio
accession numbers are reported for each sequence, and a
(https://www.ncbi.nlm.nih.gov/bioproject/380127). T
increase in heterozygosity across populations (approxim
increase in heterozygosity from 10 to 5 percentage po
D8S1179, D21S11, D5S818, D12S391, and D2S441. Th
centage point increase in average heterozygosity. Discu
traditional CE results, such as informing stutter mode
siderations for population sampling strategies in light
sequence-based STR loci. This NIST 1036 data set is exp
forensic casework by providing high-confidence sequenc
are already the basis for population statistics in many

1. Introduction

In forensic casework, Short Tandem Repeat (STR) allele frequency data is used to calculate statistical weight when a person of interest cannot be excluded as a possible contributor of genetic material recovered from an item of evidence. This statistical weight should be derived from the same level of information which was used for comparison; therefore, implementation of STR sequencing into forensic casework necessitates the development of appropriate allele frequency data sets.

Several recent publications have reported sequence-based allele frequency data for autosomal STR loci [1–6]. This manuscript reports high-confidence autosomal STR sequence-based allele frequencies for N = 1036 across 27 autosomal STR loci: D1S1656, TPOX, D2S441, D2S1338, D3S1358, D4S2408, FGA, D5S818, CSF1PO, D6S1043, D7S820, D8S1179, D9S1122, D10S1248, TH01, vWA, D12S391, D13S317, Penta E, D16S539, D17S1301, D18S51, D19S433, D20S482, D21S11, Penta D, a
confidence to this d
ment; analysis with
quality flanking seq
sample at every locu
firmation of all null
same length but diffe

The preceding "
allele frequencies for
by sequence in this
based NIST 1036 ar
F13B, FESFPS, LPL, P
data for NIST 1036
tional loci are reporte
original data set: D4S
these four loci was pr
and the remaining s

* Corresponding author at: National Institute of Standards and Technology, Biomolecular Measurement Div
USA.
E-mail addresses: katherine.gettings@nist.gov (K.B. Gettings), lisa.borsuk@nist.gov (L.A. Borsuk), becky.
kevin.kiesler@nist.gov (K.M. Kiesler), peter.vallone@nist.gov (P.M. Vallone).

https://doi.org/10.1016/j.fsigen.2018.07.013
Received 26 April 2018; Received in revised form 26 June 2018; Accepted 16 July 2018
Available online 19 July 2018
1872-4973/

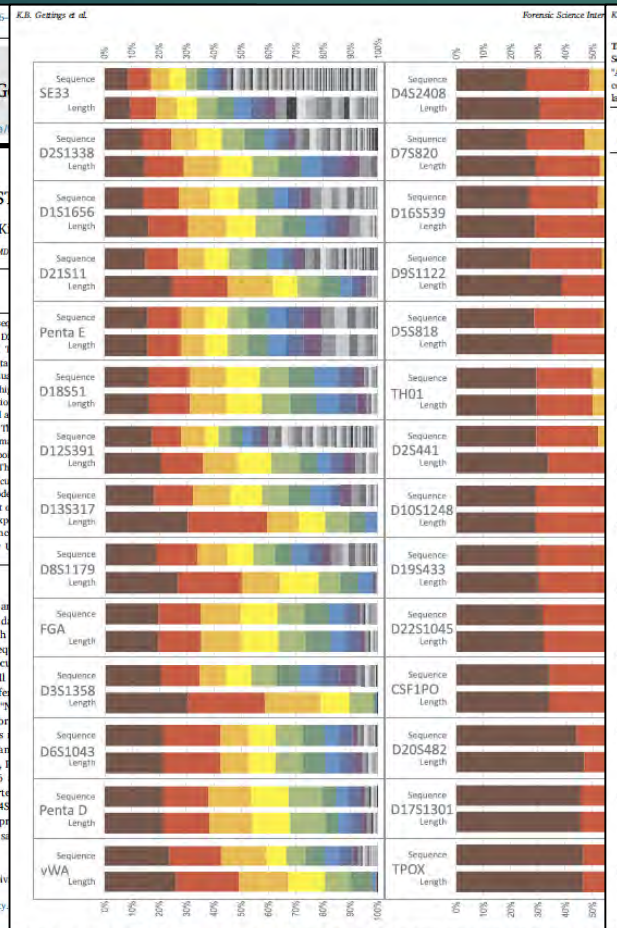


Fig. 2. Across-population allele frequency distribution per locus, by sequence and by length in N = 1036. Loci are sorted in as
frequency of the most common allele at each locus (first column top to bottom followed by second column top to bottom). The f
colored to facilitate comparisons within and across loci, with any remaining alleles shown in grayscale. Sequence data for thes
reported in [9].

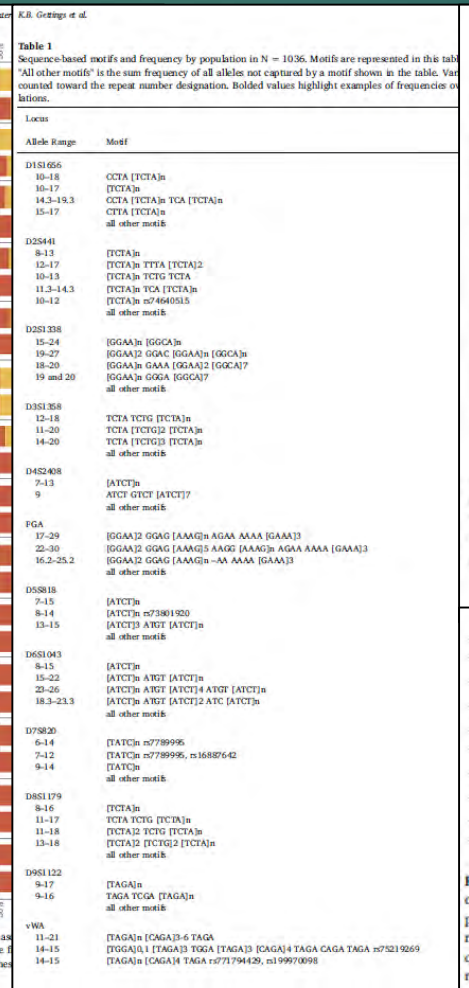


Table 1
Sequence-based motifs and frequency by population in N = 1036. Motifs are represented in this tabl
"All other motifs" is the sum frequency of all alleles not captured by a motif shown in the table. Var
counted toward the repeat number designation. Bolded values highlight examples of frequencies o
lations.

Locus / Allele Range	Motif
D1S1656	
10–18	CCTA [TCTA]n
10–17	[TCTA]n
14.3–19.3	CCTA [TCTA]n TCA [TCTA]n
15–17	CTTA [TCTA]n
	all other motifs
D2S441	
8–13	[TCTA]n
12–17	[TCTA]n TTTA [TCTA]2
10–13	[TCTA]n TCTG TCTA
11.3–14.3	[TCTA]n TCA [TCTA]n
10–12	[TCTA]n rs74640515
	all other motifs
D2S1338	
15–24	[GGAA]n [GGCA]n
19–27	[GGAA]2 GGAC [GGAA]n [GGCA]n
18–20	[GGAA]n GAAA [GGAA]2 [GGCA]7
19 and 20	[GGAA]n GGGA [GGCA]7
	all other motifs
D3S1358	
12–18	TCTA TCTG [TCTA]n
11–20	TCTA [TCTG]2 [TCTA]n
14–20	TCTA [TCTG]3 [TCTA]n
	all other motifs
D4S2408	
7–13	[ATCT]n
9	ATCT GTCT [ATCT]7
	all other motifs
FGA	
17–29	[GGAA]2 GGAG [AAAG]n AGAA AAAA [GAAA]3
22–30	[GGAA]2 GGAG [AAAG]5 AAGG [AAAG]n AGAA AAAA [GAAA]3
16.2–25.2	[GGAA]2 GGAG [AAAG]n –AA AAAA [GAAA]3
	all other motifs
D5S818	
7–15	[ATCT]n
8–14	[ATCT]n rs73801920
13–15	[ATCT]3 ATGT [ATCT]n
	all other motifs
D6S1043	
8–15	[ATCT]n
15–22	[ATCT]n ATGT [ATCT]n
23–26	[ATCT]n ATGT [ATCT]4 ATGT [ATCT]n
18.3–23.3	[ATCT]n ATGT [ATCT]2 ATC [ATCT]n
	all other motifs
D7S820	
6–14	[TATC]n rs7789995
7–12	[TATC]n rs7789995, rs16887642
9–14	[TATC]n
	all other motifs
D8S1179	
8–16	[TCTA]n
11–17	TCTA TCTG [TCTA]n
11–18	[TCTA]2 TCTG [TCTA]n
13–18	[TCTA]2 [TCTG]2 [TCTA]n
	all other motifs
D9S1122	
9–17	[TAGA]n
9–16	TAGA TCGA [TAGA]n
	all other motifs
vWA	
11–21	[TAGA]n [CAGA]3-6 TAGA
14–15	[TGGA]0,1 [TAGA]3 TGGA [TAGA]3 [CAGA]4 TAGA CAGA TAGA rs75219269
14–15	[TAGA]n [CAGA]4 TAGA rs771794429, rs199970098

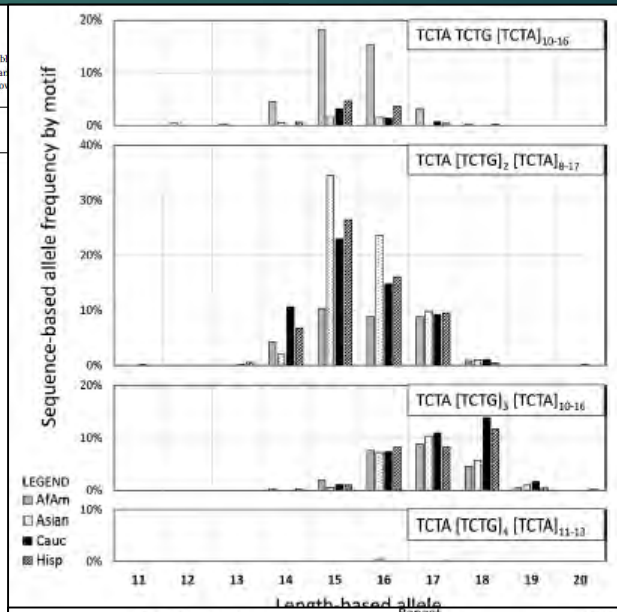
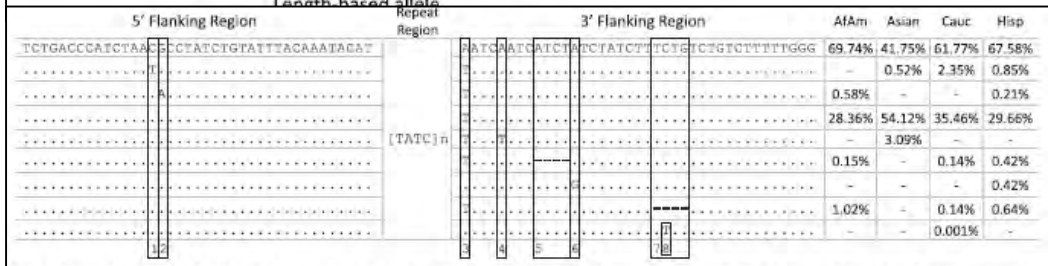


Fig. 3. D3S1358 frequency distribution among the primary motifs by length-based allele and population in N = 1036. The motif is defined as: the first subunit is fixed TCTA, the second subunit is definitive of the motif with TCTG varying from one to four, and the third subunit contains a widely varying number of TCTA repeats. For simplicity, seven additional rare motif alleles present in the data set have been excluded from this figure.



AfAm	Asian	Cauc	Hisp
69.74%	41.75%	61.77%	67.58%
–	0.52%	2.35%	0.85%
0.58%	–	–	0.21%
28.36%	54.12%	35.46%	29.66%
–	3.09%	–	–
0.15%	–	0.14%	0.42%
–	–	–	0.42%
1.02%	–	0.14%	0.64%
–	–	0.001%	–

Fig. 4. D13S317 frequency distribution by population of the nine flanking region motifs identified in N = 1036. The first row of 5′ and 3′ flanking sequence is consistent with GRCh38, and is the most common sequence found in this data set. Dots in subsequent rows represent bases matching the first row. Flanking polymorphisms are identified by numbers one through eight in the bottom row: 1) rs73250432 C > T, 2) rs146621667 G > A, 3) rs9546005 A > T, 4) rs202043589 A > T, 5) rs1442523705 delATCT, 6) ss2137543825 A > G, 7) rs561167308 delTCTG, and 8) rs768323113 C > T. Variation in repeat region length, combined with these flanking region polymorphisms, results in 32 sequence-based alleles at this locus. Three additional D13S317 alleles in this data set result from repeat region sequence variants, each observed once, and have been excluded from this figure.

STRAND *working group*

align | name | define

NIST Public Data Repository 1.1.0-beta

https://doi.org/10.18434/T4/1500024

Public Data Resource

Sequence-based U.S. population data for 27 autosomal STR loci

Contact: Katherine Gettings..

Identifier: doi:10.18434/t4/1500024

Version: 1.0.1... Last modified: 2018-06-14

Description

This information and data are supplemental files associated with: K.B. Gettings, L.A. Borsuk, C.R. Steffen, K.M. Genetics 37 (2018) 106-115. The primary data consists of sequence-based allele frequencies for N=1036 ano D2S1338, D3S1358, D4S2408, FGA, D5S818, CSF1PO, D6S1043, D7S820, D8S1179, D9S1122, D10S1248, TH0 This information is expected to support the implementation of sequence-based STR analysis in forensic appli the 42 sequencing runs performed to generate the allele frequency data (S1 - Run Metrics); coverage per locus Coverage); allele frequency data (S3 - Frequencies); GRCh38 reference coordinates for genomic regions report (S5 - Flank Polymorph); number of alleles, expected and observed heterozygosity, and p-values associated wit disequilibrium (S7 - LD p-values); and pairwise Fst values by population for the 27 auSTR loci (Supp Table 8 - P quality control of the data.

Subject Keywords: STR, forensic, sequence, population, allele frequency

Data Access

These data are public.

Files Click on the file/row in the table below to view more details

Name
NIST1036_auSTR_Seq_SuppFile1.pdf
NIST1036_auSTR_Seq_SuppFile1.pdf.sha256
NIST1036_auSTR_Seq_SuppTables.xlsx
NIST1036_auSTR_Seq_SuppTables.xlsx.sha256

References

This data is referenced in :

https://doi.org/10.1016/j.fsigen.2018.07.013

11 pages of methods, including:

Population, Sample Type	Locus	CE	UAS	SR	Source
African American, Blood	Penta D	2.2,13.4	2.2,14	**2.2,13.4**	1 bp deletion rs536566765
African American, Buccal	D5S818	**7,12***	8,12	**8,12***	Assumed 4 bp deletion outside ForenSeq amplicon
Hispanic, Blood	D7S820	10.3,11	11,11	**10.3,11**	1 bp deletion, rs897512434
Caucasian, Blood	D9S1122	(12),14	12,14	**11.2,14**	2 bp deletion rs754976988, overlaps with CE primer binding site

Table D. Discordance between CE and sequence data observed in N=1036. Bolded genotypes used in allele frequency calculations. *8,12 used in 1036 sequence-based frequencies, 7,12 used in 1036 CE-based frequencies [2].

STRAND *working group*

align | name | define

NIST Public Data Repository 1.1.0-beta

https://doi.org/10.18434/T4/1500024

Public Data Resource

Sequence-based U.S. population data for 27 autosomal STR loci

Contact: Katherine Gettings..

Identifier: doi:10.18434/t4/1500024

Visit Home Page

~800 unique sequences

NIST1036_auSTR_Seq_SuppTables.xlsx	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet	
NIST1036_auSTR_Seq_SuppTables.xlsx.sha256	text/plain	

References

This data is referenced in :

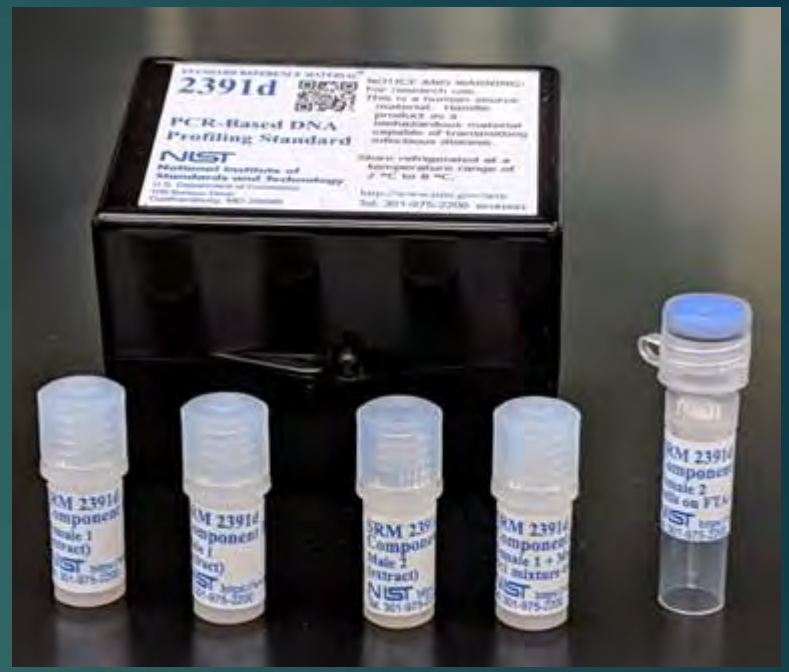
https://doi.org/10.1016/j.fsigen.2018.07.013

STRAND *working group*

align | name | define

Controls for STR Sequencing

- NIST SRM 2391d
- Expected release Summer 2019
- STR sequence data:
  - ForenSeq
  - Precision ID GlobalFiler NGS STR Panel v2
  - PowerSeq 46GY (prototype)



STRAND *working group*

align | name | define

Controls for STR Sequencing

- NIST GiaB
  - 7 Coriell cell lines
  - One individual and two trios
  - PCR-free prep, HiSeq 150 and 250, PacBio
- STR project
  - Any “novel” marker can be characterized
  - ISFG Poster targeting ~500 “novel” STRs

JCMB
The Joint Initiative for Metrology in Biology

STRAND *working group*

align | name | define

Lisa Borsuk - NIST Applied Genetics Group

STRSeq Update

- BioProject Structure
- Record Structure
- Current Sample Sets
- BioProject Status

STRAND working group

align | name | define

STRSeq BioProject Structure

The STR Sequencing Project (human)
Umbrella project
National Institute of Standards and Technology
Accession: PRJNA380127

NIH U.S. National Library of Medicine
National Center for Biotechnology Information

https://www.ncbi.nlm.nih.gov/bioproject/380127

24

Homo sapiens
STRSeq Commonly Used Autosomal STR Loci
Umbrella project
National Institute of Standards and Technology
Accession: PRJNA380345 ID: 380345

11

Homo sapiens
STRSeq Alternate Autosomal STR Loci
Umbrella project
National Institute of Standards and Technology
Accession: PRJNA380346 ID: 380346

7

Homo sapiens
STRSeq X-Chromosomal STR Loci
Umbrella project
National Institute of Standards and Technology
Accession: PRJNA380348 ID: 380348

26

Homo sapiens
STRSeq Y-Chromosomal STR Loci
Umbrella project
National Institute of Standards and Technology
Accession: PRJNA380347 ID: 380347

STRAND *working group*

align | name | define

Homo sapiens

Accession: PRJNA380345 ID: 380345

STRSeq Commonly Used Autosomal STR Loci

This sub-project of the STR Sequencing Project encompasses the data pertaining to 24 autosomal STR loci which are commonly targeted in human identification assays.

Accession	PRJNA380345
Type	Umbrella project
Publications (total 5)	1. Borsuk LA *et al.*, "Sequence-based US population data for the SE33 locus.", *Electrophoresis*, 2018 Jun 1;39(21):2694-2701 More...
Submission	Registration date: 24-Mar-2017 National Institute of Standards and Technology
Related Resources	• STRSeq • STRidER
Relevance	Human Identification

Navigate Up

This project is a component of the The STR Sequencing Project (human)

Navigate Across

3 additional projects are components of the The STR Sequencing Project (human).

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (Genomic DNA)	1251
PUBLICATIONS	
PubMed	5

Homo sapiens encompasses the following 24 sub-projects:

Project Type	Number of Projects
targeted loci	24

BioProject accession	Organism	Title
PRJNA380553	Homo sapiens	STRSeq D1S1656 Sequence-Based Alleles (National Institute of Standards...)
PRJNA380554	Homo sapiens	STRSeq TPOX Sequence-Based Alleles (National Institute of Standards...)
PRJNA380555	Homo sapiens	STRSeq D2S441 Sequence-Based Alleles (National Institute of Standards...)
PRJNA380556	Homo sapiens	STRSeq D2S1338 Sequence-Based Alleles (National Institute of Standards...)
PRJNA380558	Homo sapiens	STRSeq D3S1358 Sequence-Based Alleles (National Institute of Standards...)

List all 24 'targeted loci' projects...

The STR Sequencing Project (human)

Accession: PRJNA380127 ID: 380127

The purpose of STRSeq is to facilitate the description of sequence-based alleles at the Short Tandem Repeat (STR) loci targeted in human identification assays. This collaborative effort of the international forensic DNA community, which has been endorsed by the executive board of the ISFG (International Society of Forensic Genetics), provides a framework for communication among laboratories. Each record contains: (a) observed sequence of an STR region, (b) annotation of the repeat region ("bracketing") and flanking region polymorphisms, (c) information regarding the sequencing assay and data quality, and (d) backward compatible length-based allelic designation. Data within the umbrella project is organized into locus sub-projects, and can be accessed by browsing, BLAST searching, or ftp download at NCBI. For comments or questions, please contact strseq@nist.gov.

Accession	PRJNA380127
Type	Umbrella project
Publications (total 5)	1. Borsuk LA *et al.*, "Sequence-based US population data for the SE33 locus.", *Electrophoresis*, 2018 Jun 1;39(21):2694-2701 More...
Submission	Registration date: 22-Mar-2017 National Institute of Standards and Technology
Related Resources	• STRSeq • STRidER
Relevance	Human Identification

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (Genomic DNA)	1320
PUBLICATIONS	
PubMed	5

The STR Sequencing Project (human) encompasses the following 4 sub-projects:

Project Type	Number of Projects
Umbrella project	4

BioProject accession	Name	Title
PRJNA380345	Homo sapiens	STRSeq Commonly Used Autosomal STR Loci (National Institute of Standards...)
PRJNA380346	Homo sapiens	STRSeq Alternate Autosomal STR Loci (National Institute of Standards...)
PRJNA380347	Homo sapiens	STRSeq Y-Chromosomal STR Loci (National Institute of Standards...)
PRJNA380348	Homo sapiens	STRSeq X-Chromosomal STR Loci (National Institute of Standards...)

Links to BioProjects

STRAND *working group*

align | name | define

STRSeq Record Structure

GenBank ▾ Send to: ▾

Homo sapiens microsate
rs1019813099 sequence

GenBank: MH174843.

Graphics ▾ Send to: ▾

Homo sapiens microsatellite D1S1656 14 CCTA [TCTA]13

Graphics

```
FEATURES             Location/Qualifiers
     source          1..138
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
     misc_feature    1..117
                     /note="Illumina ForenSeq DNA Signature Prep Kit"
     misc_feature    6..138
                     /note="Promega PowerSeq 46GY System"
     variation       38
                     /note="C/T SNP"
                     /db_xref="dbSNP:rs1019813099"
     repeat_region   62..117
                     /rpt_type=tandem
                     /satellite="microsatellite:D1S1656"
ORIGIN
        1 ttcagagaaa tagaatcact agggaaccaa atatatacac atacaattaa acacacacac
       61 acctatctat ctatctatct atctatctat ctatctatct atctatctat ctatctacat
      121 cacacagttg acccttga
//
```

62..117
/rpt_type=tandem
/satellite="microsatellite:D1S1656"

repeat_region Feature 1 of 1 MH174843 : 1 segment Details Display: FASTA GenBank Help ×

of Standards and Technolo
Gaithersburg, Maryland 20

collaborative effort of the international forensic DNA community.
The purpose of this project is to facilitate the description of
sequence-based STR alleles. Additional resources can be found at
strseq.nist.gov. For questions or feedback, please contact
strseq@nist.gov. Allele frequency data can be accessed in the
strider.online database.

GenBank ▾ Send to ▾

Homo sapiens microsatellite D1S1656 14 CCTA [TCTA]13
rs1019813099 sequence

GenBank: MH174843.1

FASTA Graphics

Go to:

```
LOCUS       MH174843                138 bp    DNA     linear   PRI 04-SEP-2018
DEFINITION  Homo sapiens microsatellite D1S1656 14 CCTA [TCTA]13 rs1019813099
            sequence.
ACCESSION   MH174843
VERSION     MH174843.1
DBLINK      BioProject: PRJNA380553
KEYWORDS    STRSeq; STR; D1S1656.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 138)
  AUTHORS   Gettings,K.B., Borsuk,L.A., Ballard,D., Bodner,M., Budowle,B.,
            Devesse,L., King,J., Parson,W., Phillips,C. and Vallone,P.M.
  TITLE     STRSeq: A catalog of sequence diversity at human identification
            Short Tandem Repeat loci
  JOURNAL   Forensic Sci Int Genet 31, 111-117 (2017)
   PUBMED   28888135
REFERENCE   2  (bases 1 to 138)
  AUTHORS   NIST,A.G.G.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-APR-2018) Applied Genetics Group, National Institute
            of Standards and Technology, 100 Bureau Drive, MS-8314,
            Gaithersburg, Maryland 20899, United States of America
COMMENT     Annotation ('bracketing') of the repeat region is consistent with
            the guidance of the ISFG (International Society of Forensic
            Genetics), PMID: 26844919.  Lower case letters in the 'Bracketed
            repeat' region below denote uncounted bases.   The given
            length-based allele value was determined using the designated
            length-based technology.  Variation in the length-based allele
            between individuals or assays can result from indels in flanking
            regions.  The length of reported sequence is dependent on the assay
            and the quality of the flanking sequence. This information is
            provided as part of the STR Sequencing Project (STRseq), a
            collaborative effort of the international forensic DNA community.
            The purpose of this project is to facilitate the description of
            sequence-based STR alleles.  Additional resources can be found at
            strseq.nist.gov.  For questions or feedback, please contact
            strseq@nist.gov.  Allele frequency data can be accessed in the
            strider.online database.

            ##HumanSTR-START##
            STR locus name        :: D1S1656
            Length-based allele   :: 14
            Bracketed repeat      :: CCTA |TCTA|13
            Sequencing technology :: MiSeq FGx; MiSeq
            Coverage              :: >30X
            Length-based tech.    :: PowerPlex Fusion, 3130xl
            Assembly              :: GRCh38 (GCF_000001405)
            Chromosome            :: 1
            RefSeq Accession      :: NC_000001.11
            Chrom. Location       :: 230769555..230769704
            Repeat Location       :: 230769616..230769683
            Cytogenetic Location  :: 1q42.2
            ##HumanSTR-END##
FEATURES             Location/Qualifiers
     source          1..138
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
     misc_feature    1..117
                     /note="Illumina ForenSeq DNA Signature Prep Kit"
     misc_feature    6..138
                     /note="Promega PowerSeq 46GY System"
     variation       38
                     /note="C/T SNP"
                     /db_xref="dbSNP:rs1019813099"
     repeat_region   62..117
                     /rpt_type=tandem
                     /satellite="microsatellite:D1S1656"
ORIGIN
        1 ttcagagaaa tagaatcact agggaaccaa atatatacac atacaattaa acacacacac
       61 acctatctat ctatctatct atctatctat ctatctatct atctatctat ctatctacat
      121 cacacagttg acccttga
//
```

STRAND *working group*

align | name | define

STRSeq Current Sample Sets

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

1786 Samples

ForenSeq Verogen,
PowerSeq 46GY (prototype) Promega, and GlobalFiler NGS Thermo Fisher (and Sanger 😗)
CE Information

King's College London

1043 Samples

ForenSeq Verogen,
CE Information

UNT Health Science Center

839 Samples

ForenSeq Verogen,
CE Information

USC Universidade de Santiago de Compostela

944 Samples

ForenSeq Verogen,
CE Information

Working towards including more sets of data
Collaborating with STRidER

STRAND working group

align | name | define

STRSeq BioProject Status

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (Genomic DNA)	1320
PUBLICATIONS	
PubMed	5

- Currently sequences for autosomal loci have been submitted
  - Additional autosomal sequences continue to be loaded
- Y and X data is going to be submitted in the next few months
  - 2019 Summer - Fall

STRAND *working group*

align | name | define

Jonathan King

STRAND *working group*

align | name | define

- **Current stable release**
  - v3
    - fastq processing (C++ script)
    - Data visualization and processing (Microsoft Excel)
- **Development version**
  - R package
    - Combination of C++ script and functions
      - Full UI
      - Data visualization
      - Allele-calling

STRAND *working group*

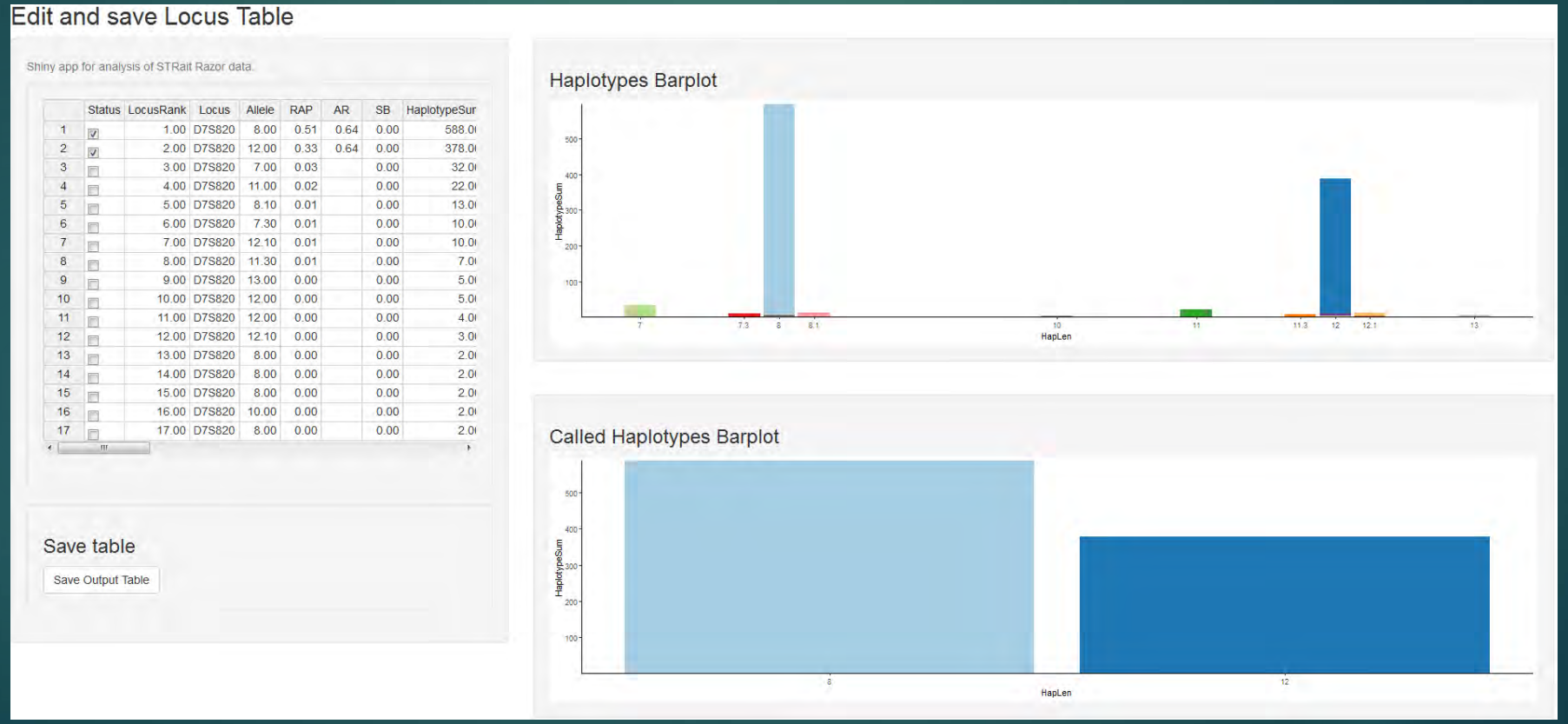
align | name | define

- **Default settings** (ForenSeq-STRs only)
  - ~2s per sample
    - (~15s from fastq)
  - ~97.9% of alleles called automatically
    - (>99% excluding D22S1045…)
  - Stutter Assessment
    - ~6s per sample
- **Population-level processing**

STRAND *working group*

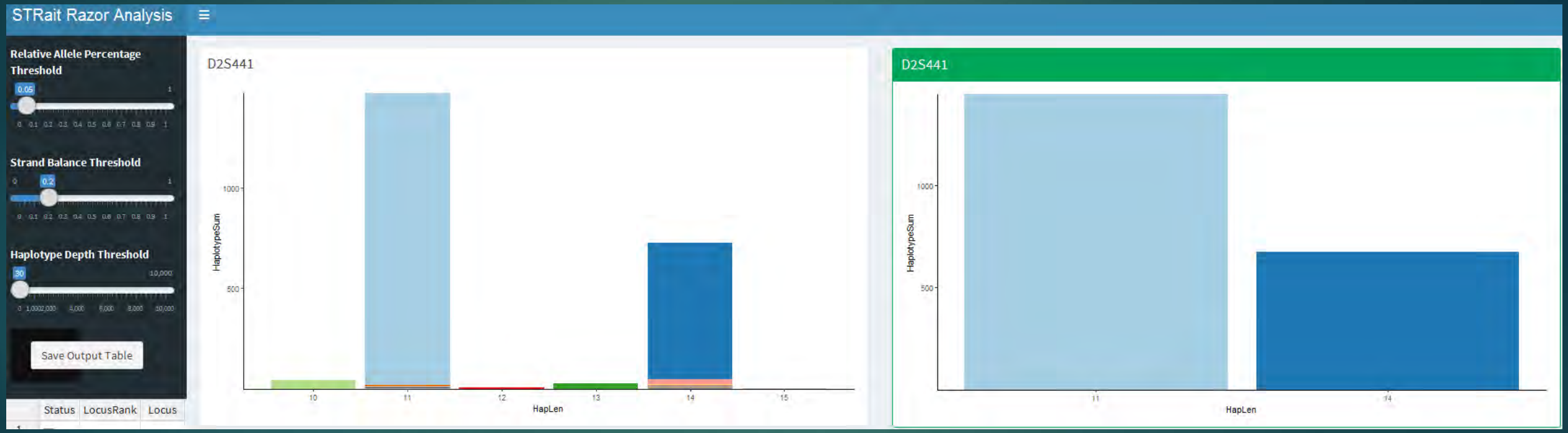
align | name | define

- In Development UI



STRAND *working group*

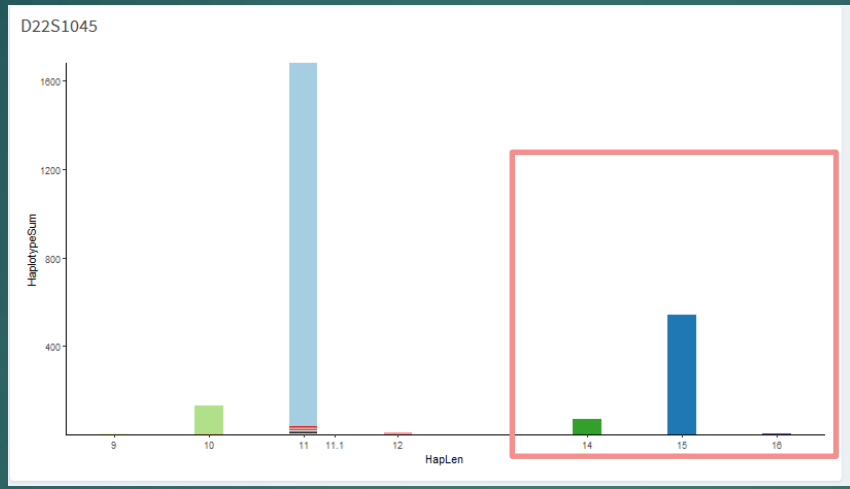
align | name | define

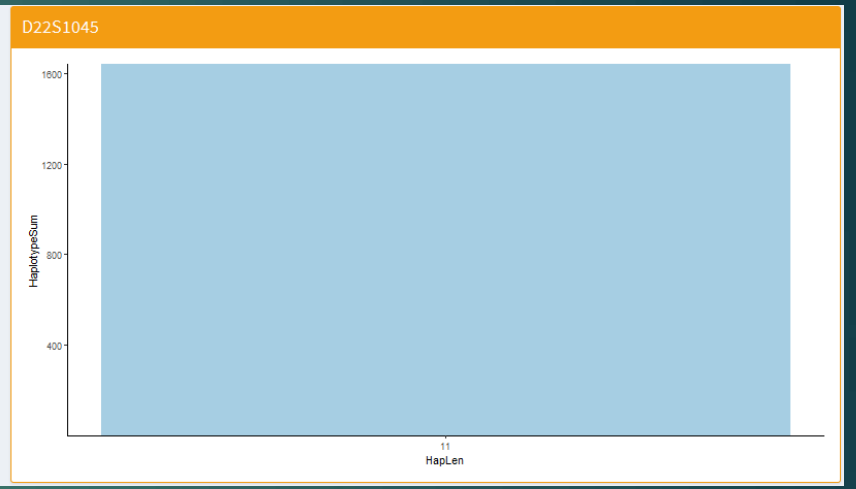


STRAND *working group*

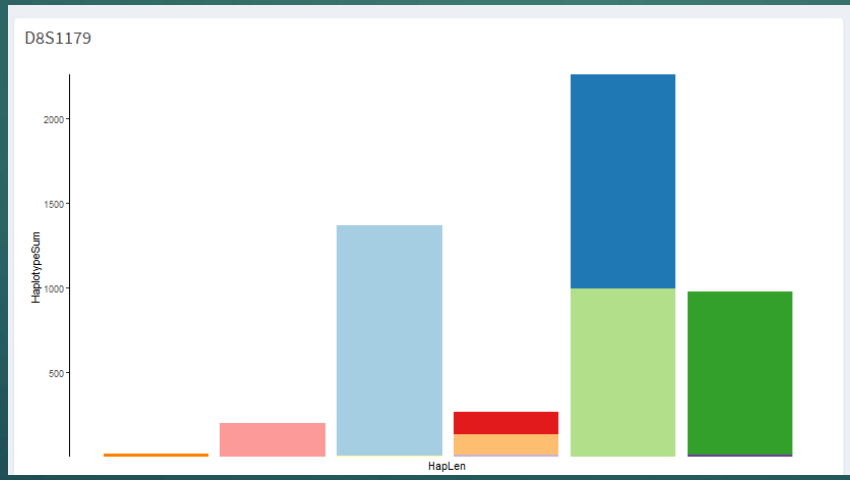
align | name | define

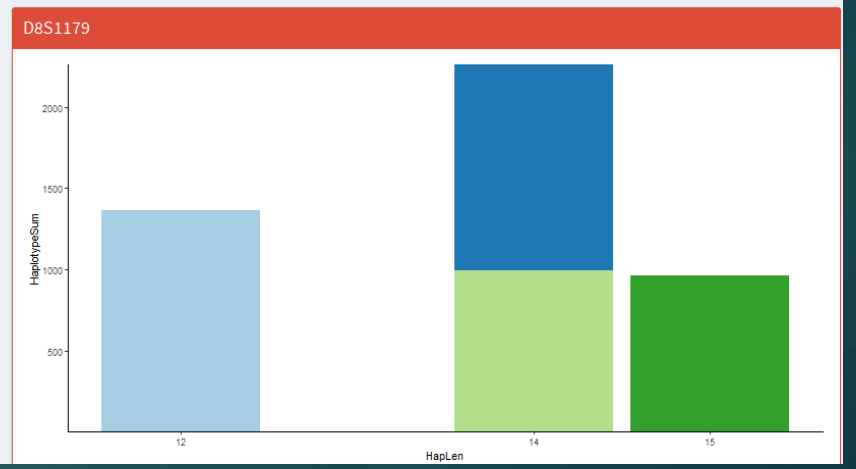
Caution





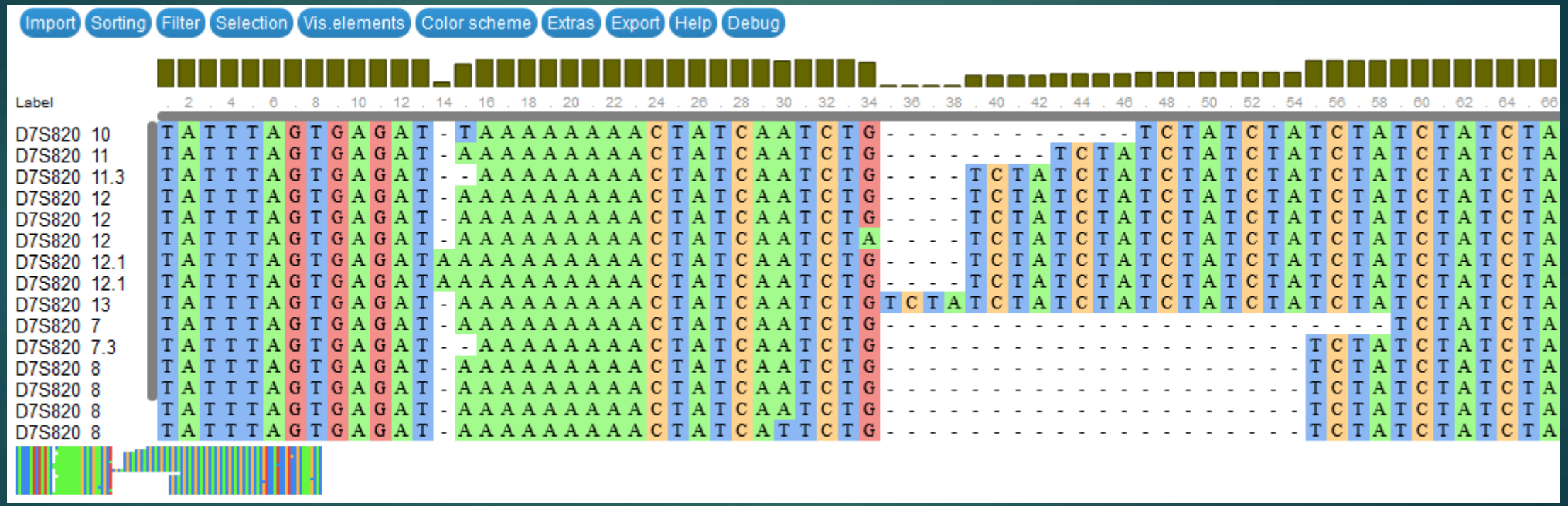
Warning





STRAND *working group*

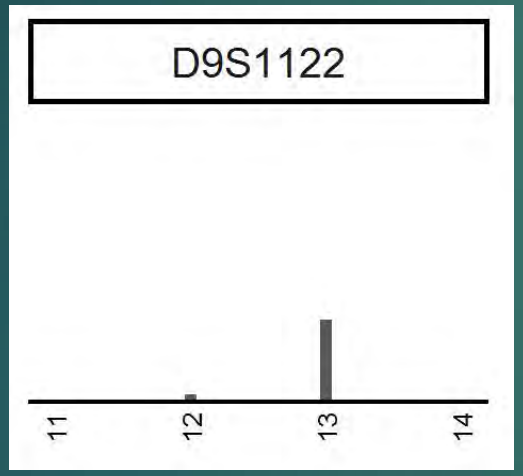
align | name | define

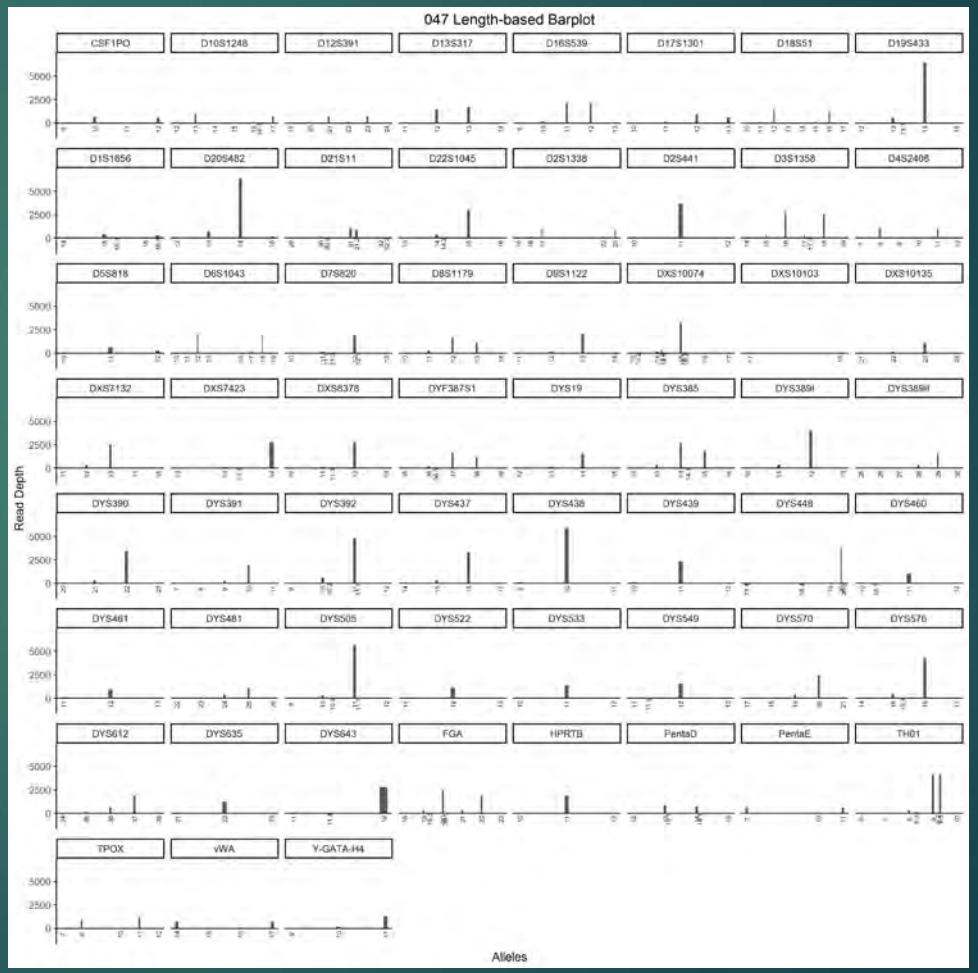


STRAND *working group*

align | name | define

- Length-based Allele Calling

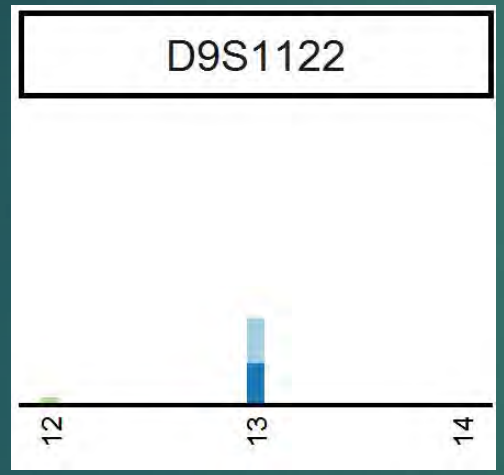


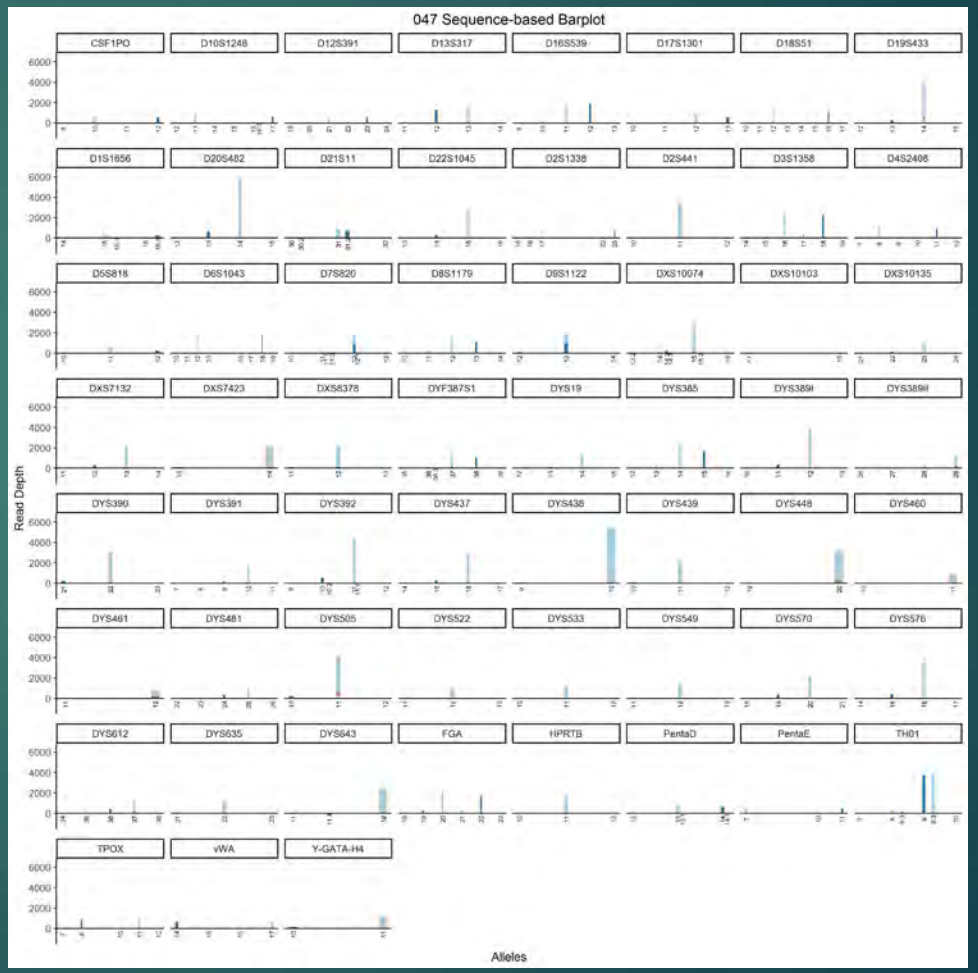


STRAND *working group*

align | name | define

- Sequence-based





STRAND *working group*

align | name | define

- Stutter Assessment
- Refine Levels
  - Locus
    - Most widely implemented
  - Length-based allele
    - Most widely studied
  - Haplotype
    - Most useful ultimately…(probably)

STRAND *working group*

align | name | define

- **Parent Allele:** D2S1338 [CE 17]…[GGAA]11 [GGCA]6
  - ***Primary Stutter Product (~12% of parent allele)***
    - [GGAA]10 [GGCA]6
    - GAGGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAA GG----CAGGCAGGCAGGCAGGCAGGCAAGGCCAAGCCATTT
  - ***Secondary Stutter Product (~1% of parent allele)***
    - [GGAA]11 [GGCA]5
    - GAGGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAA GGAAGG----CAGGCAGGCAGGCAGGCAAGGCCAAGCCATTT
  - ----GAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGGCAGGCAAGGCCAAGCCATTT

STRAND *working group*

align | name | define

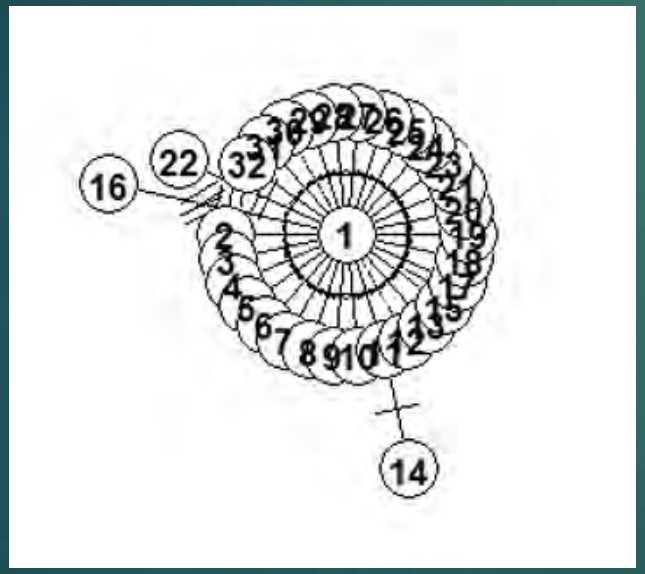
- D21S11
- PA: [TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]9
  - [TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA **[TCTA]8**
    - >6%
  - [TCTA]4 **[TCTG]5** [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]9
    - >1%
  - **[TCTA]3** [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]9
    - <1%
  - [TCTA]4 [TCTG]6 [TCTA]3 TA **[TCTA]2** TCA [TCTA]2 TCCA TA [TCTA]9
    - <1%

STRAND *working group*

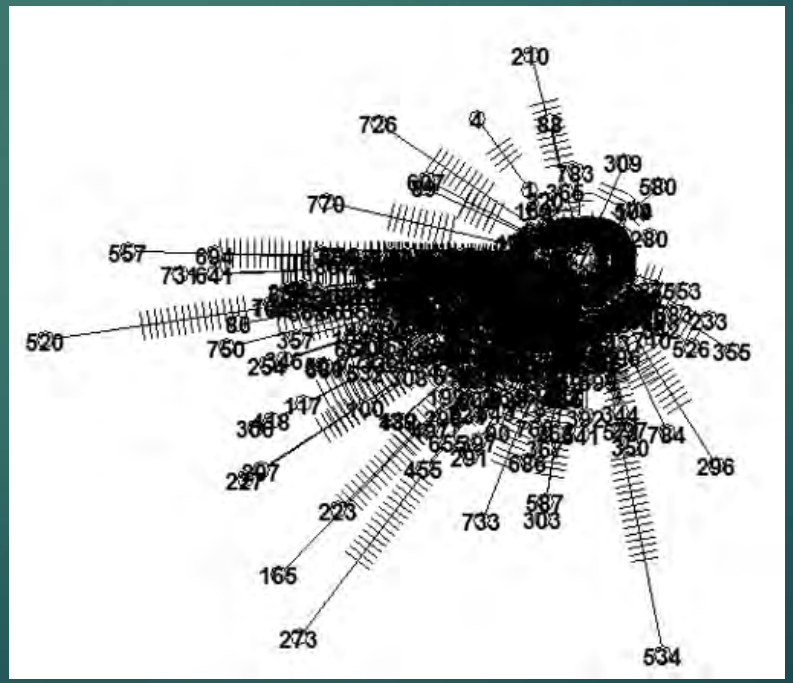
align | name | define

- Understanding Error via Haplotype Networks

CSF1PO



D19S433



STRAND *working group*

align | name | define

- Future Directions
  - Long-term vision of STRait Razor
    - Alignment-based
    - Dynamic flanking region processing/anchor assignment
    - Web interface

STRAND *working group*

align | name | define

UNT Center for Human ID Research and Development Unit

STRAND *working group*

align | name | define

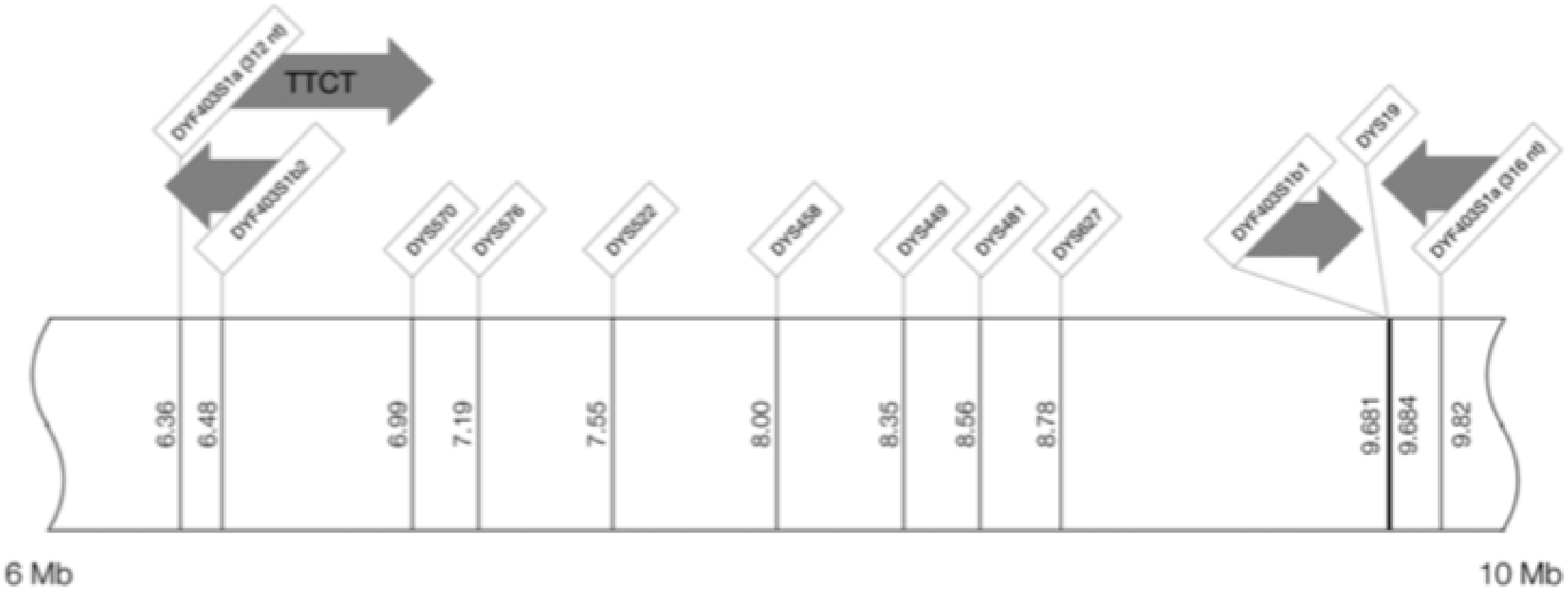
Chris Phillips

Reference genomes, existing databases

GRCh38 has regular updates but no ‘audits’ reveal the sequence build or alignments for forensic STRs have changed

GRCh37 and GRCh 38 do have different sequence builds in one or two loci - most complex comparison of sequences is DXS10146

No distinction can be made between the fragments of duplication-inversion Y-STRs when we move away from the single strand RefSeq framework


DYF403S1a (D12 rd)
TTCT
DYF403S1b2
DYS570
DYS576
DYS522
DYS458
DYS449
DYS481
DYS627
DYF403S1b1
DYS19
DYF403S1a (D16 rd)
6.36
6.48
6.99
7.19
7.55
8.00
8.35
8.56
8.78
9.681
9.684
9.82
6 Mb
10 Mb
DYF399S1 F1
AAT
DYF399S1 F2
DYS448
DYS626
DYF399S1 F3
DYF387S1 F1
TTTC
DYF404S1 F1
AAAG
DYF404S1 F2
DYF387S1 F2
DYF399S1 F4
22.02
22.22
22.27
22.56
22.63
22.79
23.81
25.86
25.88
26.04
21.8 Mb
26.2 Mb

Reference genomes, existing databases

GRCh38 has regular updates but no ‘audits’ reveal the sequence build or alignments for forensic STRs have changed

GRCh37 and GRCh 38 do have different sequence builds in one or two loci - most complex comparison of sequences is DXS10146

No distinction can be made between the fragments of duplication-inversion Y-STRs when we move away from the single strand RefSeq framework

Handling of insertions is quite low key when these are above a certain length

Short Indels can be insertions or deletions of the sequence element - this is rarely fixed in 1000 Genomes / dbSNP annotations

[A/-] is rarely given as [-/A]

[G/GA] rarely [GA/G]

longer sequence elements are given as insertions and this includes STR alleles when these are compiled in 1000 Genomes (not always)

mina reported	GRCh 38	1	2	3	4	5	6	7	8	9	10	11	12						13	Repeat structure	Rpt
		124894865–124894868	124894869–124894872	124894873–124894876	124894877–124894880	124894881–124894884	124894885–124894888	124894889–124894892	124894893–124894896	124894897–124894900	124894901–124894904	124894905–124894908	124894909–124894912						124894913–124894916		
		T C T A	T C T G	T C T A	T C T A	T C T A	T C T A	T C T A	T C T A	T C T A	T C T A	T C T A	T C T A						T C T A		
8	1	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA											TCTA	[TCTA]8	8
9	2	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA										TCTA	[TCTA]9	9
10	3	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA									TCTA	[TCTA]10	10
11	4	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA								TCTA	[TCTA]11	11
11	5	TCTA	TCTA	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA								TCTA	[TCTA]2 TCTG [TCTA]8	11
11	6	TCTA	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA								TCTA	TCTA TCTG [TCTA]9	11
12	7	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA							TCTA	[TCTA]12	12
12	8	TCTA	TCTA	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA							TCTA	[TCTA]2 TCTG [TCTA]9	12
12	9	TCTA	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA							TCTA	TCTA TCTG [TCTA]10	12
13	10	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA						TCTA	[TCTA]13	13
13	11	TCTA	TCTA	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA						TCTA	[TCTA]2 TCTG [TCTA]10	13
13	12	TCTA	TCTA	TCTG	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA						TCTA	[TCTA]2 [TCTG]2 [TCTA]9	13
13	13	TCTA	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA						TCTA	TCTA TCTG [TCTA]11	13
14	14	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA					TCTA	[TCTA]14	14
14	15	TCTA	TCTA	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA					TCTA	[TCTA]2 TCTG [TCTA]11	14
14	16	TCTA	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA					TCTA	TCTA TCTG [TCTA]12	14
15	17	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA				TCTA	[TCTA]15	15
15	18	TCTA	TCTA	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA				TCTA	[TCTA]2 TCTG [TCTA]12	15
15	19	TCTA	TCTA	TCTG	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA				TCTA	[TCTA]2 [TCTG]2 [TCTA]11	15
15	20	TCTA	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA				TCTA	TCTA TCTG [TCTA]13	15
16	21	TCTA	TCTA	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA			TCTA	[TCTA]2 TCTG [TCTA]13	16
16	22	TCTA	TCTA	TCTG	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA			TCTA	[TCTA]2 [TCTG]2 [TCTA]12	16
16	23	TCTA	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA			TCTA	TCTA TCTG [TCTA]14	16
16	24	TCTA	TCTG	TCTG	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA			TCTA	TCTA [TCTG]3 [TCTA]12	16
17	25	TCTA	TCTA	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA		TCTA	[TCTA]2 TCTG [TCTA]14	17
17	26	TCTA	TCTA	TCTG	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA		TCTA	[TCTA]2 [TCTG]2 [TCTA]13	17
18	27	TCTA	TCTA	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	[TCTA]2 TCTG [TCTA]15	18
18	28	TCTA	TCTA	TCTG	TCTG	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	TCTA	[TCTA]2 [TCTG]2 [TCTA]14	18

al framework

		1	2	3	4	5	6	7	8	9	10	11	12						13
8	1	. . . .	. . . A	. . . .	. . . .	. . . .	. . . .	. . . .	— — — —	— — — —	— — — —	— — — —	— — — —						. . . .
9	2	. . . .	. . . A	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	— — — —	— — — —	— — — —	— — — —						. . . .
10	3	. . . .	. . . A	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	— — — —	— — — —	— — — —						. . . .
11	4	. . . .	. . . A	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	— — — —	— — — —						. . . .
11	5	. . . .	. . . A	. . . G	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	— — — —	— — — —						. . . .
11	6	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	— — — —	— — — —						. . . .
12	7	. . . .	. . . A	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	— — — —						. . . .
12	8	. . . .	. . . A	. . . G	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	— — — —						. . . .
12	9	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	— — — —						. . . .
13	10	. . . .	. . . A	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .						. . . .
13	11	. . . .	. . . A	. . . G	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .						. . . .
13	12	. . . .	. . . A	. . . G	. . . G	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .						. . . .
13	13	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .						. . . .
14	14	. . . .	. . . A	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	+ + + +					. . . .
14	15	. . . .	. . . A	. . . G	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	+ + + +					. . . .
14	16	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	+ + + +					. . . .
15	17	. . . .	. . . A	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	+ + + +	+ + + +				. . . .
15	18	. . . .	. . . A	. . . G	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	+ + + +	+ + + +				. . . .
15	19	. . . .	. . . A	. . . G	. . . G	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	+ + + +	+ + + +				. . . .
15	20	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	+ + + +	+ + + +				. . . .
16	21	. . . .	. . . A	. . . G	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	+ + + +	+ + + +	+ + + +			. . . .
16	22	. . . .	. . . A	. . . G	. . . G	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	+ + + +	+ + + +	+ + + +			. . . .
16	23	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	+ + + +	+ + + +	+ + + +			. . . .
16	24	. . . .	. . . .	. . . G	. . . G	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	+ + + +	+ + + +	+ + + +			. . . .
17	25	. . . .	. . . A	. . . G	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	+ + + +	+ + + +	+ + + +	+ + + +		. . . .
17	26	. . . .	. . . A	. . . G	. . . G	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	+ + + +	+ + + +	+ + + +	+ + + +		. . . .
18	27	. . . .	. . . A	. . . G	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	+ + + +	+ + + +	+ + + +	+ + + +	+ + + +	. . . .
18	28	. . . .	. . . A	. . . G	. . . G	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	+ + + +	+ + + +	+ + + +	+ + + +	+ + + +	. . . .

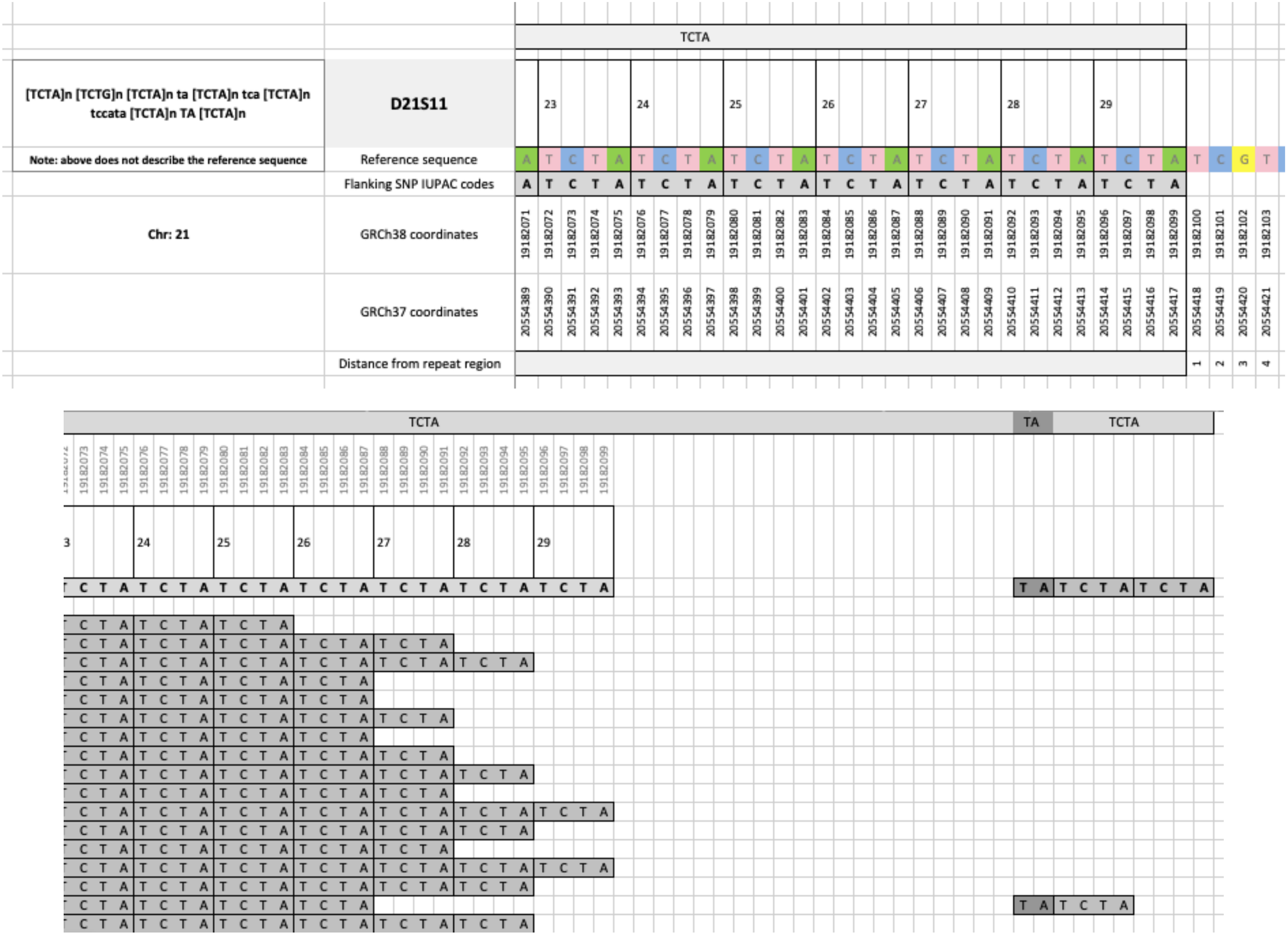
CEPH
D8S1179

Insertions are generally rare but can be difficult to align to RefSeq

22.2	14			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
23	15			GGAG	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
23	16			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
23.2	17			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
24	18			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
24	19			GGAA	GGAA	GGAG	AAAG	AAAA	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
24.2	20			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
24.3	21			GGAA	GGAA	GGAG	AAAG	AAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
25	22			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
25	23			GGAA	GGAA	GGAG	AAAG	AAAA	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
25.2	24			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
26	25			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAGG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
26	26			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
26	27			GGAA	GGAA	GGAG	AAAG	AAAA	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
27	28			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAGG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
27	29			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
27.1	30			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
27.2	31			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
28	32			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAGG	AAAG	AAAG	AAAG	AAAG	AAAG	GAAG	A
28	33			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAGG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
28	34			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
28.1	35			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
29	36			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAGG	AAAG	AAAG	AAAG	AAAG	AAAG	GAAG	A
29	37			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAGG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
29	38			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
30	39			GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAGG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A
31.2	40	GGAA	GGAA	GGAA	GGAA	GGAG	AAAG	AAAG	AAAG	GAAG	GAAG	GAAG	AAAG	AAAG	AAAG	AAAG	AAAG	AAAG	A

FGA

The final 3’ repeat motifs of D21S11 are the only common STR alleles not represented in RefSeq



Reference genomes, existing databases

The two sets of co-ordinates and sequence linked to them won't change substantially

D9S2157										
Reference sequence	C	T	G	T	C	T	C	A	A	T
Flanking SNP IUPAC codes										
GRCh38 coordinates	133160272	133160273	133160274	133160275	133160276	133160277	133160278	133160279	133160280	133160281
GRCh37 coordinates	136035659	136035660	136035661	136035662	136035663	136035664	136035665	136035666	136035667	136035668
Distance from repeat region	10	9	8	7	6	5	4	3	2	1

D9S2157	1			2			3			4			5		
Reference sequence	A	T	A	A	T	A	A	T	A	A	T	A	A	T	A
Flanking SNP IUPAC codes	A	T	A	A	T	A	A	T	A	A	T	A	A	T	A
GRCh38 coordinates	133160282	133160283	133160284	133160285	133160286	133160287	133160288	133160289	133160290	133160291	133160292	133160293	133160294	133160295	133160296
GRCh37 coordinates	136035669	136035670	136035671	136035672	136035673	136035674	136035675	136035676	136035677	136035678	136035679	136035680	136035681	136035682	136035683
Distance from repeat region															

D9S2157	6			7			8			9			10		
Reference sequence	A	T	A	A	T	A	A	T	A	A	T	A	A	T	A
Flanking SNP IUPAC codes	A	T	A	A	T	A	A	T	A	A	T	A	A	T	A
GRCh38 coordinates	133160297	133160298	133160299	133160300	133160301	133160302	133160303	133160304	133160305	133160306	133160307	133160308	133160309	133160310	133160311
GRCh37 coordinates	136035684	136035685	136035686	136035687	136035688	136035689	136035690	136035691	136035692	136035693	136035694	136035695	136035696	136035697	136035698
Distance from repeat region															

D9S2157										
Reference sequence	A	A	T	A	A	A	A	T	T	T
Flanking SNP IUPAC codes										
GRCh38 coordinates	133160312	133160313	133160314	133160315	133160316	133160317	133160318	133160319	133160320	133160321
GRCh37 coordinates	136035699	136035700	136035701	136035702	136035703	136035704	136035705	136035706	136035707	136035708
Distance from repeat region	1	2	3	4	5	6	7	8	9	10

We can accommodate whatever framework for defining the repeat region we agree to

Reference genomes, existing databases

Databases of genomic details pertinent to a variant

- Sequence databases generally compile good quality data for SNPs/Indels
- STR structure data is poor - description and detail varies between loci
- There may be a need to systematise some forensic STR locus names (e.g. FGA) - less regularised than SNPs/Indels, so locating novel loci very difficult

Databases of the population variation observed in a variant

- In last two years >10K genome projects have found many rare variants
- No STR sequence variation database exists - we will have to build our own
- Forensic STR databases of CE-alleles have functioned well up to now

STR.Base *popSTR* *ALFRED*

Databases of genomic details pertinent to a variant

Databases of the population variation observed in a variant

Maps of clustered variants:

Ensembl

STRs and their accompanying flanking SNPs

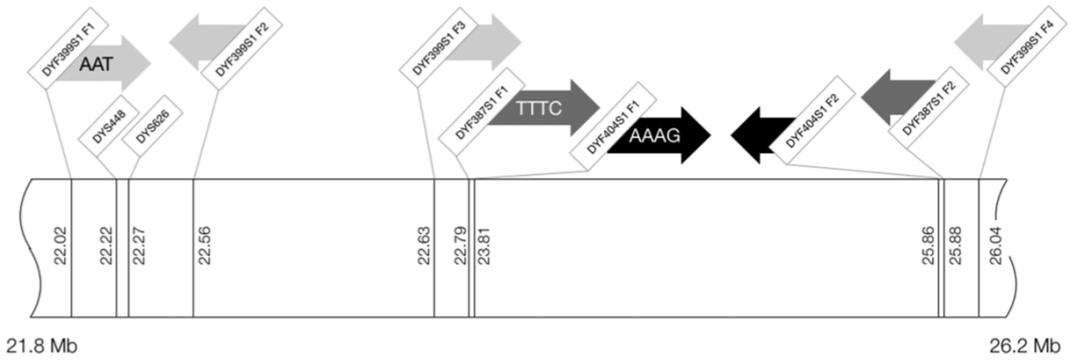
Multiple STRs in close proximity

NCBI Genbank

STR-specific BioProjects

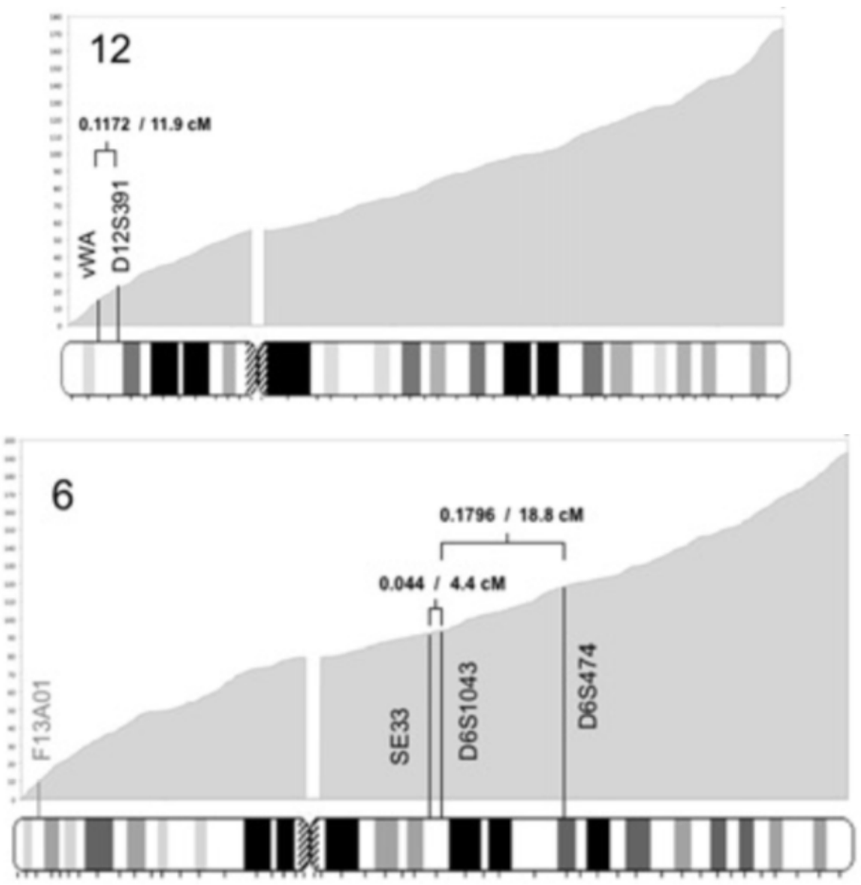
Indels that might influence CE size estimates

Arrangement of multiple fragment STRs



Gauging linkage between close STRs:

HapMap cM estimates of syntenic loci



Identifying less well established STRs as unique

NCBI Probe D5S2500 (in a CE kit) vs D5S2800 (in an MPS kit)

Chrom.	Core STRs	dbSNP rs-number identifier for STR	Chrom.	Supplementary STRs	Kit	dbSNP rs-number identifier for STR
C1	D1S1656	rs113633160	C1	F13B	Promega CS7	*rs10643350*
C2	TPOX	rs113475620	C2	D2S1360	Qiagen HD-plex	rs113680434
C2	D2S1338	rs112111672				
C2	D2S441	*rs10203882*				
C3	D3S1358	rs111694514	C3	D3S1744	Qiagen HD-plex	rs113865588
C4	FGA	*rs67296980*	C4	D4S2366	Qiagen HD-plex	rs113820309
C5	D5S818	rs112497490	C5	D5S2500	Qiagen HD-plex	rs111362704
C5	CSF1PO	rs113729910				
C6	SE33	*rs71021371*	C6	D6S474	Qiagen HD-plex	rs113991233
C7	D7S820	rs112714641	C7	D7S1517	Qiagen HD-plex	rs112397288
C8	D8S1179	*rs67563232*	C8	D8S1132	Qiagen HD-plex	*rs71307053*
			C9	Penta C	Promega CS7	*rs72398274*
C10	D10S1248	rs113518246	C10	D10S2325	Qiagen HD-plex	no SNPs found
C11	TH01	*rs71029110*				
C12	D12S391	rs113002069				
C12	vWA	*rs10579907*				
C13	D13S317	rs111980288				
C15	Penta E	*rs8036258*	C15	FES-FPS	Promega CS7	*rs6229*
C16	D16S539	rs112689398				
C18	D18S51	*rs10560567*				
C19	D19S433	rs113951851				
C21	D21S11	rs113145752	C21	D21S2055	Qiagen HD-plex	rs113225349
C21	Penta D	*rs7279663*				
C22	D22S1045	rs112790319				

Databases of genomic details pertinent to a variant

Databases of the population variation observed in a variant

Maps of clustered variants:

Ensembl

STRs and their accompanying flanking SNPs

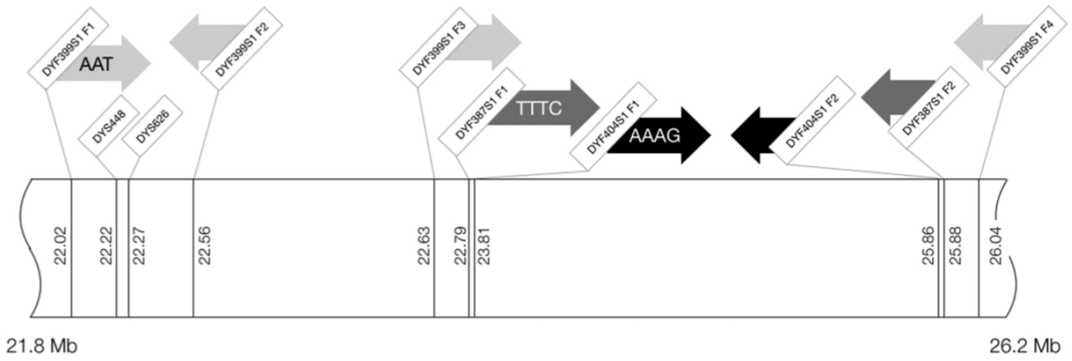
Multiple STRs in close proximity

NCBI Genbank

STR-specific BioProjects

Indels that might influence CE size estimates

Arrangement of multiple fragment STRs



Gauging linkage between close STRs:

HapMap cM estimates of syntenic loci

Identifying less well established STRs as unique

NCBI Probe D5S2500 (in a CE kit) vs D5S2800 (in an MPS kit)

Agreement to use the RefSeq reference strand:

STR Sequence Guide

5’ to 3’ single sequence with fixed coordinates for both current builds

Consensus amongst MPS users for loci and level of detail

Consensus amongst STRAND members for each STR’s START-END nucleotides and repeat region structures

Guide Wiki Database

‘Red-point’ SNPs

Mobility-shift SNPs

Rare 4-nt Indels

√ Flanking variants even at low frequency are easy to compile

STRSeq compiling full sequences including flanks

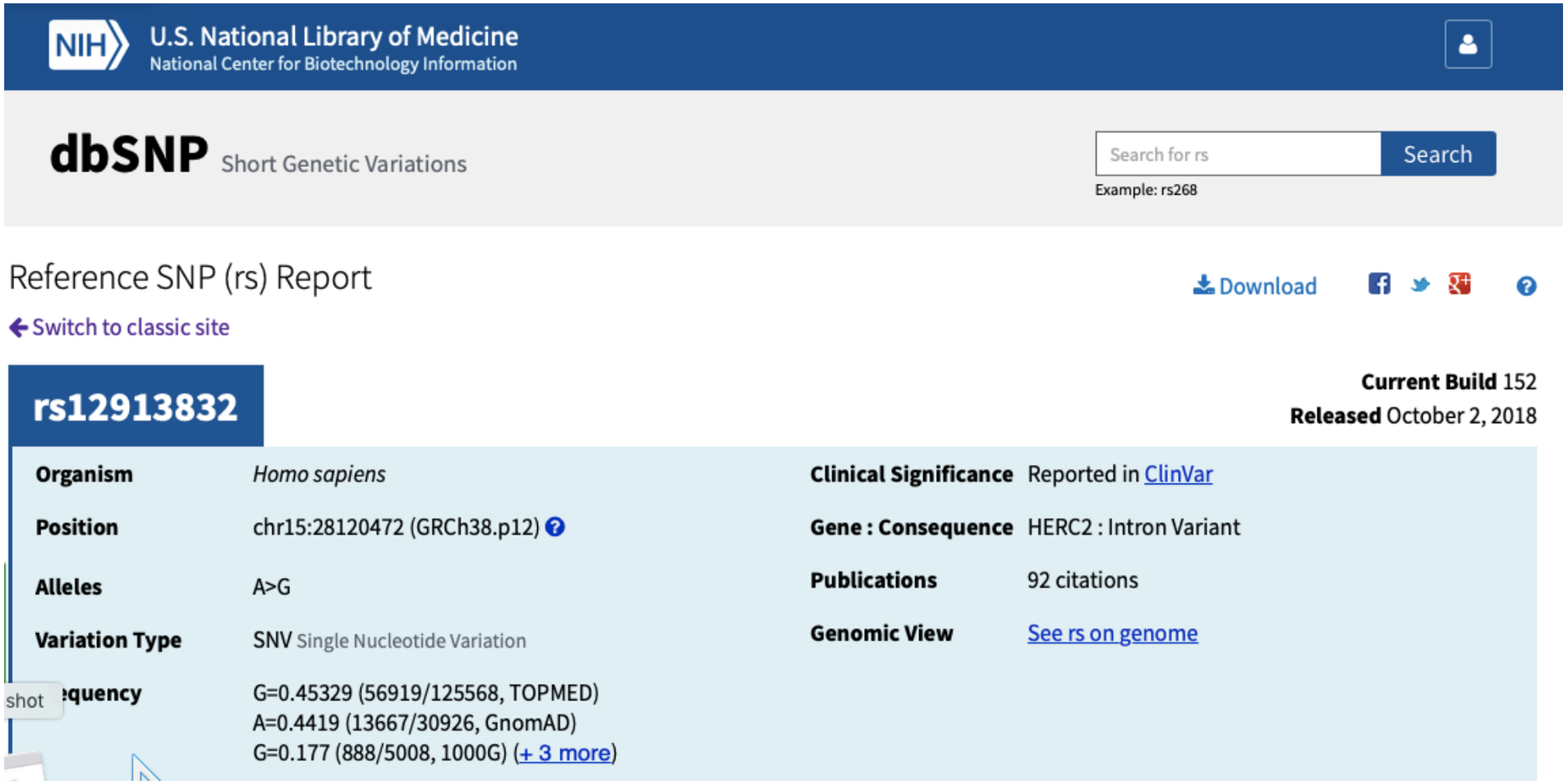
dbSNP is accelerating the uptake of Indels and assignment of rs-numbers

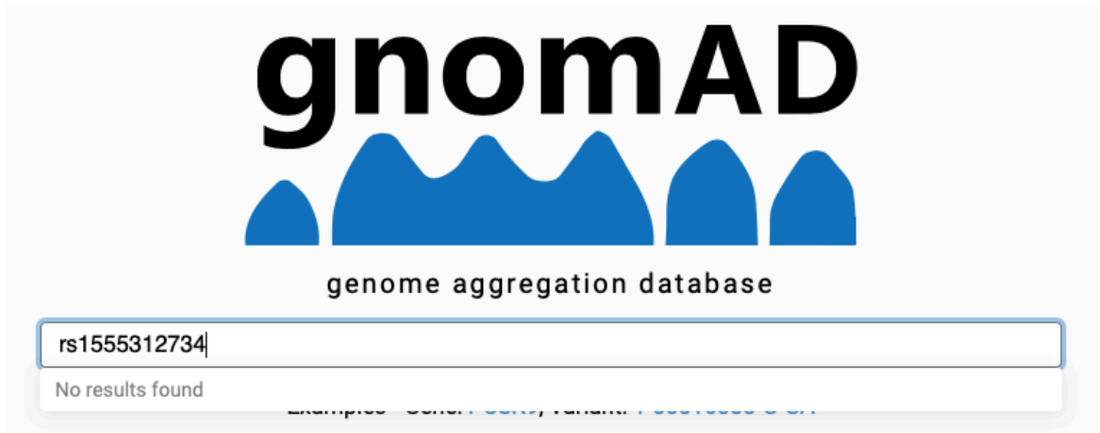
1000 genomes is now supplemented by gnomAD / TOPMed - tens of thousands of samples

Databases of genomic details pertinent to a variant

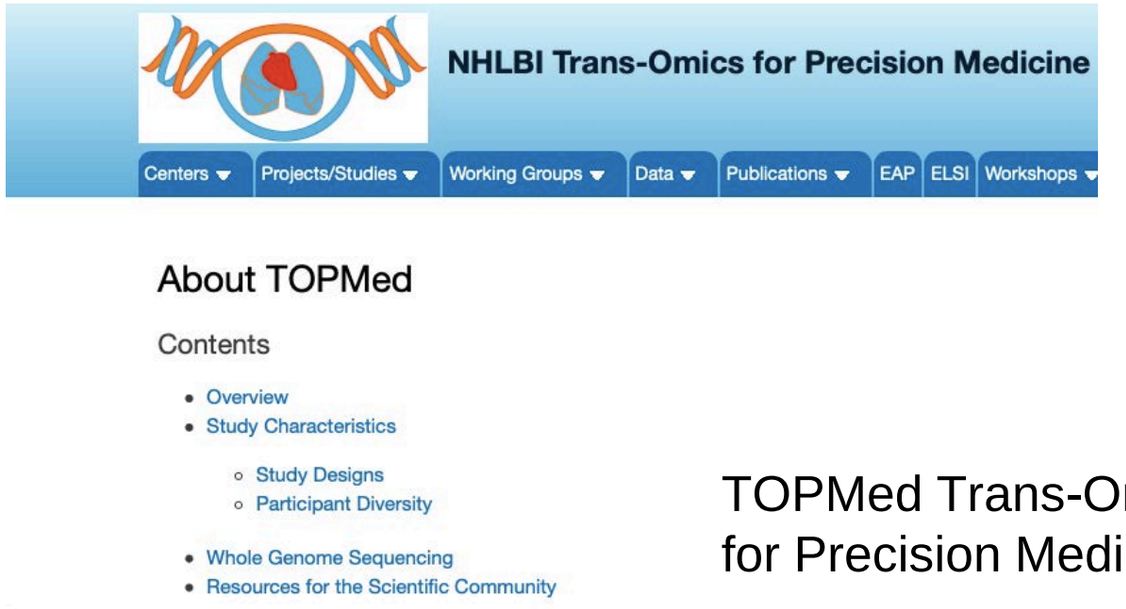
Databases of the population variation observed in a variant

dbSNP is undergoing a transition in presentation and scope to accommodate many more rare SNPs identified by >10,000 sample projects





gnomAD Genome Aggregation Database



TOPMed Trans-Omics for Precision Medicine

gnomAD Data

gnomAD genomes (8)

Show/hide columns

Population	Allele: frequency (count)	
gnomADg:ALL	ATCT: 0.999551253 (31184)	-: 0.000448747 (14)
gnomADg:afr	ATCT: 0.999769319 (8668)	-: 0.000230681 (2)
gnomADg:amr	ATCT: 1.000 (836)	-: 0.000
gnomADg:asj	ATCT: 0.997 (285)	-: 0.003 (1)
gnomADg:eas	ATCT: 1.000 (1538)	-: 0.000
gnomADg:fin	ATCT: 1.000 (3458)	-: 0.000
gnomADg:nfe	ATCT: 0.999 (15317)	-: 0.001 (11)
gnomADg:oth	ATCT: 1.000 (1082)	-: 0.000

dbSNP — SNP ss2137543798 — Create alert | Advanced

Display Settings: Summary

rs1442523705 *[Homo sapiens]*

1.

Variant type:	DELINS
Alleles:	ATCTATCTATCT>ATCTATCT
Chromosome:	13:82148077
Validated:	by cluster
HGVS:	NC_000013.11:g.82148077_82148080ATCT[2],

TOPMed (1)

Show/hide columns

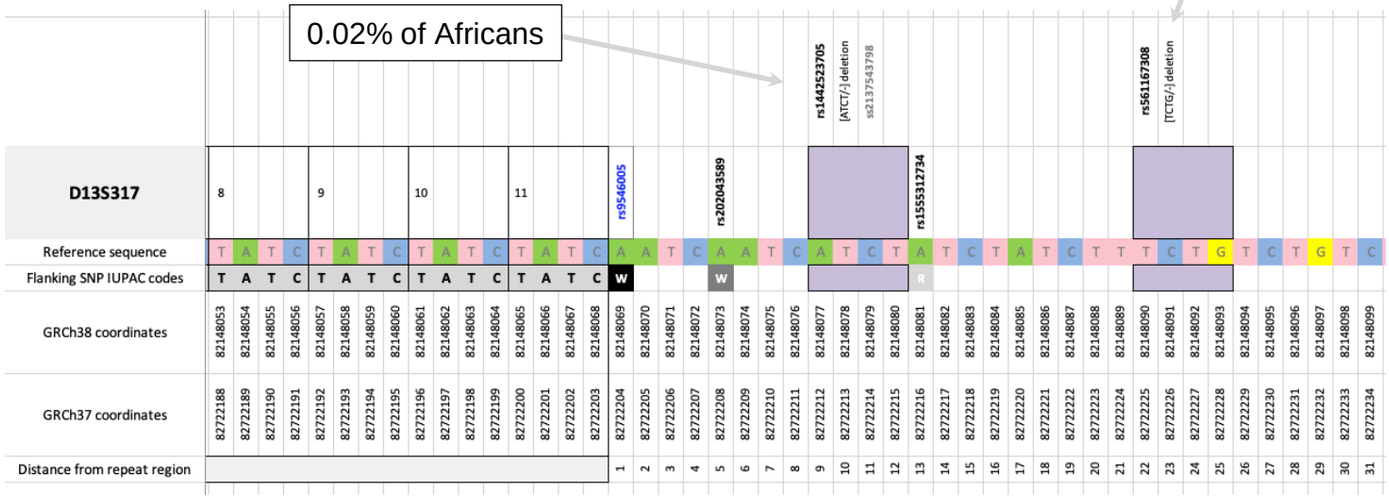
Population	Allele: frequency (count)
TOPMed	ATCT: 0.999601809 -: 0.000398191

rs1442523705 DELETION

Most severe consequence	intergenic variant
Alleles	ATCT/- | Ancestral: ATCT | Highest population MAF: < 0.01
Change tolerance	GERP: 0.58
Location	Chromosome 13:82148077-82148080 (forward strand) | VCF: 13 82148076 rs1442523705 CATCT C
Evidence status	
HGVS name	NC_000013.11:g.82148085_82148088del
Synonyms	ClinGen Allele Registry CA611275396 (-)
Original source	Variants (including SNPs and indels) imported from dbSNP (release 151) | View in dbSNP

11% of Africans

0.02% of Africans



gnomAD genomes (8)

Show/hide columns

Population	Allele: frequency (count)	
gnomADg:ALL	TCTT: 0.994 (30443)	-: 0.006 (169)
gnomADg:afr	TCTT: 0.980 (8320)	-: 0.020 (166)
gnomADg:amr	TCTT: 0.999 (809)	-: 0.001 (1)
gnomADg:asj	TCTT: 1.000 (286)	-: 0.000
gnomADg:eas	TCTT: 1.000 (1544)	-: 0.000
gnomADg:fin	TCTT: 1.000 (3228)	-: 0.000
gnomADg:nfe	TCTT: 0.9999342278 (15203)	-: 6.57722e-05 (1)
gnomADg:oth	TCTT: 0.999 (1053)	-: 0.001 (1)

dbSNP — SNP ss2137535201 — Create alert | Advanced

Display Settings: Summary

rs1371483225 *[Homo sapiens]*

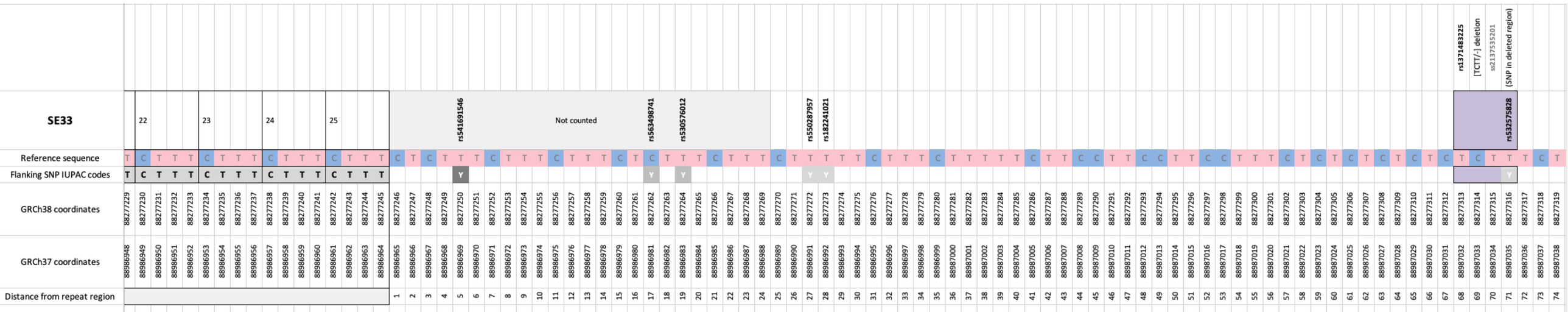
1.

Variant type:	DELINS
Alleles:	TCTTTCTTTCTTTCT>TCTTTCTTTCT
Chromosome:	6:88277313
Gene:	LOC105377884 (GeneView)
Functional Consequence:	downstream_transcript_variant
Validated:	by frequency,by cluster
Global MAF:	-=0.0057/173 (GnomAD)
HGVS:	NC_000006.12:g.88277316_88277319TTCT[2]

rs1371483225 DELETION

Most severe consequence	intron variant | See all predicted consequences
Alleles	TCTT/- | Ancestral: TCTT | Highest population MAF: 0.02
Change tolerance	GERP: 0.40
Location	Chromosome 6:88277313-88277316 (forward strand) | VCF: 6 88277312 rs1371483225
Evidence status	gnomAD
HGVS names	This variant has 2 HGVS names - Show
Synonyms	ClinGen Allele Registry CA568842946 (-)
Original source	Variants (including SNPs and indels) imported from dbSNP (release 151) | View in dbSNP

2% of Africans

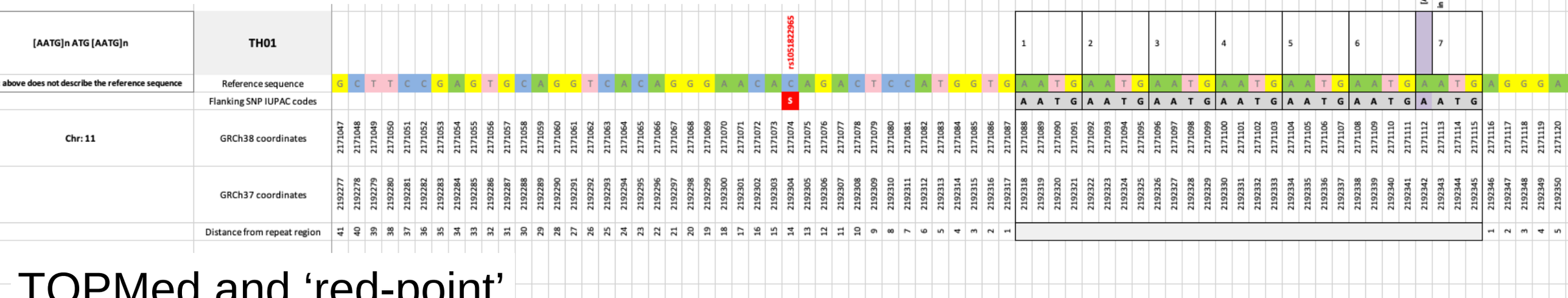


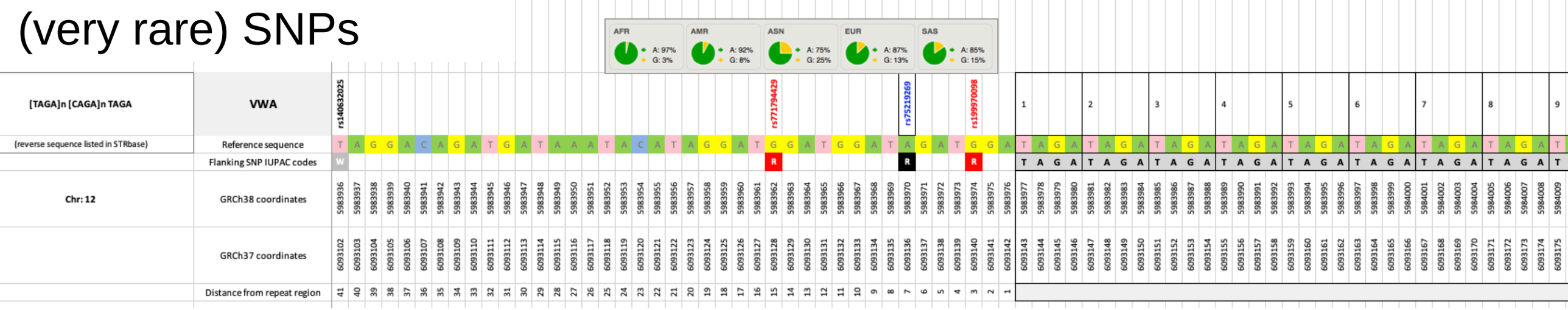
TOPMed and ‘red-point’ (very rare) SNPs

rs1051822965 SNP

Most severe consequence	intron variant | See all predicted consequences
Alleles	C/G | Ancestral: C | Highest population MAF: < 0.01
Change tolerance	CADD: G:3.952 | GERP: -3.35
Location	Chromosome 11:2171074 (forward strand) | VCF: 11 2171074 rs1051822965 C G
Evidence status	gnom AD
HGVS names	This variant has 10 HGVS names - Show
Synonyms	ClinGen Allele Registry CA216229061 (G)
Original source	Variants (including SNPs and indels) imported from dbSNP (release 151) | View in dbSNP
About this variant	This variant overlaps 6 transcripts.

gnomADg:ALL	C: 0.999899112 (29733)	G: 0.000100888 (3)
gnomADg:afr	C: 0.999635569 (8229)	G: 0.000364431 (3)
gnomADg:amr	C: 1.000 (832)	G: 0.000
gnomADg:asj	C: 1.000 (264)	G: 0.000
gnomADg:eas	C: 1.000 (1412)	G: 0.000
gnomADg:fin	C: 1.000 (3432)	G: 0.000
gnomADg:nfe	C: 1.000 (14546)	G: 0.000
gnomADg:oth	C: 1.000 (1018)	G: 0.000





rs771794429 SNP

gnomAD Genome Aggregation Database

Most severe consequence	intron variant | See all predicted consequences
Alleles	G/A | Ancestral: G | MAF: 0.01 (A) | Highest population MAF: < 0.01
Change tolerance	CADD: A:1.395 | GERP: -2.86
Location	Chromosome 12:5983962 (forward strand) | VCF: 12 5983962 rs771794429 G A
Evidence status	1K
HGVS names	This variant has 4 HGVS names - Show
Synonyms	ClinGen Allele Registry CA232291691 (A)
Original source	Variants (including SNPs and indels) imported from dbSNP (release 151) | View in dbSNP
About this variant	gnomADg:afr G: 0.999 (8041) A: 0.001 (5)

rs199970098 SNP

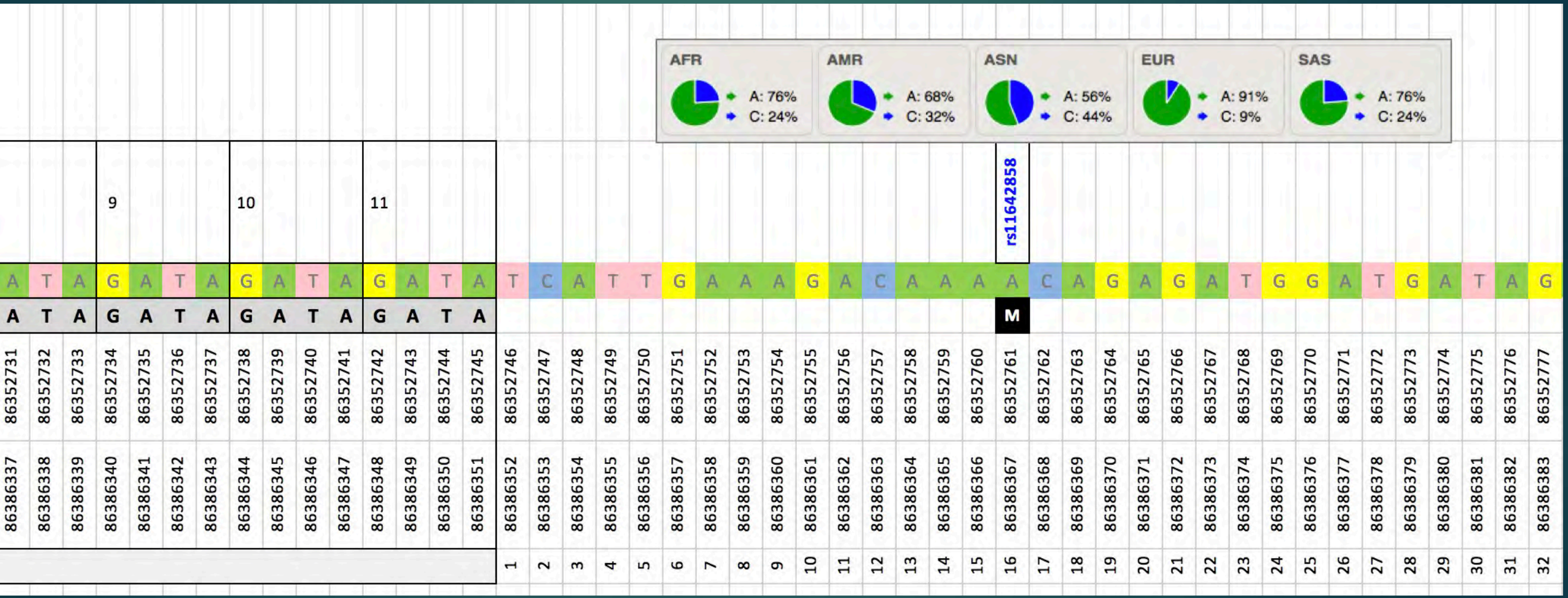
TOPMed Trans-Omics for Precision Medicine Program

Most severe consequence	intron variant | See all predicted consequences
Alleles	G/A | Ancestral: A | Highest population MAF: 0.01
Change tolerance	CADD: A:0.041 | GERP: -3.74
Location	Chromosome 12:5983974 (forward strand) | VCF: 12 5983974 rs199970098 G A
Evidence status	
HGVS names	This variant has 4 HGVS names - Show
Synonyms	ClinGen Allele Registry CA232291705 (A)
Original source	Variants (including SNPs and indels) imported from dbSNP (release 151) | View in dbSNP
About this variant	TOPMed G: 0.985 A: 0.015

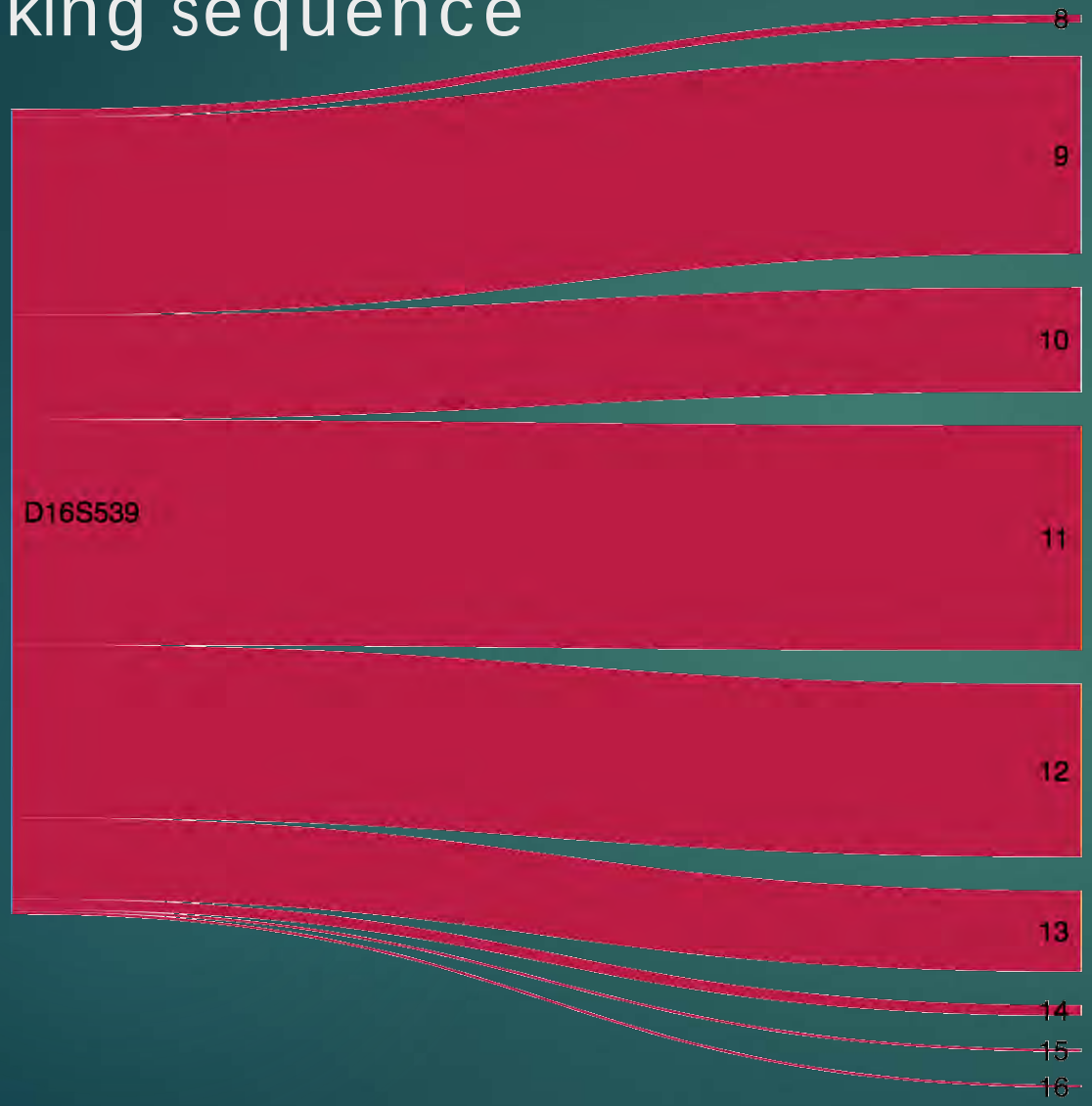
STR Flanking Region Variation

David Ballard – King's College London

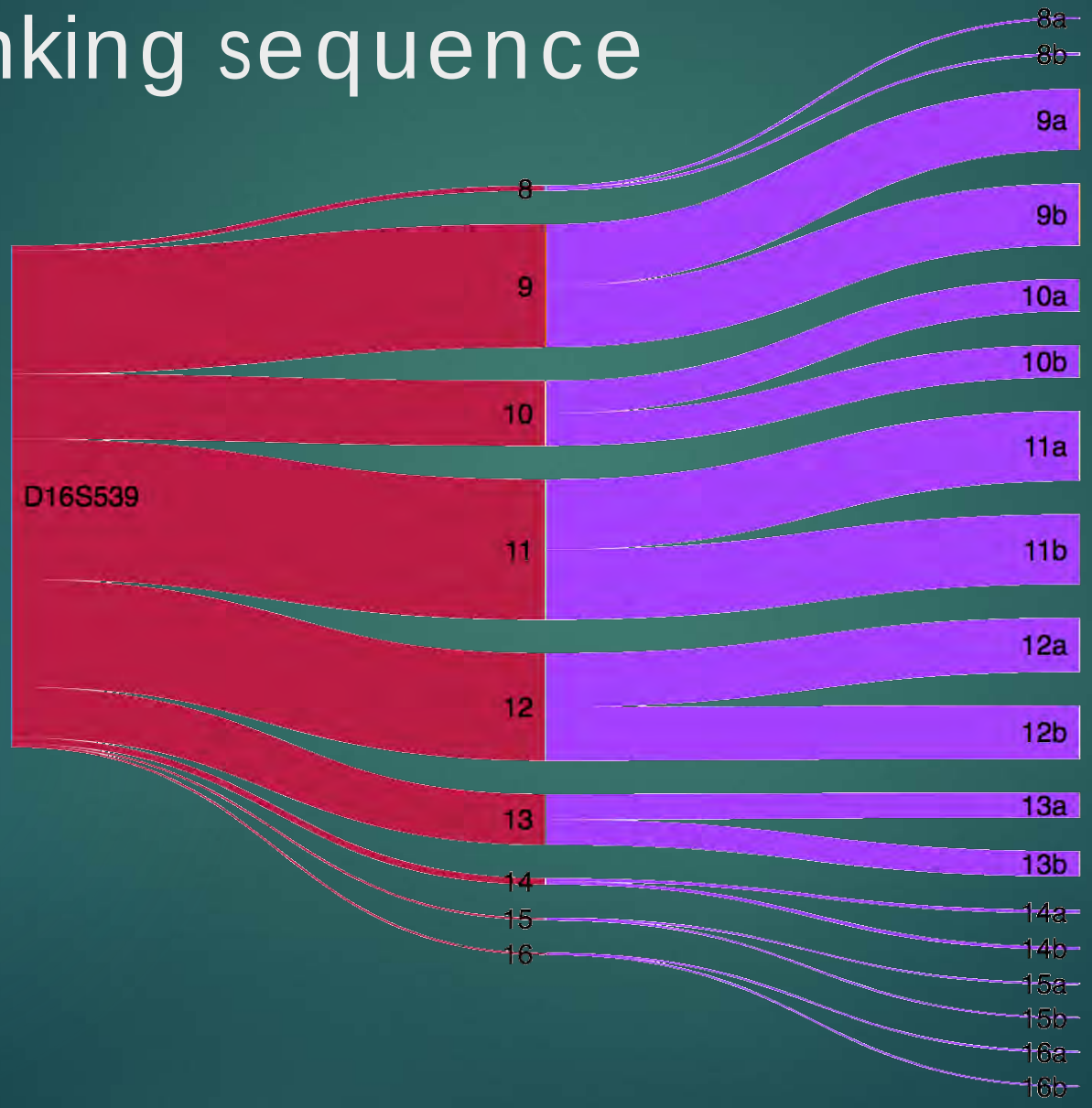
D16S539 – Known Flanking SNP Variation



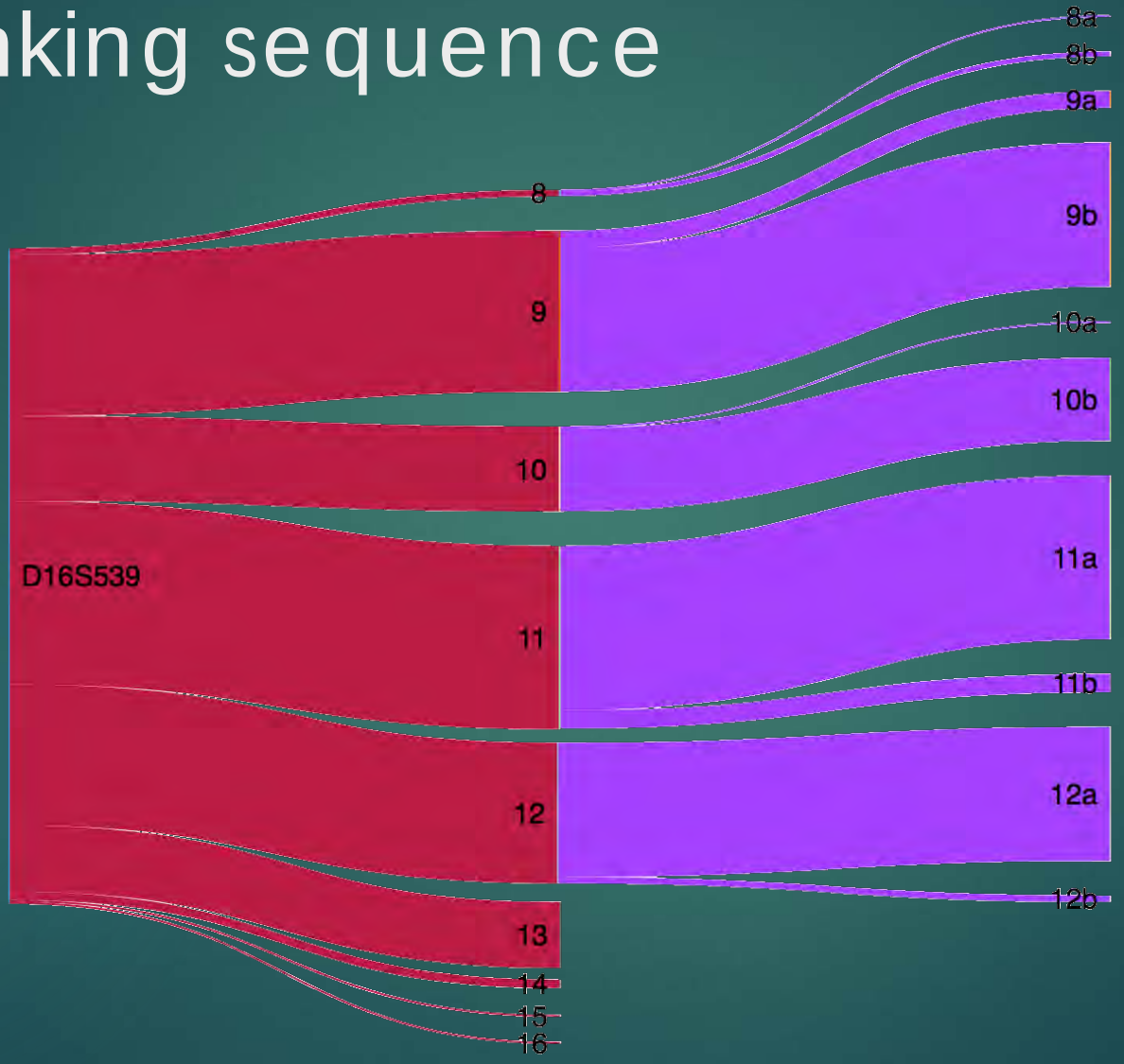
D16S539 – Expected frequency of alleles using flanking sequence



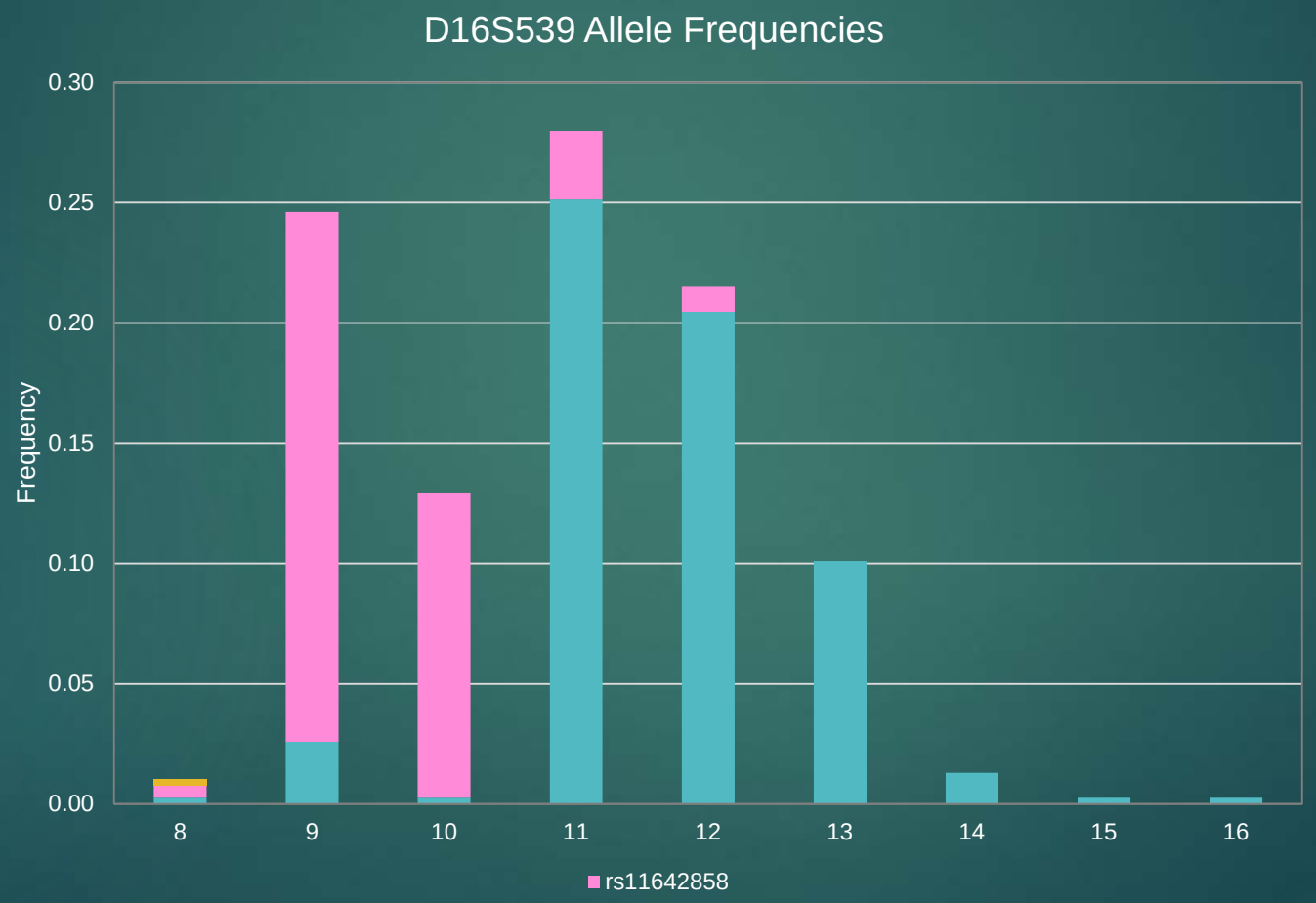
D16S539 – Expected frequency of alleles using flanking sequence



D16S539 – Actual frequency of alleles using flanking sequence



D16S539 – Why isn't the flanking region helping?



STR nomenclature

To consider?

- For discrimination purposes, extended flanking region variation is of limited use due to linkage with specific STR repeat sequence alleles
- A nomenclature system just describing the repeat region (or a defined region around the repeat) would be simple and capture almost all of the useful variation
- A collection of all common STR sequence allele variation is already available

UNIVERSITY OF LEICESTER

STR Sequence Nomenclature – The view from Leicester

Tunde Huszar

University of Leicester
Alec Jeffreys Forensic Genomics Unit
Department of Genetics and Genome Biology
Leicester, UK
th201@leicester.ac.uk

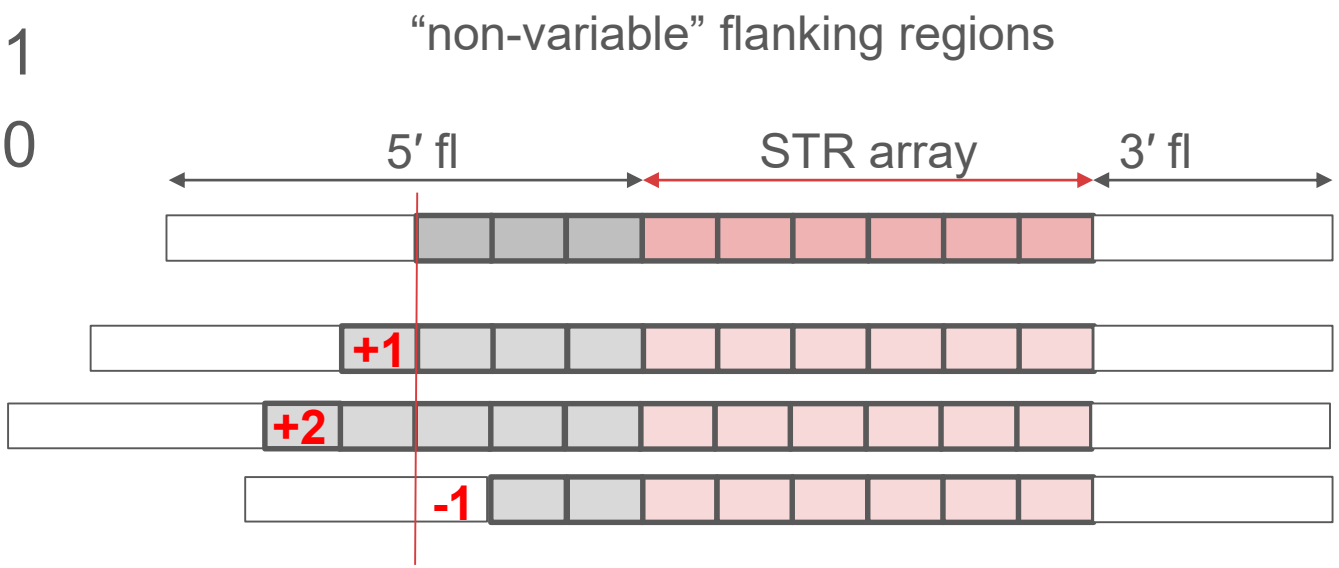
STR Sequence Nomenclature

- ✓ Lessons from a phylogenetic framework for Y-STRs
- ✓ Existing databases – using STRSeq
- ✓ Local database – LeiceSTRSeq

Phylogenetic framework

✓ variable array limits:

- DYS385a,b
- DYS481
- DYS390



Huszar et al. FSI: Genetics (2018), 35, 97-106
A phylogenetic framework facilitates Y-STR variant discovery and classification via massively parallel sequencing

UNIVERSITY OF LEICESTER

Phylogenetic framework - DYS385a,b

conventionally : (STRBase, strbase.nist.gov)

DYS385a,b: $[GAAA]_n$

2016 ISFG recommendation: (Parson et al. 2016 FSI Gen)

DYS385a : $[TTTC]_n$

DYS385b: $[GAAA]_n$

no distinction between a/b in kits – suggestion:

DYS385a,b: $[GAAA]_n$

Huszar et al. FSI: Genetics (2018), 35, 97-106
A phylogenetic framework facilitates Y-STR variant discovery and classification via massively parallel sequencing

Phylogenetic framework - DYS385a,b

✓ Variable flanking regions:

recognise repeat structure, rather than calling several SNPs

Allele	Observed #	General structure of alleles including variable flanking sequences	CE allele name designation
canonical	193	AAGG[6]GAAA[n]	n
variant	4	AAGG[5]GAAA[n]	n − 1
variant	2	AAGG[7]GAAA[n]	n + 1
variant	2	AAGG[8]GAAA[n]	n + 2
variant	*	AAGG[9]GAAA[n]	n + 3

* observed in Novroski et al. 2016, FSI Gen

Huszar et al. FSI: Genetics (2018), 35, 97-106
A phylogenetic framework facilitates Y-STR variant discovery and classification via massively parallel sequencing

UNIVERSITY OF LEICESTER

Phylogenetic framework

- ✓ high sequence variability – flexible software

  FDSTools (Hoogenboom et al. 2017, FSI Gen)

  non-standard populations, new variants, non-human STRs

- ✓ Multiple software/analysis – against bioinformatic nulls

  UAS / STRaitRazor / FDSTools /

  / commercial software / in-house scripts

Huszar et al. FSI: Genetics (2018), 35, 97-106
A phylogenetic framework facilitates Y-STR variant discovery and classification via massively parallel sequencing

UNIVERSITY OF LEICESTER

Existing databases

- ✓ STRSeq: (Gettings et al. 2017, FSI Gen)

  GenBank records at NCBI - (unique Acc#)

  BioProjects (Auto, Auto+, Y- ,X-STRs)

  "Project Data:
  No public data is linked to this project. Any recently released data that cites this project will be linked to it within a few days."

- ✓ STRidER: (Bodner et al. 2016, FSI Gen)

  STR sequence guide v4 (Phillips et al. 2018, FSI Gen)

  user interface, pathway for submission, QC

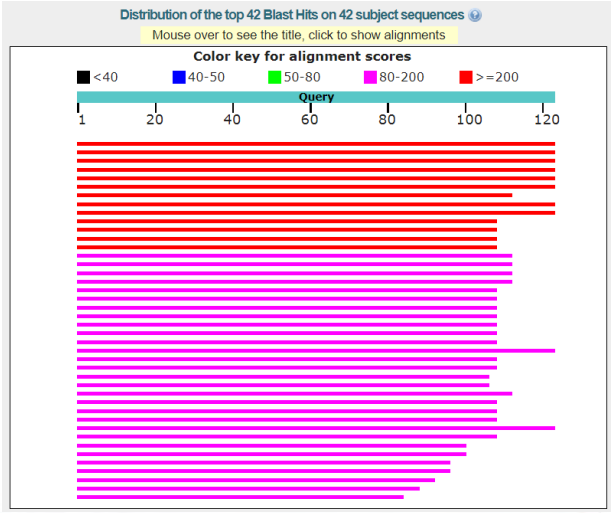
Existing databases

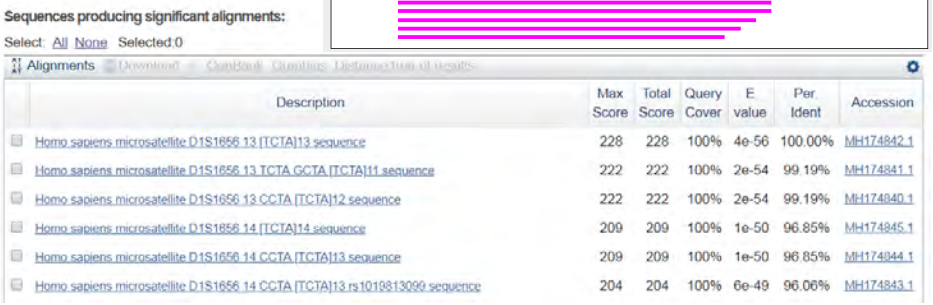
- ✓ STRSeq: (Auto, Auto+, Y- ,X-STRs)

  search by BLASTn

  detailed, comparable,

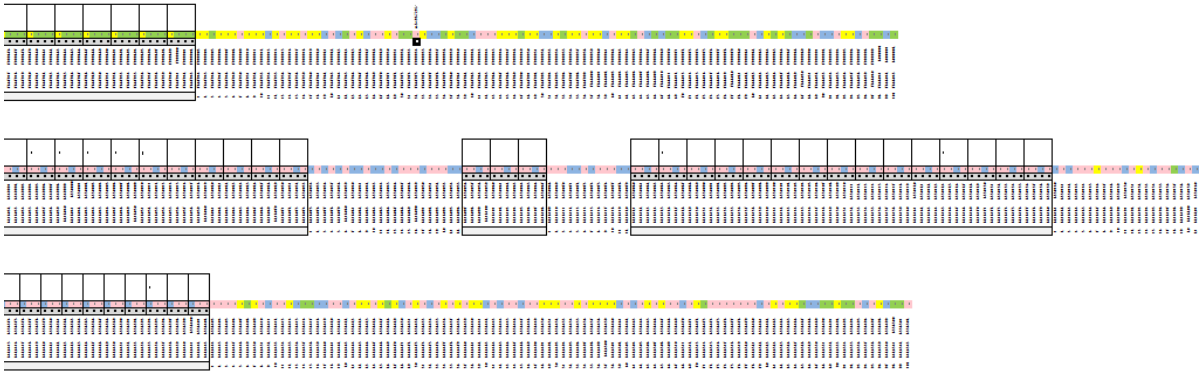
  sequence identifiers





- ✓ STRidER:

  compendium for variation



UNIVERSITY OF LEICESTER

Existing databases - Issues

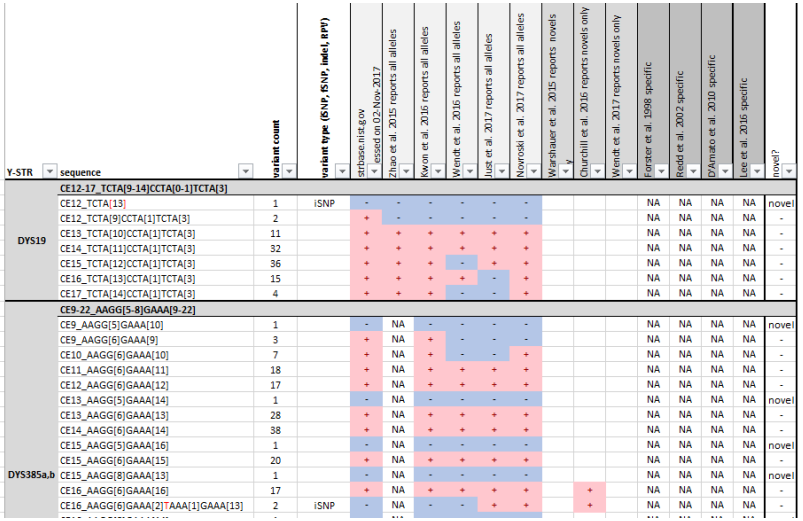
- ✓ STRSeq: not yet applicable for Auto+, Y- ,X-STRs

  instead: literature search

  redundant task

  no unique sequence ID

  LeiceSTRSeq – local DB



Y-STR	sequence	variant count	variant type: (ISNP, ISNP, indel, RPF)	strbase.nist.gov	essed on 02-Nov-2017	Zhao et al. 2015 reports all alleles	Kwon et al. 2016 reports all alleles	Wendt et al. 2016 reports all alleles	Just et al. 2017 reports all alleles	Novroski et al. 2017 reports all alleles	Warshauer et al. 2015 reports novels only	Churchill et al. 2016 reports novels only	Wendt et al. 2017 reports novels only	Forster et al. 1998 specific	Redd et al. 2002 specific	D'Amato et al. 2010 specific	Lee et al. 2016 specific	novel?
	CE12-17_TCTA[9-14]CCTA[0-1]TCTA[3]																	
DYS19	CE12_TCTA[13]	1	ISNP	-	-	-	-	-	-					NA	NA	NA	NA	novel
	CE12_TCTA[9]CCTA[1]TCTA[3]	2		+	-	-	-	-	-					NA	NA	NA	NA	-
	CE13_TCTA[10]CCTA[1]TCTA[3]	11		+	+	+	+	+	+					NA	NA	NA	NA	-
	CE14_TCTA[11]CCTA[1]TCTA[3]	32		+	+	+	+	+	+					NA	NA	NA	NA	-
	CE15_TCTA[12]CCTA[1]TCTA[3]	36		+	+	+	-	+	+					NA	NA	NA	NA	-
	CE16_TCTA[13]CCTA[1]TCTA[3]	15		+	+	+	+	-	+					NA	NA	NA	NA	-
	CE17_TCTA[14]CCTA[1]TCTA[3]	4		+	+	+	-	-	+					NA	NA	NA	NA	-
	CE9-22_AAGG[5-8]GAAA[9-22]																	
	CE9_AAGG[5]GAAA[10]	1		-	NA	-	-	-	-					NA	NA	NA	NA	novel
	CE9_AAGG[6]GAAA[9]	3		+	NA	+	-	-	-					NA	NA	NA	NA	-
	CE10_AAGG[6]GAAA[10]	7		+	NA	+	-	-	+					NA	NA	NA	NA	-
	CE11_AAGG[6]GAAA[11]	18		+	NA	+	+	+	+					NA	NA	NA	NA	-
	CE12_AAGG[6]GAAA[12]	17		+	NA	+	+	+	+					NA	NA	NA	NA	-
	CE13_AAGG[5]GAAA[14]	1		-	NA	-	-	-	-					NA	NA	NA	NA	novel
	CE13_AAGG[6]GAAA[13]	28		+	NA	+	+	+	+					NA	NA	NA	NA	-
	CE14_AAGG[6]GAAA[14]	38		+	NA	+	+	+	+					NA	NA	NA	NA	-
	CE15_AAGG[5]GAAA[16]	1		-	NA	-	-	-	-					NA	NA	NA	NA	novel
	CE15_AAGG[6]GAAA[15]	20		+	NA	+	+	+	+					NA	NA	NA	NA	-
DYS385a,b	CE15_AAGG[8]GAAA[13]	1		-	NA	-	-	-	-					NA	NA	NA	NA	novel
	CE16_AAGG[6]GAAA[16]	17		+	NA	+	+	+	+		+			NA	NA	NA	NA	-
	CE16_AAGG[6GAAA[2]TAAA[1]GAAA[13]	2	ISNP	-	NA	-	-	+	+		+			NA	NA	NA	NA	-

- ✓ STRidER:

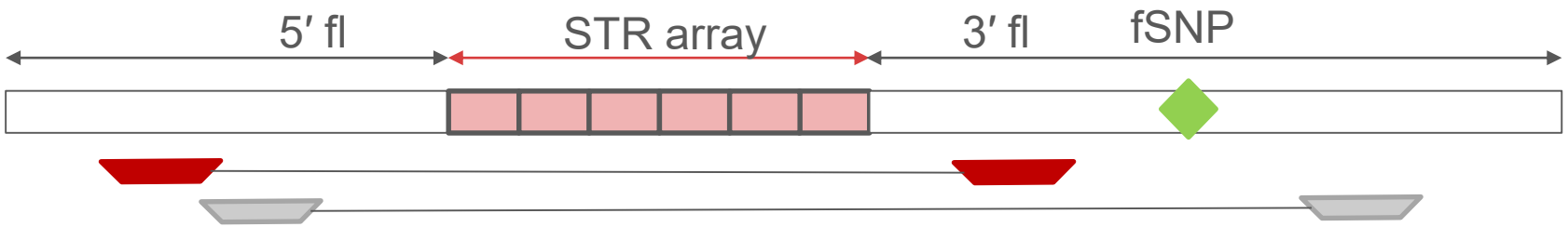
  great QC for CE-based submission,

  but no MPS submission pathway yet

UNIVERSITY OF LEICESTER

Existing databases – Potential issues

✓ STRSeq: GenBank Acc# - unique sequence ID



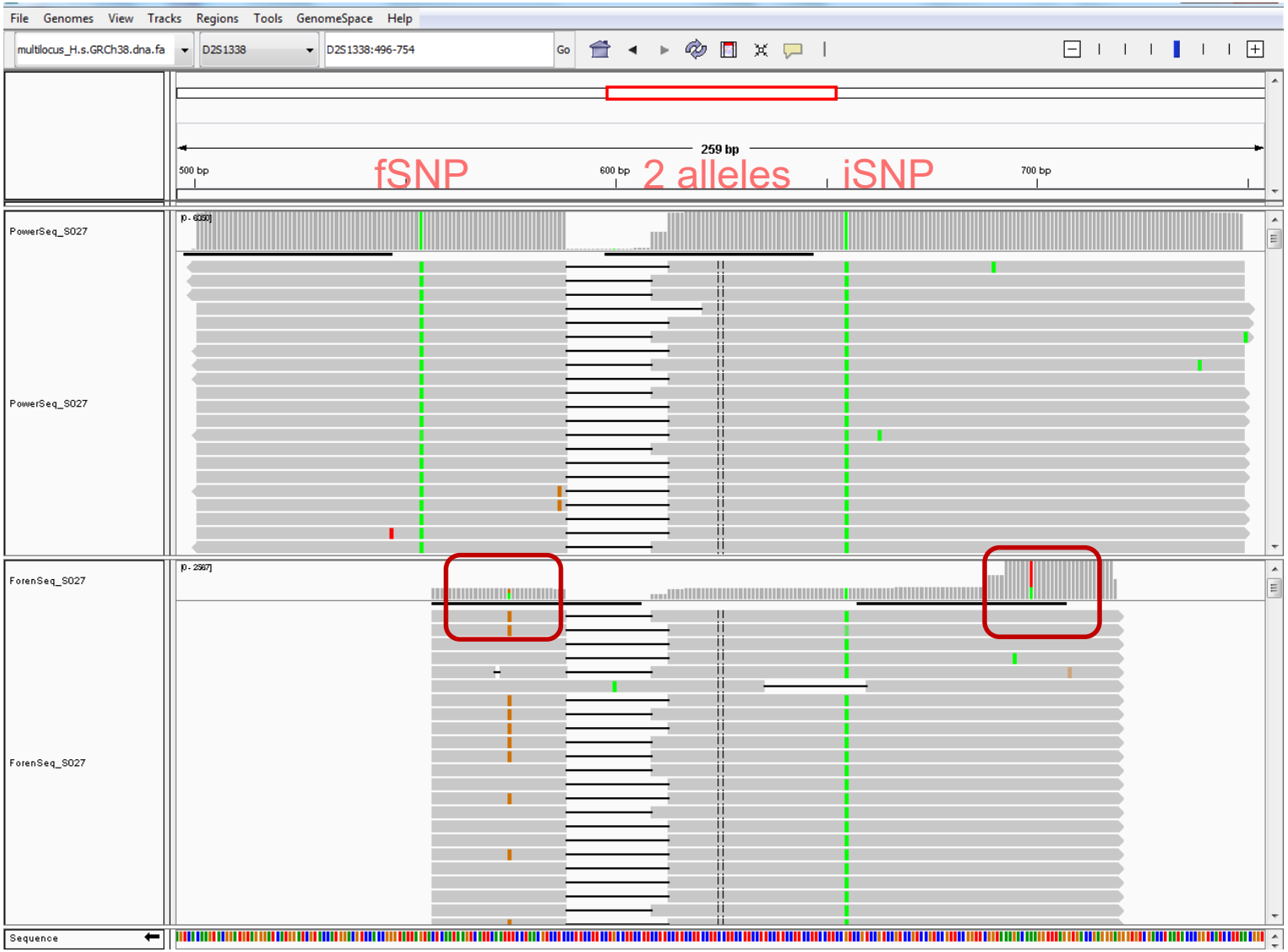
Kit1 STR only – unique ID#1
Kit2 STR + flanking SNP – unique ID#2

D16S539 – same sample on GRCh38, chr16

PowerSeq:CE8_GATA[8] – in STRSeq as MH167241.1

ForenSeq:CE8_GATA[8]_rs11642858 – in STRSeq as MK570017.1

Existing databases – Potential issues



Primer interference – trimming! (Huszar et al. 2019, FSI Gen.)

Integrative Genomics Viewer (IGV) tool (Robinson et al. 2011, Nat. Biotechnol.)

Existing databases – Potential issues

✓ STRSeq: GenBank Acc# - unique sequence ID

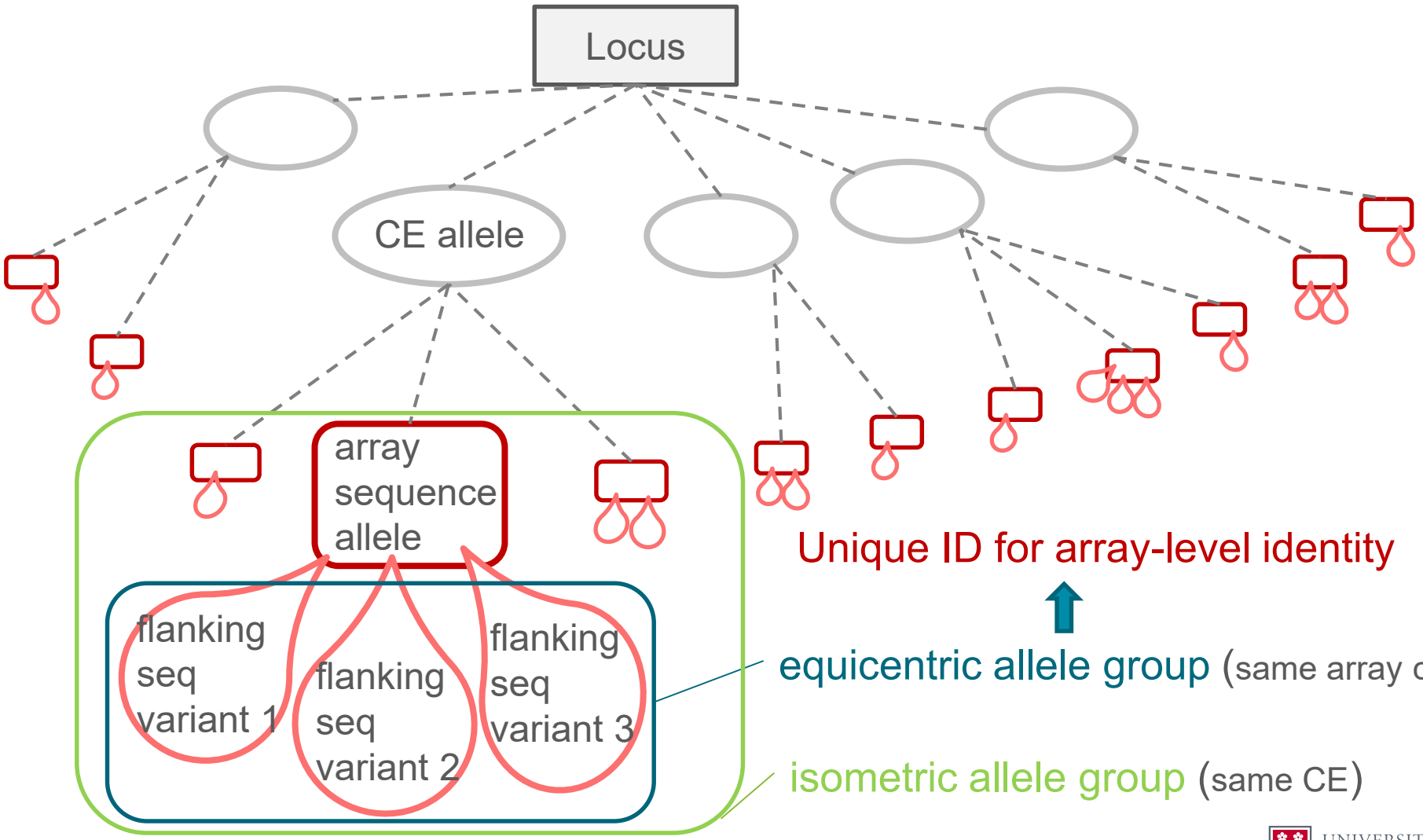
- one DNA type – several unique sequence IDs
- sequence ID groups – based on STR array seq identity
- flanking region difference – exclusion / no match

  real OR kit / software / reported region difference

- unique ID currently at flanking region variants level

  not ideal for automated matches

Unique ID for comparable reporting



Local database - LeiceSTRSeq

- ✓ To help MPS-based projects with redundant tasks:
  - literature search for allele variants
  - allele IDs from STRSeq – by array-level identity groups
  - extra: kit, software, amplified/reported coordinates
    STRSeq ID/publications for reference, annotated string
- ✓ Difficulties:
  - lack of personnel: constant screening, input, cleaning data and development
  - Excel-based user's copy, Access-based background DB

Summary

- ✓ more flexible/sensible array definition – repeat vs SNPs
- ✓ unique ID for STR array-level identity – comparable match
- ✓ flanking variants - with clear reporting coordinates (reference genome, kit type, software)
- ✓ Current state: doable, but not user-friendly
  - central QC
  - curation of submitted new alleles
  - build and maintain database
  - cross-platform searchable interface

Acknowledgements

Mark Jobling

Jon Wetton

BBSRC

Key Forensic Services Limited

Promega

UNIVERSITY OF LEICESTER

High Performance Computing cluster

NUCLEUS Genomics Lab

Other Leicester MPS projects:

UNIVERSITY OF LEICESTER

STRAND working group
align | name | define

ISFG

CE-MPS Discordances

in a study of 31 autosomal STR loci from 498 Spanish individuals

Pedro A. Barrio[1], Pablo Martin[1], Antonio Alonso[1], The DNASEQEX Consortium

[1] Servicio de Biología del Instituto Nacional de Toxicología y Ciencias Forenses (INTCF), Departamento de Madrid
pedro.barrio@justicia.es

INT CIENCIAS FORENSES

DNA SeqEx STR Sequencing & Exchange

Comparison of two MPS platforms:
Ion S5 (*Thermo Fisher Scientific*) MiSeq FGx™ (*Illumina*)

STRs Standardization typing by MPS

International Exchange of MPS data

Population Studies

Madrid

Innsbruck

Berlin

Material & Methods

498 samples

Spanish ancestries representing all the 17 Autonomous Communities of Spain (i.e. “regions”)

Precision ID GlobalFiler® NGS STR Panel v2 on Ion S5 System

Concordance Study CE/MPS

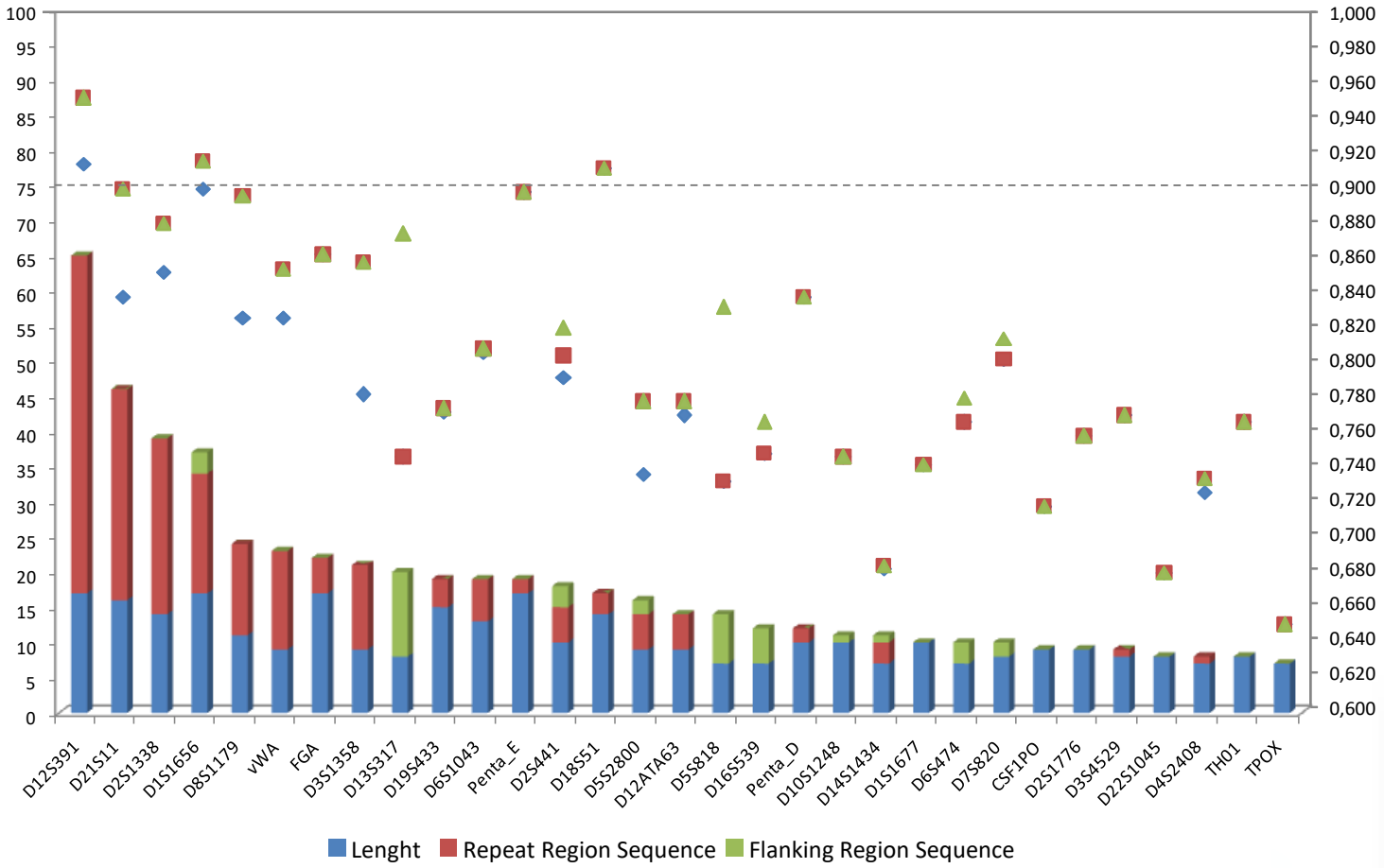
221 samples

PowerPlex Fusion 6C System *(Promega, Madison, WI, USA)*

Results. *Allele frequencies*

STR allelic gains by sequence:

Number of alleles compared to heterozygosity observed for the 31 auSTR *loci*

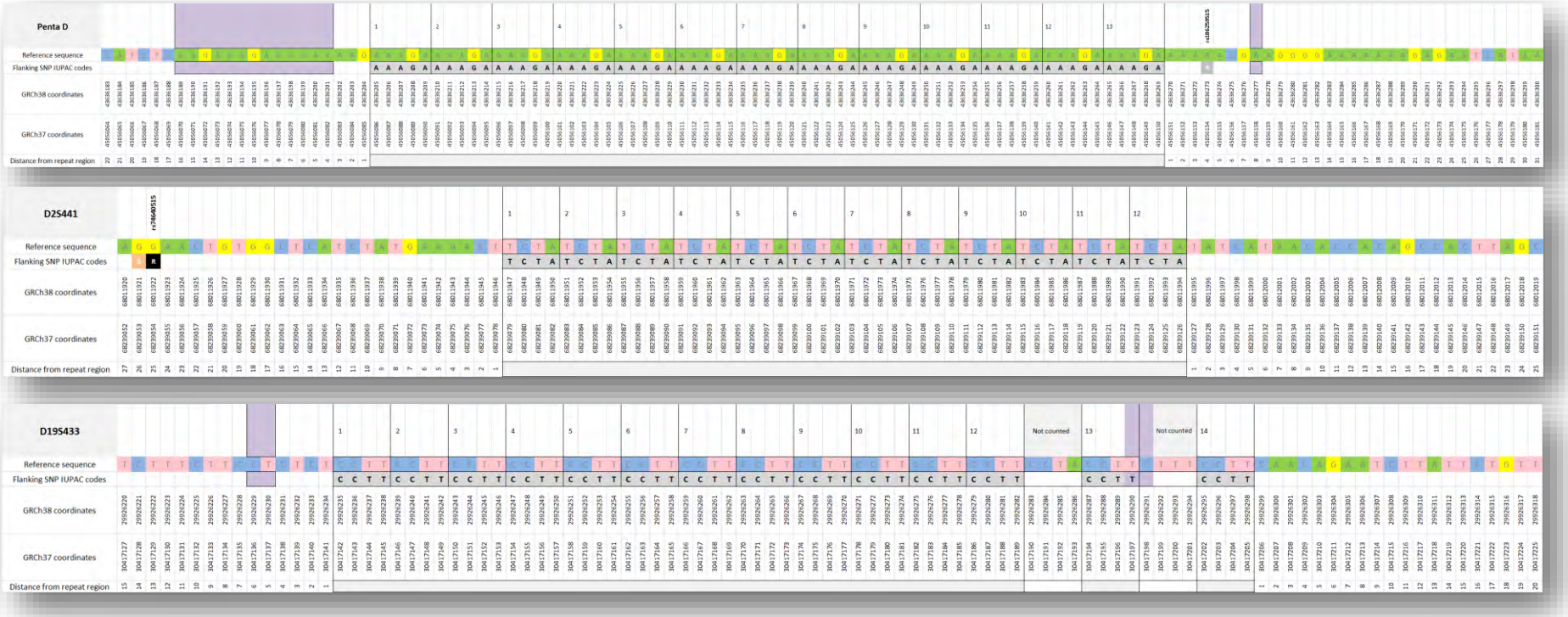


Barrio et al., manuscript in review

Discordances. *Samples*

Putative mismatches:

3 loci: Penta D, D2S441 and D19S433



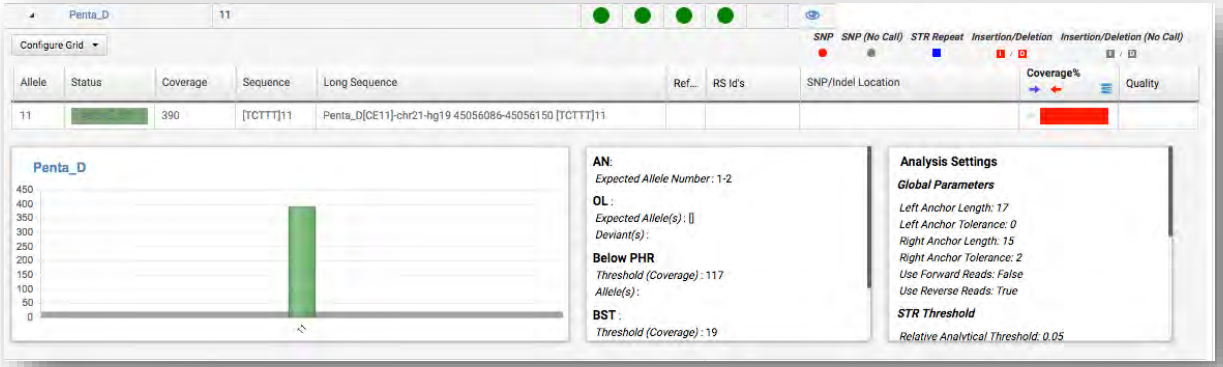
5 samples: 5 samples out of 221 (97.73 % sample concordance)

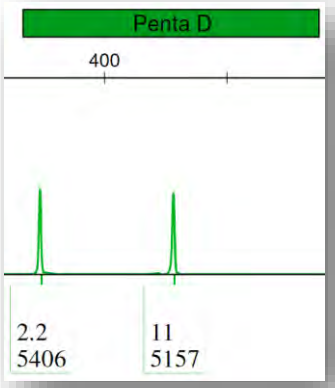
1 locus/sample: 5 loci out of 5083 (99.90 % locus concordance)

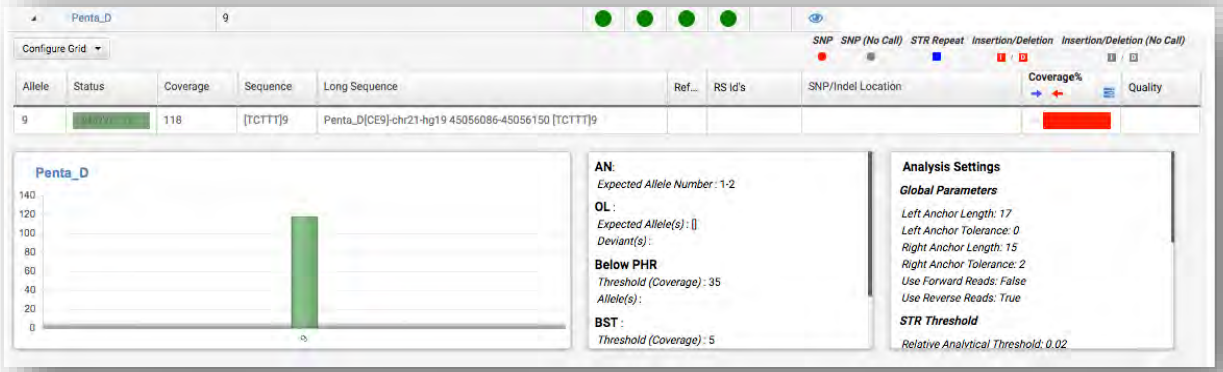
1 allele/locus: 5 alleles out of 10166 (99.95 % allele concordance)

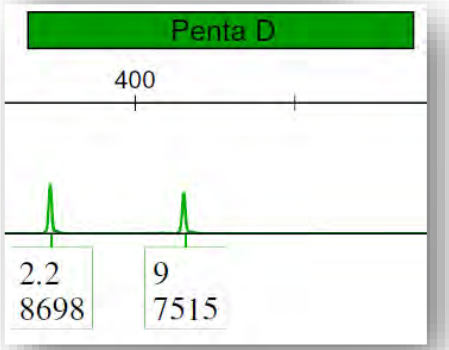
Discordances. *Penta D locus*

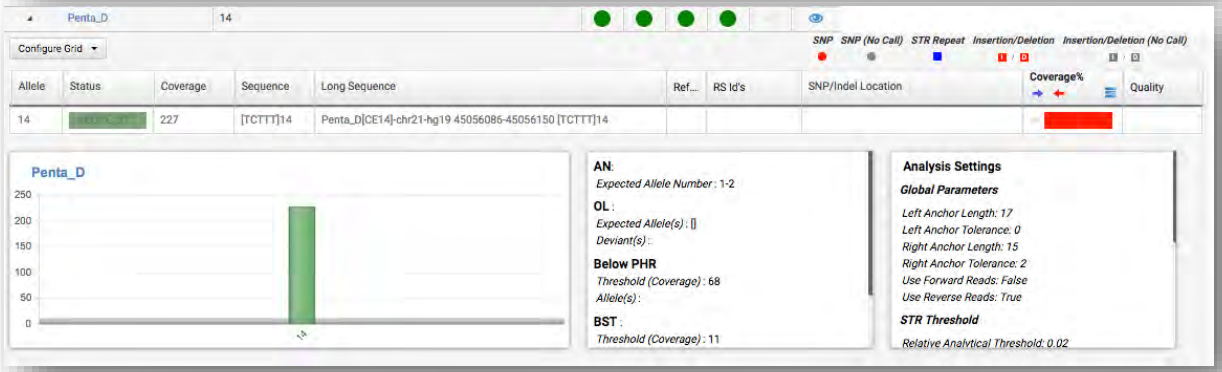
3 samples: M17-06600-47-869-EX-1 , M17-06602-78-882-EX-1 and M17-06602-84-767-EX-1

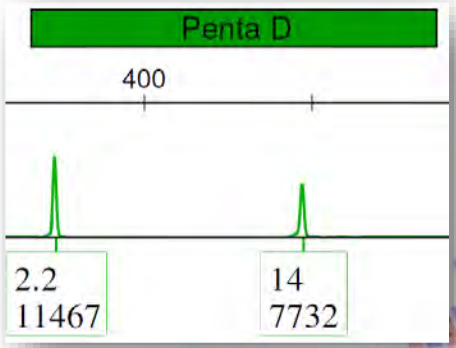






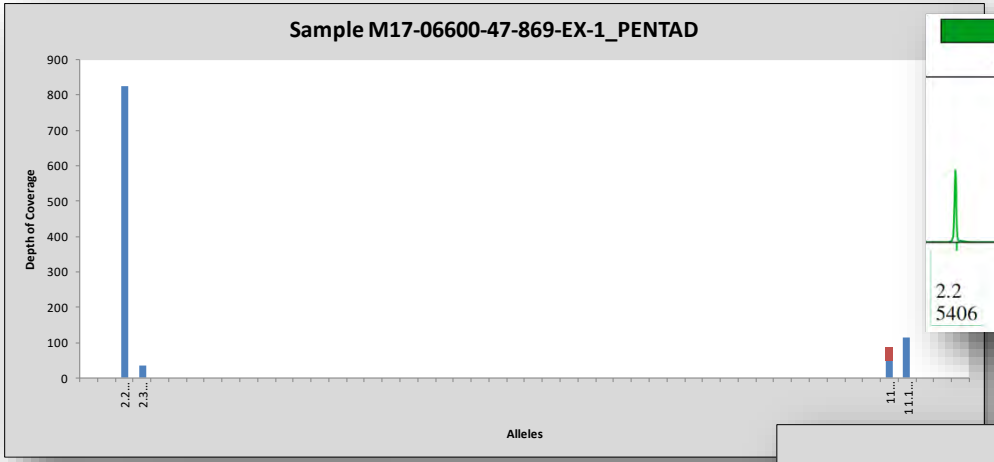


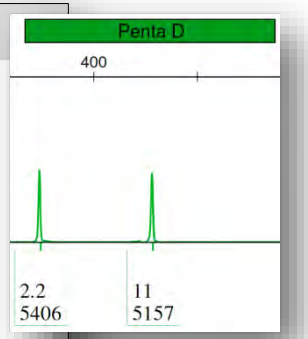


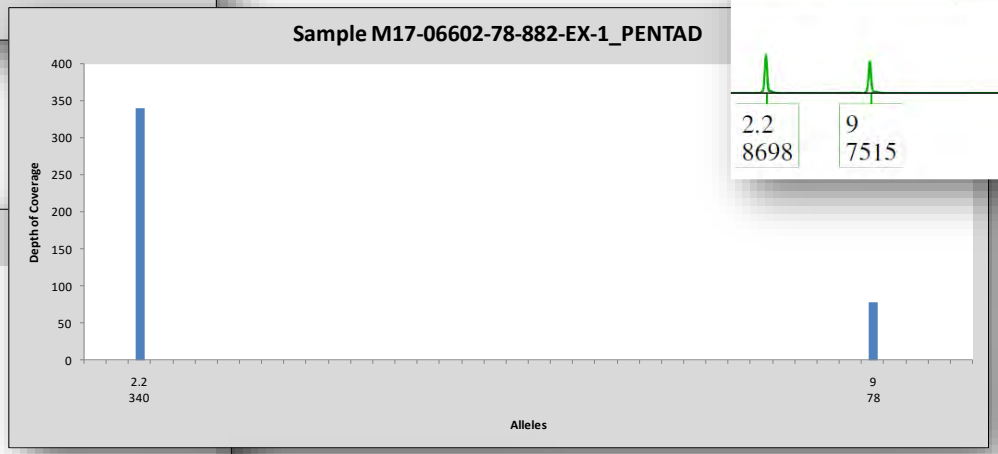


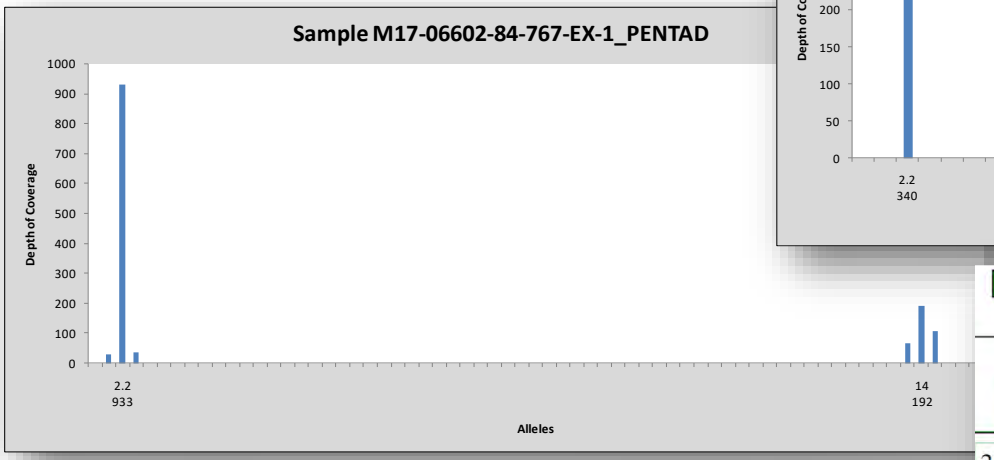
Discordances. *Penta D locus*

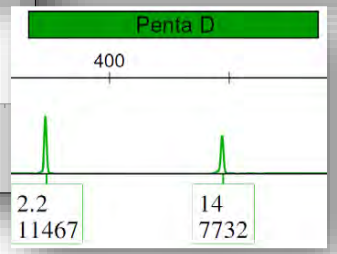
Additional analysis: *STRait Razor v3* (Woerner et al., 2017)









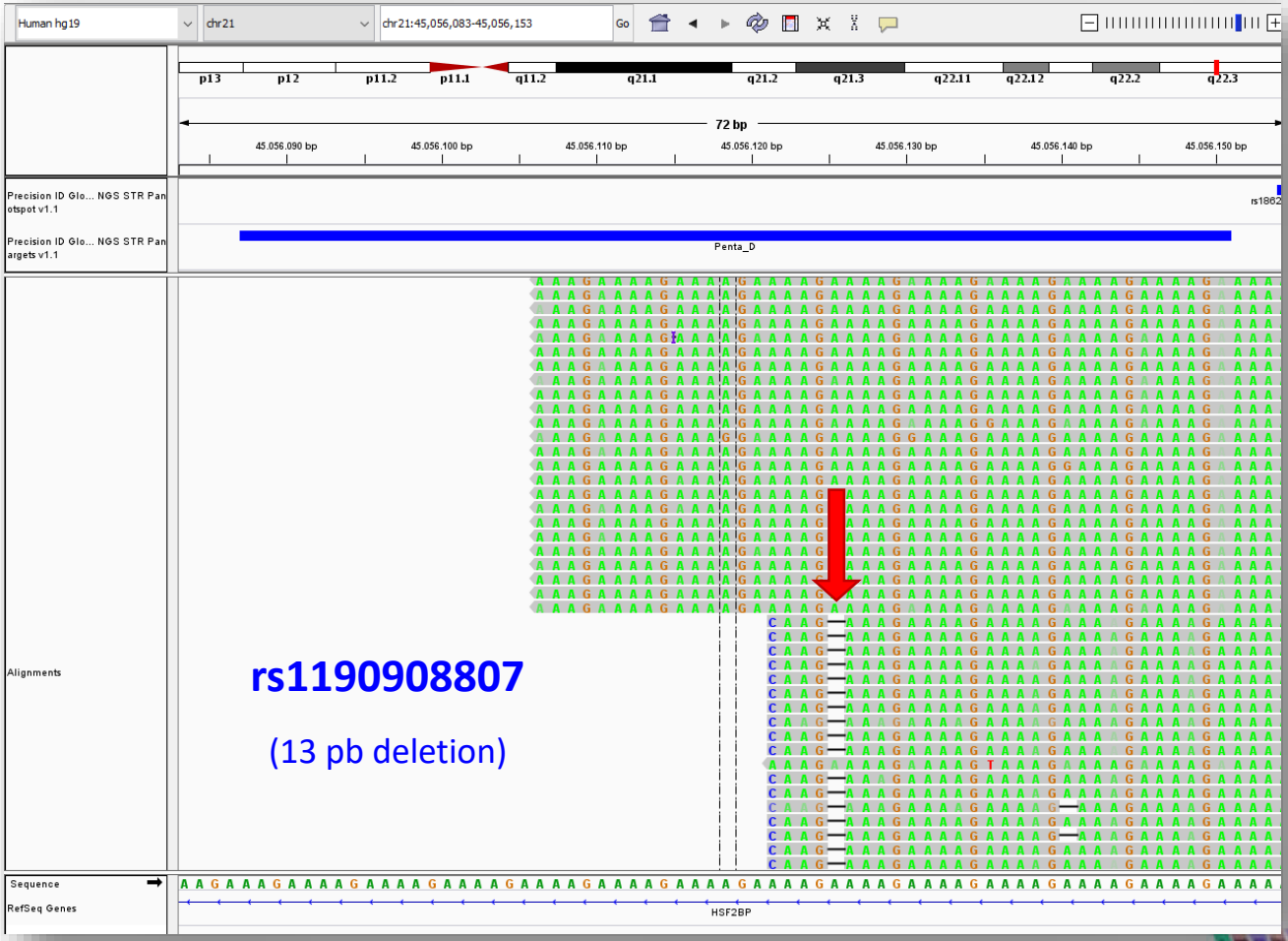


Discordances. *Penta D locus*

Additional analysis: *STRSeq catalog* (Gettings et al., 2017)

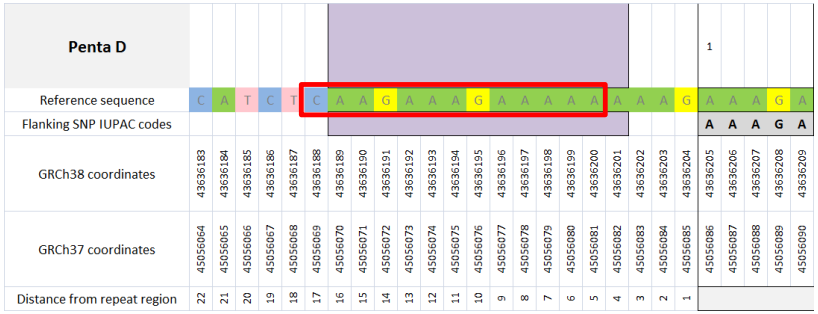
BioProject: **PRJNA380576** Penta_D[CE2.2] → [AAAGA]5 rs1190908807

Integrative Genomics Viewer - *IGV v2.4.16* (Robinson et al., 2011; Thorvaldsdottir et al., 2013)

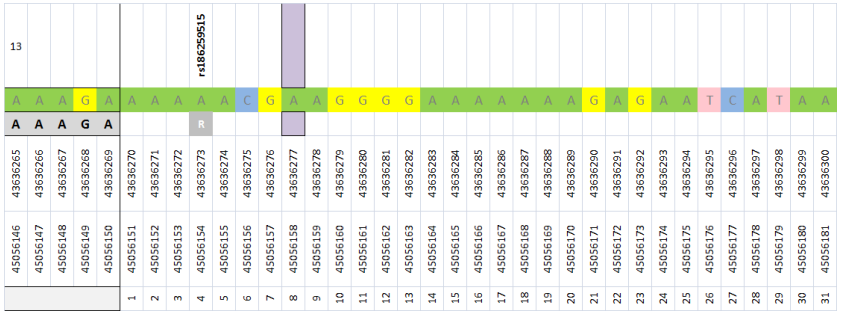


Discordances. *Penta D locus*

Additional analysis: *STRait Razor v3* (Woerner et al., 2017)



(…)



Anchor sequences of .config file

5’ C A A G A A A G A A A A A 3’

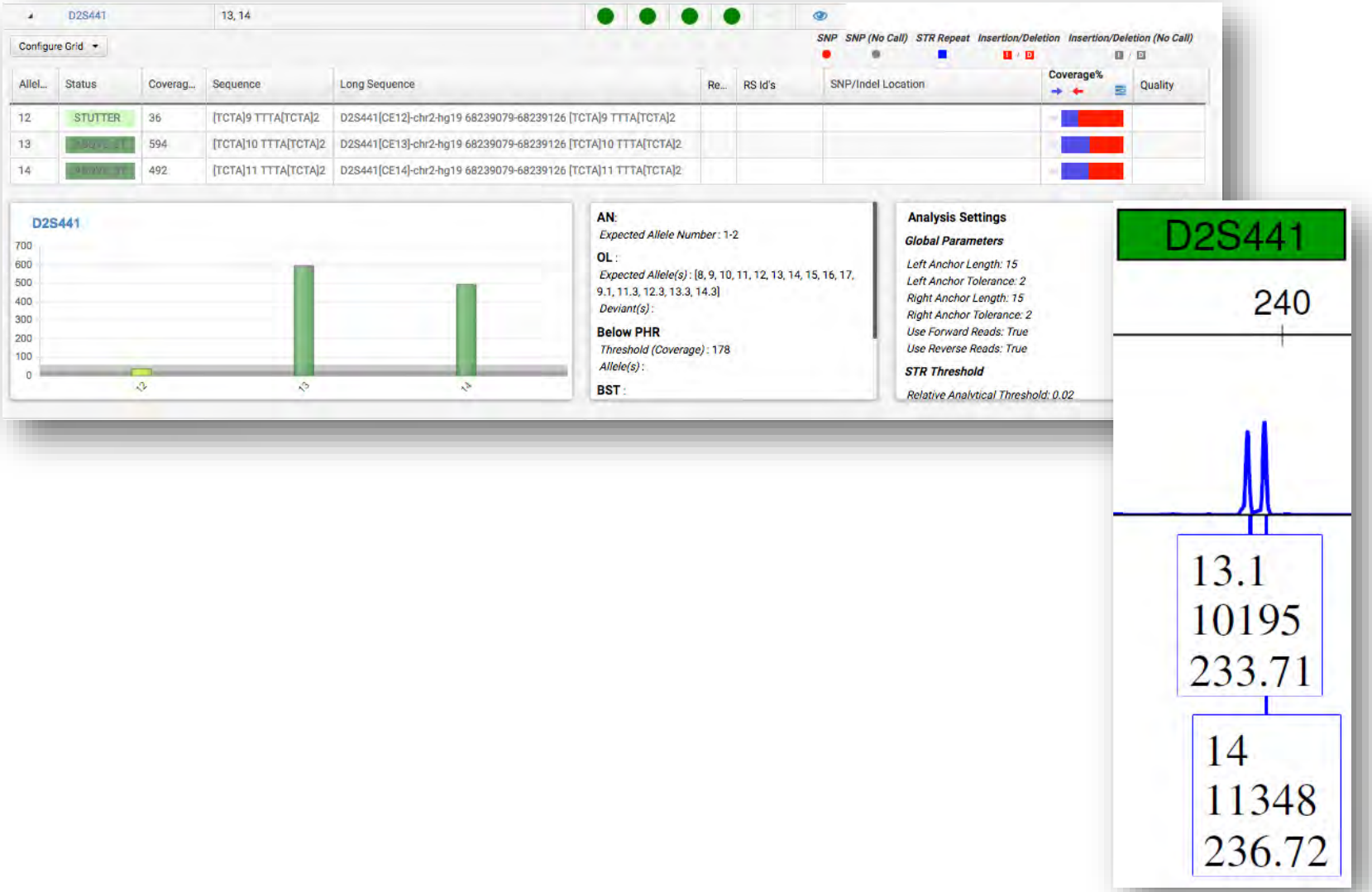
5’ G G G G A A A A A A A G A G A A T 3’

5’ anchor sequence aligns on the deleted region (rs1190908807)

The allele 2.2 is really an allele 5…. with the 13 bp deletion (rs1190908807)

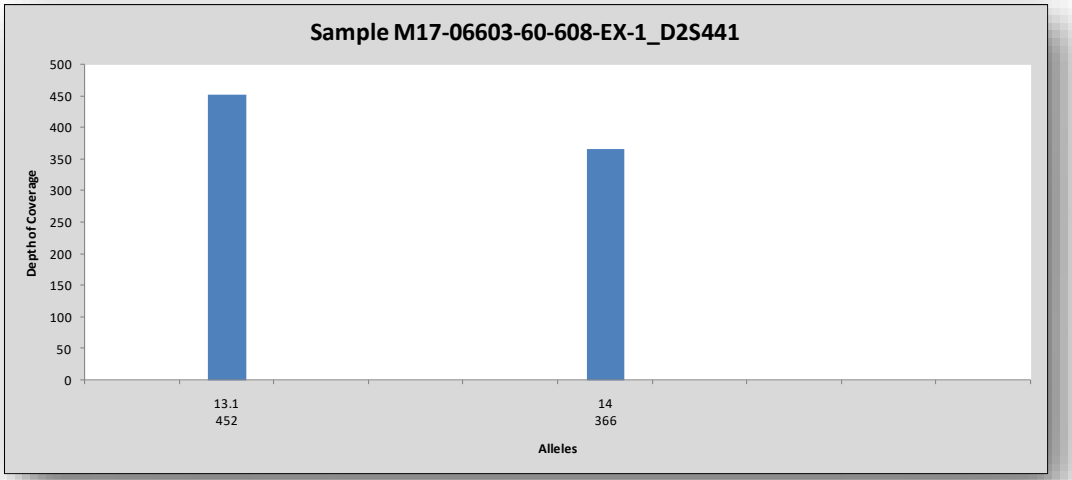
Discordances. *D2S441 locus*

1 sample: M17-06603-60-608-EX-1

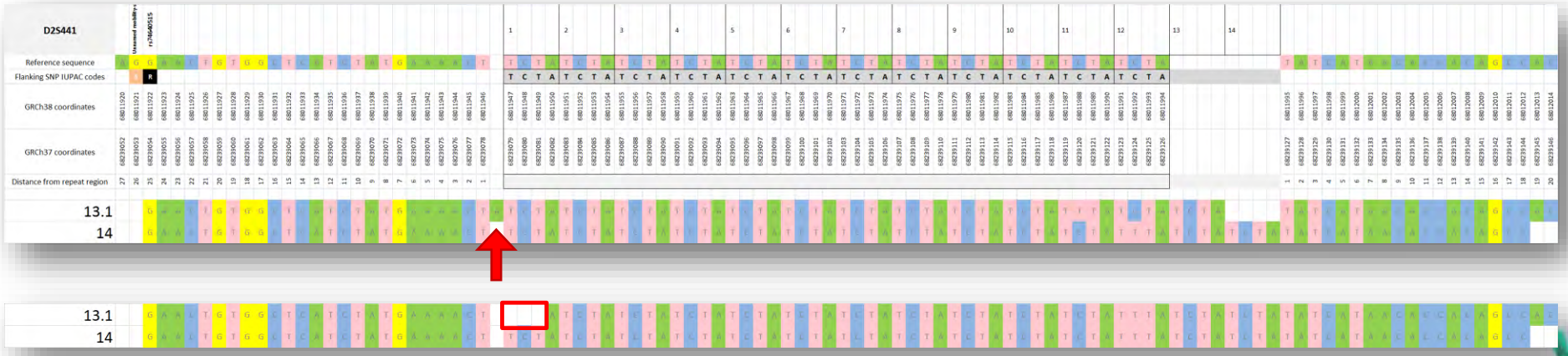


Discordances. *D2S441 locus*

Additional analysis: *STRait Razor v3* (Woerner et al., 2017)



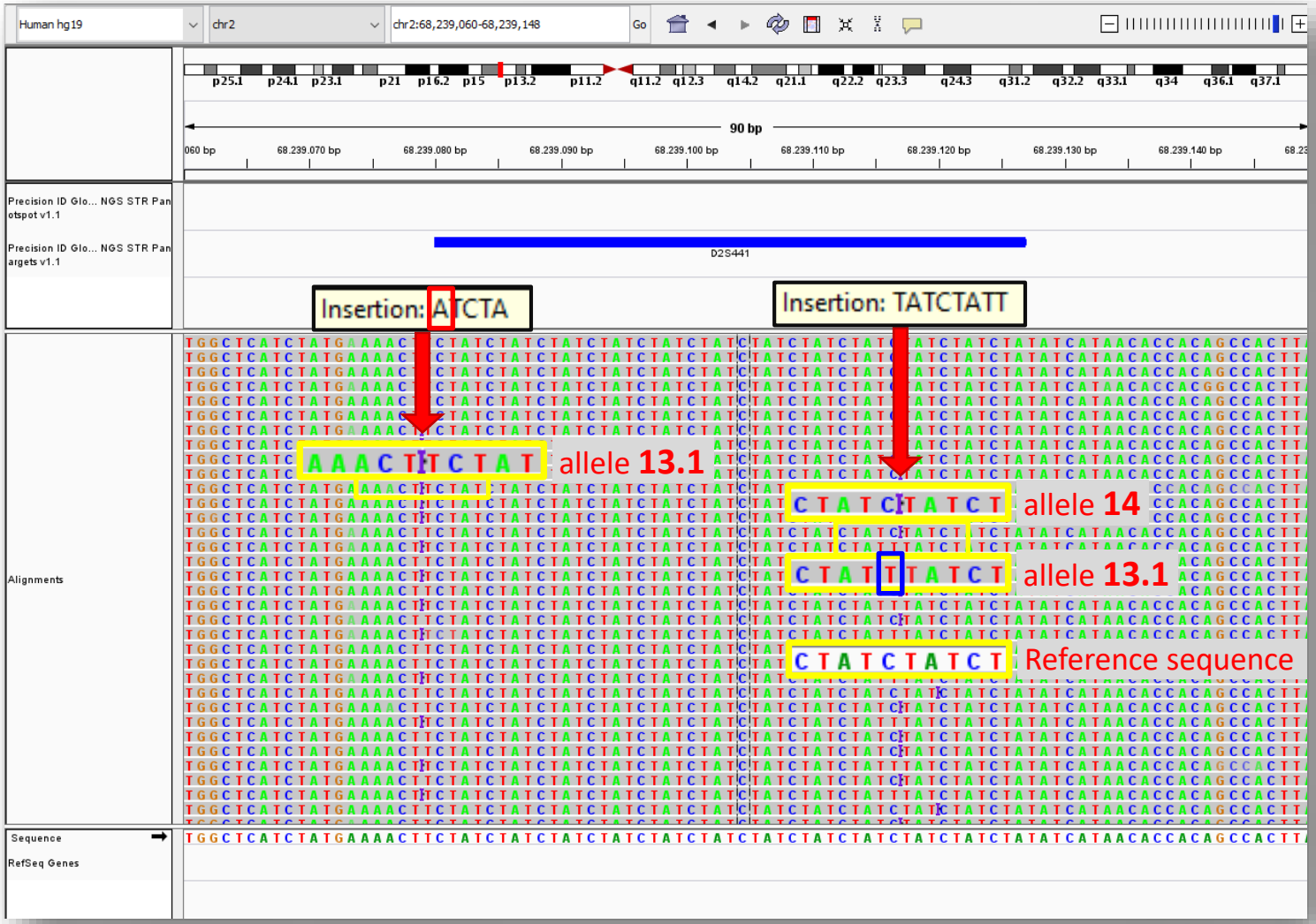
Manual alignments using *updated Forensic STR Sequence Guide v4* (Phillips et al., 2018)



Converge doesn't detect/value the 1 bp insertion inside the repeat unit.

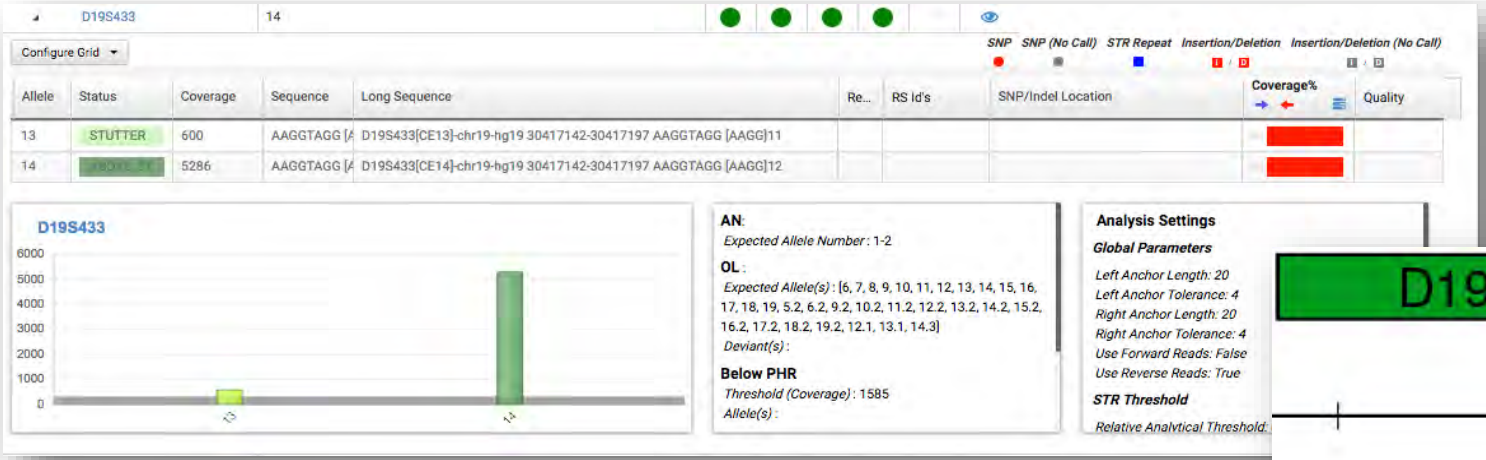
Discordances. *D2S441 locus*

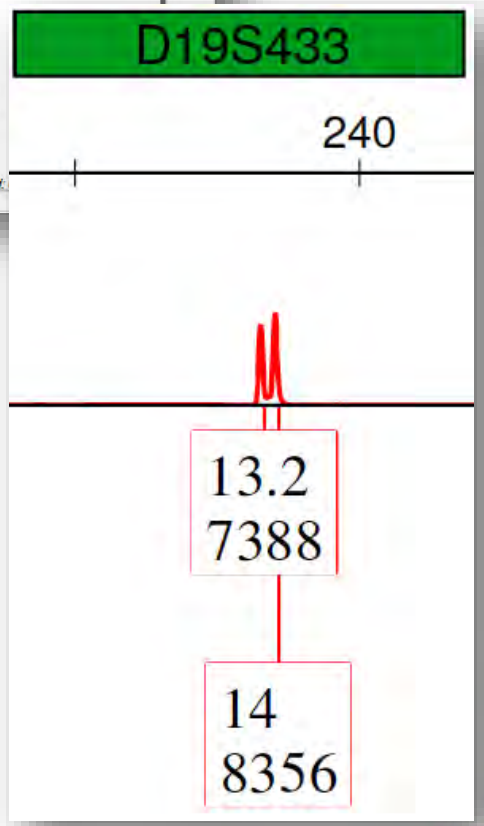
Additional analysis: *IGV v2.4.16* (Robinson et al., 2011; Thorvaldsdottir et al., 2013)



Discordances. *D19S433 locus*

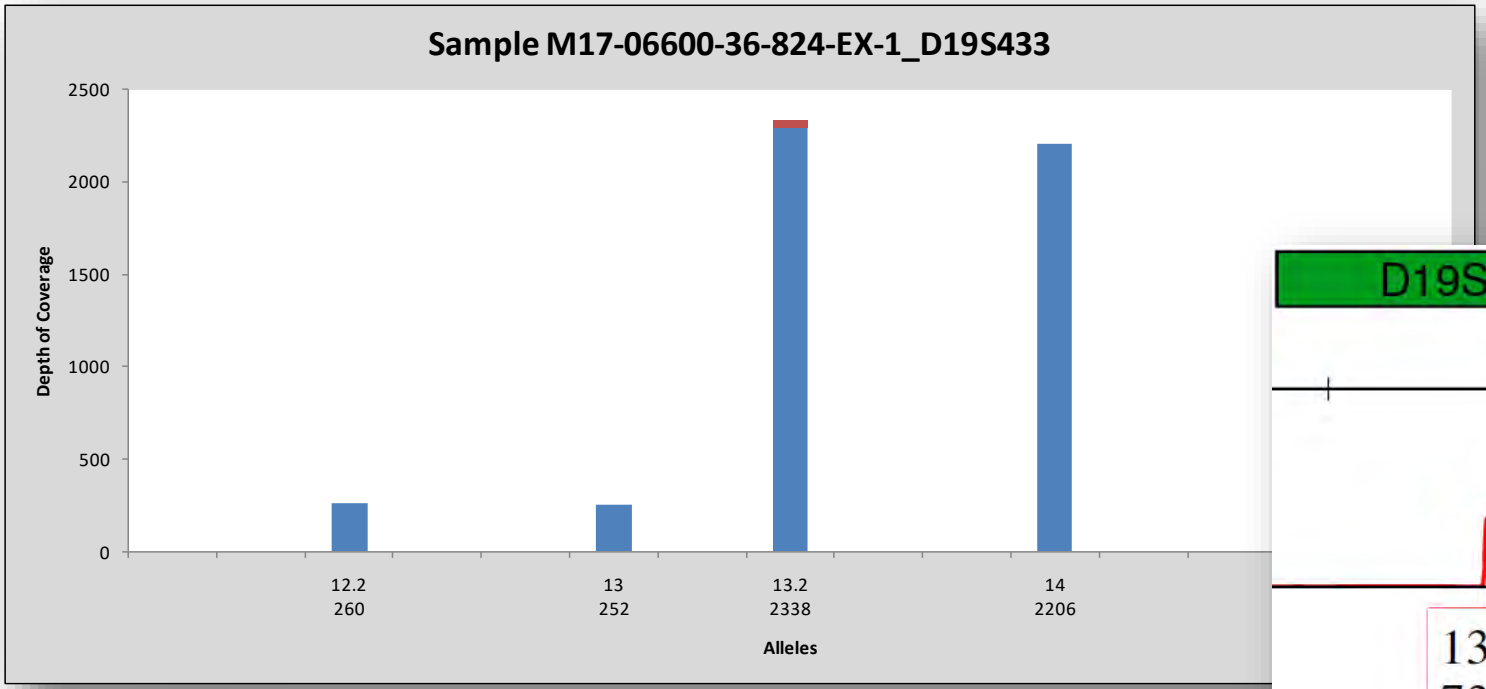
1 sample: M17-06600-36-824-EX-1

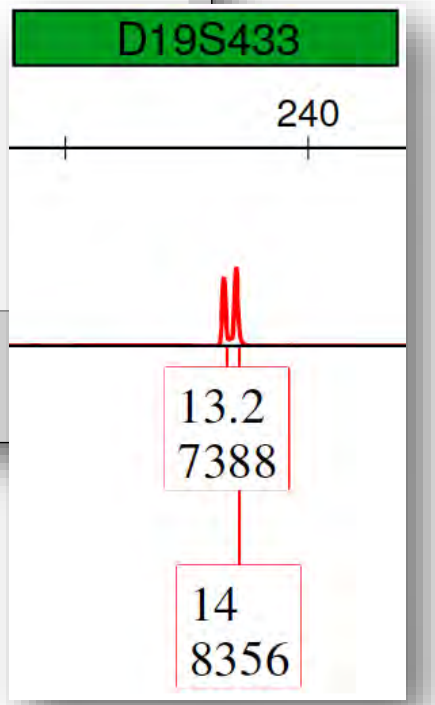




Discordances. *D19S433 locus*

Additional analysis: *STRait Razor v3* (Woerner et al., 2017)



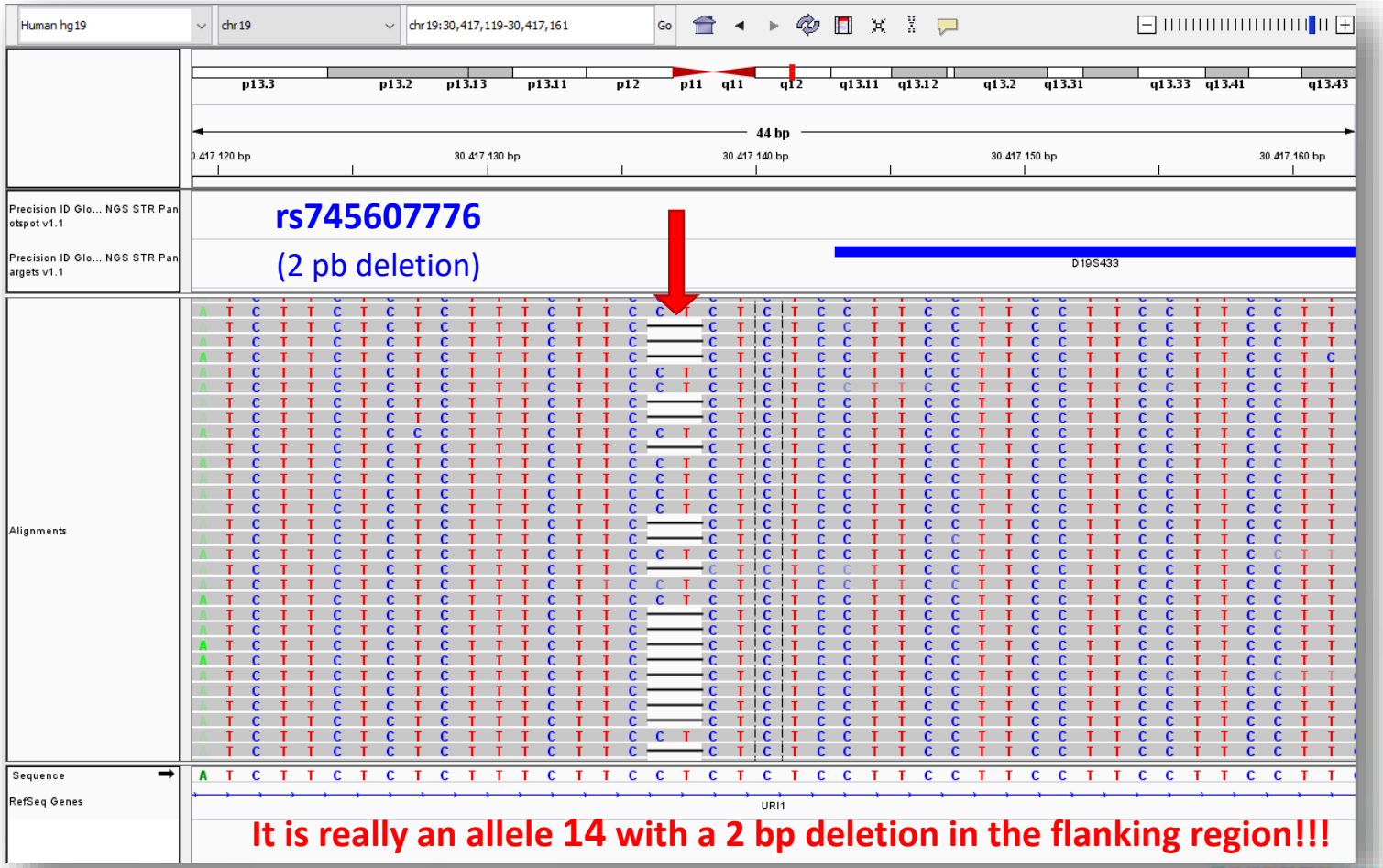


Discordances. *D19S433 locus*

Additional analysis: *STRSeq catalog* (Gettings et al., 2017)

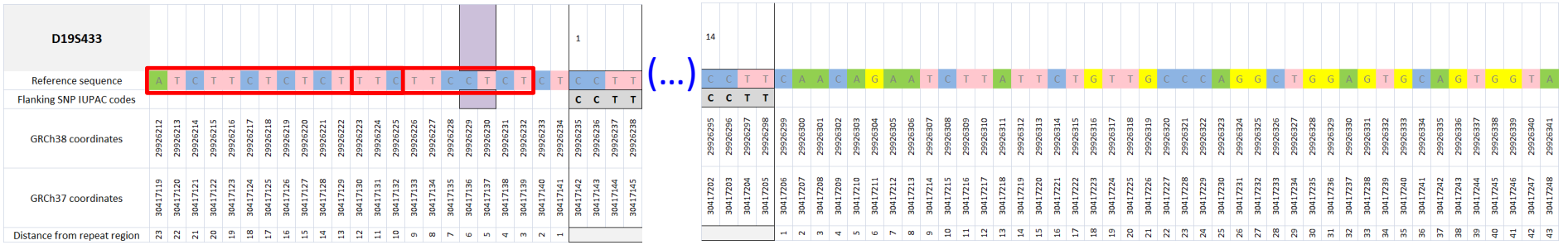
BioProject: **PRJNA380574** D19S433[CE13.2] → [CCTT]12 ccta CCTT cttt CCTT rs745607776

Integrative Genomics Viewer - *IGV v2.4.16* (Robinson et al., 2011; Thorvaldsdottir et al., 2013)



Discordances. *D19S433 locus*

Additional analysis: *STRait Razor v3* (Woerner et al., 2017)



Anchor sequences of .config file

5’ T T C T T C C T C T 3’

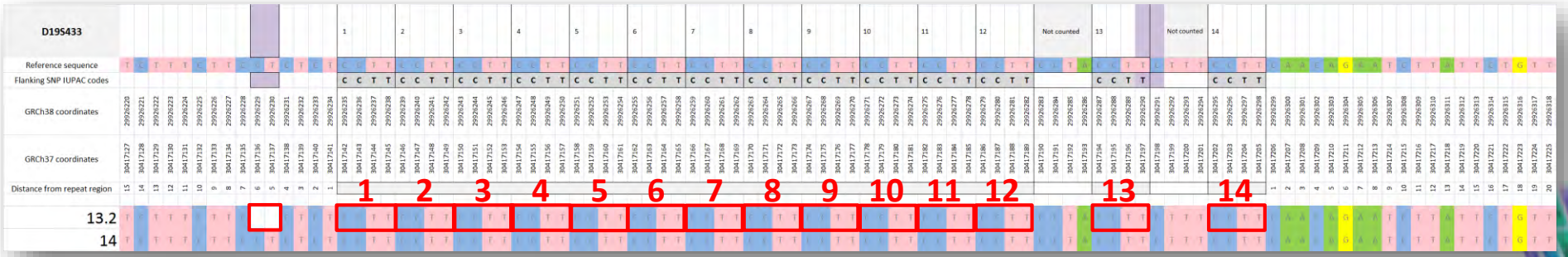
5’ G A G T G C A G T G G T 3’

5’ anchor sequence aligns on the deleted region (rs745607776)

If we move the anchor sequence to save the deleted region...

5’ A T C T T C T C T C T T T C 3’

5’ C A G G C T G G A G T G C A 3’



The allele 13.2 is really an allele 14…. with the 2 bp deletion (rs745607776)

Summary. Highlights

* Findings based on the **CE** (*Fusion 6C*) - **MPS** (*Global NGS*) concordance study:

- Deletions in the flanking region (13 bp deletion in Penta D associated to CE allele 2.2) may cause "**bioinformatic null alleles**" if the selected region matches the anchor region of sequence recognition by the software.

- Deletions and insertions in the flanking regions also generated discrepancies between the **CE micro-variant alleles** and **MPS** data for Penta D, D2S441, and D19S433.

  Penta_D[CE**2.2**]-chr21-hg19 45056086-45056150 **[AAAGA]5** 45056081-45056093-**delAAAAGAAAGAAAA**

  D2S441[CE**13.1**]-chr2-hg19 68239079-68239126 **A [TCTA]10 TTTA [TCTA]2**

  D19S433[CE**13.2**]-chr19-hg19 30417142-30417197 **[CCTT]12** ccta **CCTT** cttt **CCTT** 30417137-3041713-**delGA**

- The knowledge of the MPS sequence demonstrated that **these microvariants were erroneously called by CE**, as they were alleles with complete repeat units plus insertions or deletions in the flanking region.

- The identification and classification of these small discrepancies in the different STR markers will allow software improvement and the development of **better comparison tools between CE and MPS data**.

The DNA SeqEx STR Sequencing & Exchange Faculty

Institute of Legal Medicine, Medical University of Innsbruck. Austria

Institute of Legal Medicine and Forensic Science, Charité - Universitätsmedizin Berlin. Germany

Center for Human Identification at the University of North Texas Health Science Center. USA

National Institute of Toxicology and Forensic Sciences. Madrid Department. Spain

ResearchGate

Follow us on https://www.researchgate.net/project/DNASEQEX

DNA Seq EX
STR Sequencing & Exchange

Thanks for your attention!!

Acknowledgements

This project has been funded with support from the European Commission (grant **HOME/2014/ISFP/AG/LAWX/4000007135** under the Internal Security Funding Police programme of the European Commission---Directorate General Justice and Home Affairs). This publication reflects the views only of the authors, and the European Commission cannot be held responsible for any use which may be made of the information contained therein.

The authors would like to thank members of LIMS Administrators Team of the General Subdirectorate of New Technologies of Justice (SGNTJ) of the Ministry of Justice (Spain) for their helpful technical support.

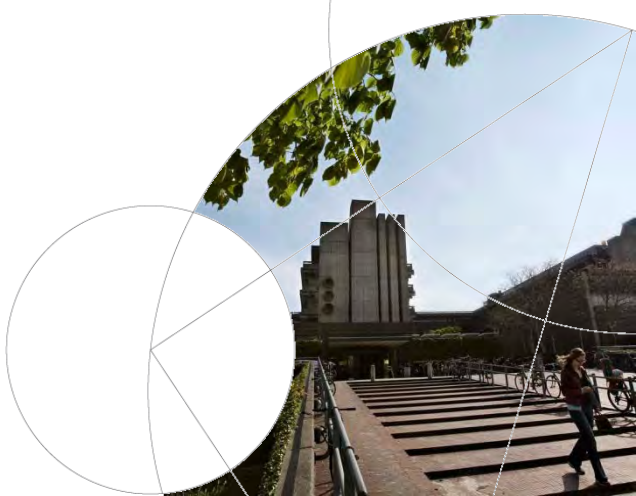
Thermo Fisher Scientific Inc. Provider of *Precision ID GlobalFiler® NGS STR Panel v2* and *Converge v2.1* software. The authors would like to thank Matt Phipps for technical support.

Faculty of Health and Medical Sciences

STR sequencing in Copenhagen

Claus Børsting MSc, PhD
Senior researcher

Section of Forensic Genetics
Department of Forensic Medicine
Faculty of Health and Medical Sciences
University of Copenhagen
Denmark



STR sequencing in Copenhagen

A bit of history

Case reporting

NGS nomenclature issues

STR sequencing in Copenhagen - history

2011: 454 GS Junior (Roche)

Fordyce et al., (2011) Biotechniques 51, 127-133.
Rockenbauer et al., (2014) FSI genet. 8, 68-72.
Dalsgaard et al., (2014) FSI genet. 8, 195-199.
Gelardi et al., (2014) FSI genet. 12, 38-41.

2014: Ion PGM (Thermo Fisher Scientific)

Fordyce et al., (2015) FSI genet. 14, 132-140.
Friis et al., (2016) FSI genet. 21, 68-75.
Vilsen et al., (2017) FSI genet. 28, 82-89.

2015: MiSeq FGx (Verogen)

Hussing et al., (2018) Foren. Sci. Res. 3, 111-123.
Vilsen et al., (2018) FSI genet. 35, 107-112.
Hussing et al., (2019) Int. J. Legal Med. 133, 325-334.
Simayijiang et al., (2019) in preparation.
DNASEQEX 29 Y-STR panel test (2019).

2016: Ion S5 System (Thermo Fisher Scientific)

Precision ID Globalfiler mixture ID test (2017).
Project Iceberg: Precision ID Globalfiler NGS STR panel test (2018-19).

STR sequencing in Copenhagen - history

Nomenclature*:

- Locus name (used in forensic genetics)
- Length of repeat region/length of subunit
- Sequence(s) of subrepeat(s) followed by the number of repeats
- Variation in the flanking regions

Examples:

TH01[9] AATG[9]

D5S818[12] AGAT[9]ACAT[1]AGAT[2]rs25768[T]rs73801920[G]

DXS10135[23] AAGA[3]GAAAGGA[1]AAGA[19]AAAG[1]del:9306454-6

Python script: STRinNGS‡

*Gelardi *et al*., FSI genet. (2014) 12, 38-41

‡ Friis et al., (2016) FSI genet. 21, 68-75.

Case reporting in Copenhagen - history

Forensic Genetics in Denmark

- Accredited according to ISO 17025
- University based service
- Independent of the State administration, police, and judiciary system

Report everything we detect

- Unless the quality of the results prevent it
- Example: LPL in the FFFL panel
- Example: rs7520386 and rs576261 in the Precision ID Identity Panel

STR sequencing nomenclature

Examples:

TH01[9] AATG[9]

D5S818[12] AGAT[9]ACAT[1]AGAT[2]rs25768[T]rs73801920[G]

DXS10135[23] AAGA[3]GAAAGGA[1]AAGA[19]AAAG[1]del:9306454-6

STRidER nomenclature*:

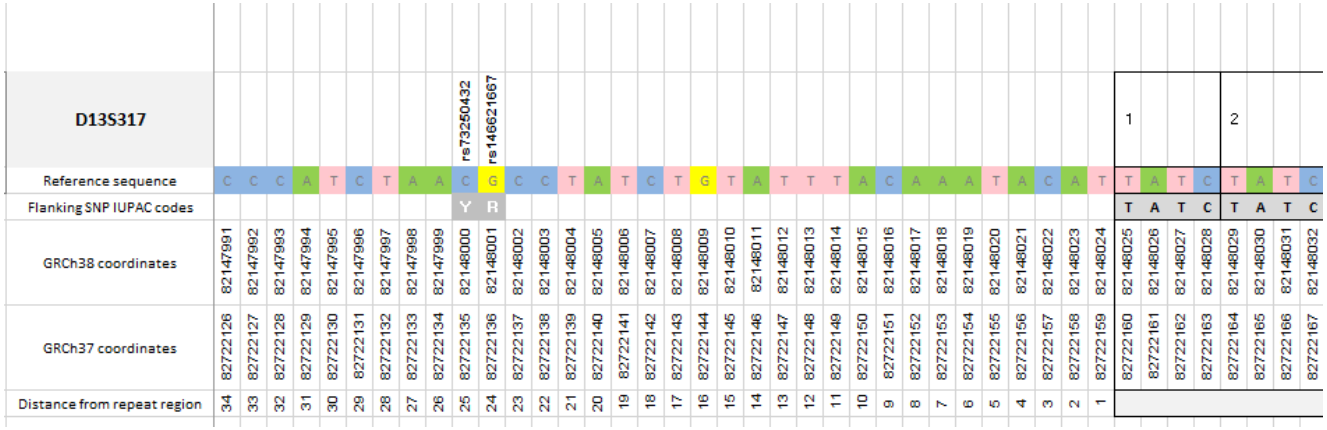
TH01 [AATG]9

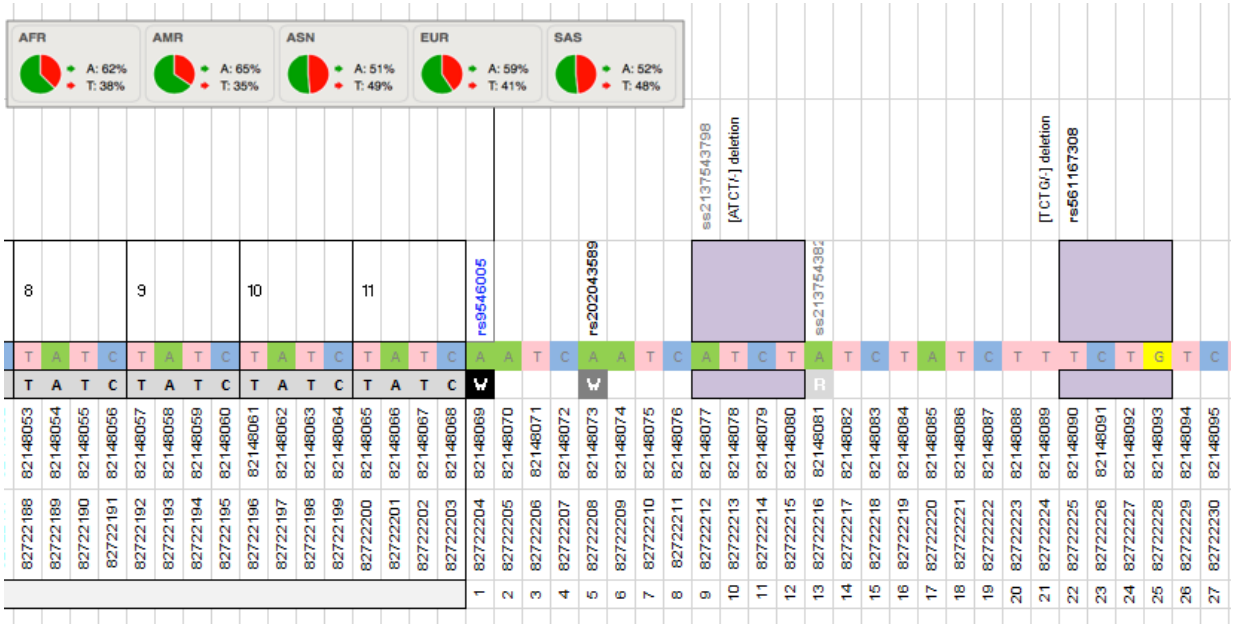
D5S818 [ATCT]9 ATGT [ATCT]2 rs25768[T] rs73801920[G]

DXS10135 [AAGA]3 gaaagga [AAGA]19 AAAG del:9306454-6

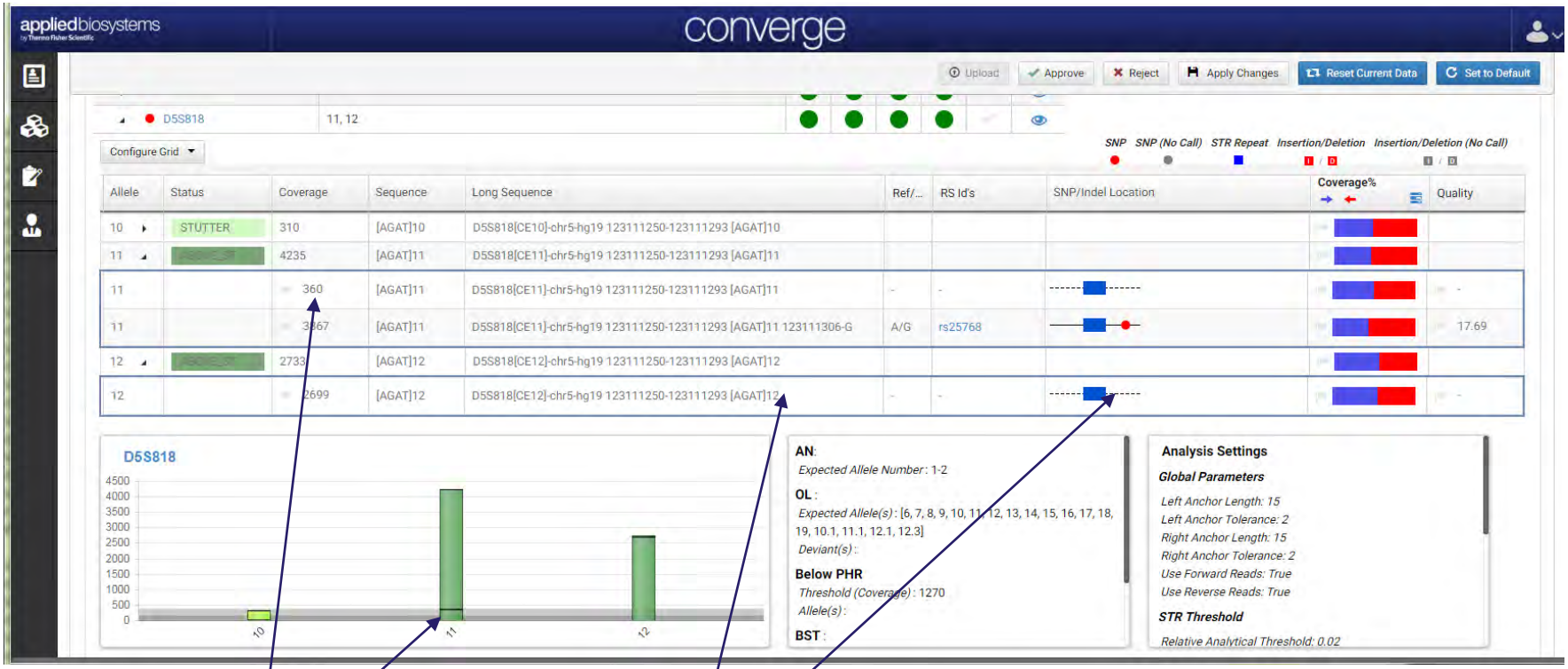
*Bodner et al., (2016) FSI genet. 24, 97-102.
Phillips et al., (2018) FSI genet. 34, 162-169.

STR sequencing nomenclature





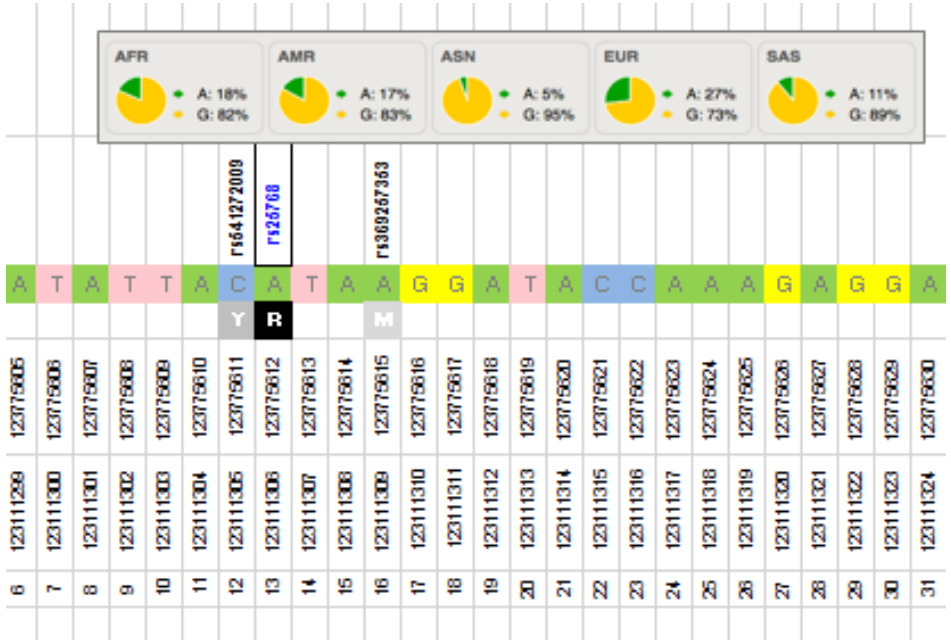
STR sequencing nomenclature



360 stutter reads

What is the SNP-STR haplotype

STR sequencing nomenclature



Summary

Nomenclature

STR region (clear definitions in STRidER)

- Start and end (genome build)
- Forward strand
- Subrepeat format

Flanking regions

- InDels should be incorporated for back-compatibility with CE
  - PCR primers should be made available (old and new kits)
- SNPs?
  - Global nomenclature for SNP-STR - which SNPs?
  - Restrict multiplex design
  - Challenge sequencing length

Database in place (STRSeq)

Thank you for the invitation

MixtureAce
. . .For Analysis of Mixed MPS Data

THE SEQUENCE IDENTIFIER (SID) OPERATIONAL NOMENCLATURE FOR MIXED-DNA CASEWORK USING MPS

STRAND Working Group Nomenclature Meeting

11 April 2019

Brian Young, Ph.D.

brian@nichevision.com

NicheVision Forensics, Inc.

NicheVision Forensics, LLC

Outline

1. Motivation
2. The Challenge of MPS Artifacts in Mixed DNA Interpretation
3. Sequence Identifier (SID) Nomenclature
4. Why We Need SID: Labeling of Artifacts by ‘Phylogeny’
5. Mixture Interpretation of Sequence-Based MPS Data Using MixtureAce™ Plugin to ArmedXpert™ Software

Motivation

- Casework by MPS Has Been Hindered by a Lack of Available Methods for Interpreting Mixed DNA Samples
- We Developed a Mixed DNA Interpretation Method But it Requires an ‘Operational’ Nomenclature

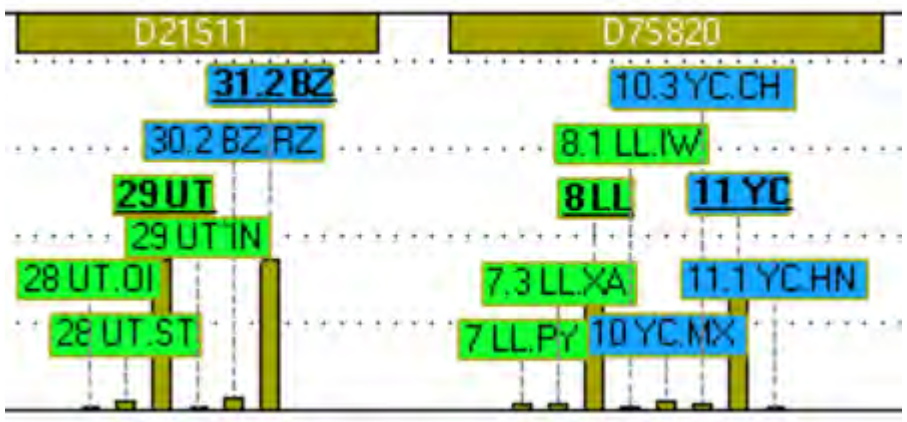
Outline

1. Motivation
2. The Challenge of MPS Artifacts in Mixed DNA Interpretation
3. Sequence Identifier (SID) Nomenclature
4. Why We Need SID: Labeling of Artifacts by ‘Phylogeny’
5. Mixture Interpretation of Sequence-Based MPS Data Using MixtureAce™ Plugin to ArmedXpert™ Software

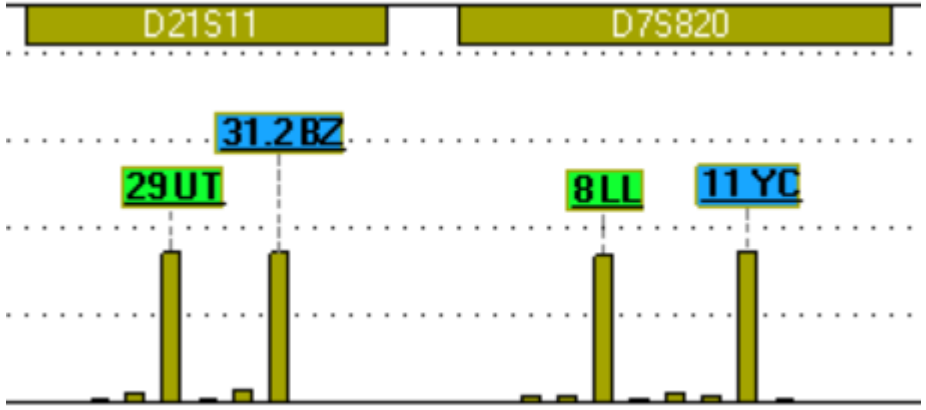
Challenges

1. Space in Computer Displays is Limited
2. Both Artifacts and Alleles Must be Labeled In Mixture Analysis
3. Artifacts are Errors, Not Polymorphisms (rs# Not Available)
4. Universal: Every Laboratory Gets Equivalent Labels for Equivalent Sequences
5. Sequence Type Labeling That is Easy to Vocalize When Discussing Profiles

Example Screen Captures from ArmedXpert/MixtureAce



Show Stutter and Non-Stutter Artifacts



Filter Stutter and Non-Stutter Artifacts

2800M sample 141, PowerSeq, Syracuse Univ.

Any Sequence is a Potential Allele in Mixtures
All Sequences Must be Labeled, and Attributed

MixtureAce
. . .For Analysis of Mixed MPS Data

Marker	SID	Count	STR	Amplicon
D2S1338		10	[GGAA]2GGAC[GGAA]10[GGCA]7	GGATTGCAGGAGGGAAGGAAGGACGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGGCAAGGCCAAGCCATTTCTGTT
		183	[GGAA]2GGAC[GGAA]11[GGCA]7	GGATTGCAGGAGGGAAGGAAGGACGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGGCAAGGCCAAGCCATTTCTGTT
		33	[GGAA]2GGAC[GGAA]12[GGCA]6	GGATTGCAGGAGGGAAGGAAGGACGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGGCAAGGCCAAGCCATTTCTGTT
	22 ZU	2000	**[GGAA]2GGAC[GGAA]12[GGCA]7**	**GGATTGCAGGAGGGAAGGAAGGACGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGGCAAGGCCAAGCCATTTCTGTT** ALLELE
		22	[GGAA]2GGAC[GGAA]12[GGCA]6GACA	GGATTGCAGGAGGGAAGGAAGGACGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGACAAGGCCAAGCCATTTCTGTT
		12	[GGAA]2GGAC[GGAA]5AGAA[GGAA]6[GGCA]7	GGATTGCAGGAGGGAAGGAAGGACGGAAGGAAGGAAGGAAGGAAAGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGGCAAGGCCAAGCCATTTCTGTT
		20	[GGAA]2GGAC[GGAA]13[GGCA]7	GGATTGCAGGAGGGAAGGAAGGACGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGGCAAGGCCAAGCCATTTCTGTT
		221	[GGAA]2GGAC[GGAA]14[GGCA]7	GGATTGCAGGAGGGAAGGAAGGACGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGGCAAGGCCAAGCCATTTCTGTT
		30	[GGAA]2GGAC[GGAA]15[GGCA]6	GGATTGCAGGAGGGAAGGAAGGACGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGGCAAGGCCAAGCCATTTCTGTT
	25 MB	1766	**[GGAA]2GGAC[GGAA]15[GGCA]7**	**GGATTGCAGGAGGGAAGGAAGGACGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGGCAAGGCCAAGCCATTTCTGTT** ALLELE
		20	[GGAA]2GGAC[GGAA]15[GGCA]7	GGATTGCAGGAGGGAAGGAAGGACGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGGCAAGGCCAAGCCATTTCTGCT
		17	[GGAA]2GGAC[GGAA]15[GGCA]7	GGATTGCAGGAGGGAAGGAAGGACGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGGCAAGGCCAAGCCATGTCTGTT
		16	[GGAA]2GGAC[GGAA]15[GGCA]7	GGATTGCAGGAGGGAAGGAAGGACGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGGCAAGGCCAAGCCATCTCTGTT
		10	[GGAA]2GGAC[GGAA]15[GGCA]7	GGATTGCAGGAGGGAAGGAAGGACGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGGCAAGGCCAAGCCATTTCTGAT
D3S1358		22	[TCTA]1[TCTG]3[TCTA]11	TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTATCTATCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACAGGGTCTTGCTC
		298	[TCTA]1[TCTG]3[TCTA]12	TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTATCTATCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACAGGGTCTTGCTC
	17 KW	3947	**[TCTA]1[TCTG]3[TCTA]13**	**TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTATCTATCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACAGGGTCTTGCTC** ALLELE
		32	[TCTA]1[TCTG]3[TCTA]13	TTTGGGGGCATCTCTTATACTTATGAAATCAACAGAGGCTTGCATGTATCTATCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACAGGGTCTTGCTC
		23	[TCTA]1[TCTG]3[TCTA]13	TTTGGGGGCATCCTTATACTCATGAAATCAACAGAGGCTTGCATGTATCTATCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACAGGGTCTTGCTC
		12	[TCTA]1[TCTG]3[TCTA]13	TTTGGGGACATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTATCTATCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACAGGGTCTTGCTC
	18 MI	2856	**[TCTA]1[TCTG]3[TCTA]14**	**TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTATCTATCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACAGGGTCTTGCTC** ALLELE
		21	[TCTA]1[TCTG]2[TCTA]15	TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTATCTATCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACAGGGTCTTGCTC
		12	[TCTA]1[TCTG]4[TCTA]13	TTTGGGGGCATCTCTTATACTCATGAAATCAACAGAGGCTTGCATGTATCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACAGGGTCTTGCTC
		15	[TCTA]1[TCTG]3[TCTA]14	TTTGGGGGCATCTCTTATACTCATGAAATCAACAGGGGCTTGCATGTATCTATCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACAGGGTCTTGCTC

Sample: 2800M/ForenSeq
AT = 10 Reads
Source: OCME

Expressed by Length		
Marker	**Length**	**Count**
D2S1338	20	10
	21	216
	22	**2034**
	23	20
	24	251
	25	**1829**
D3S1358	15	22
	16	298
	17	**4014**
	18	**2904**

Apply 1.5% AT		
Marker	**Length**	**Count**
D2S1338	21	216
	22	**2034**
	24	251
	25	**1829**
D3S1358	16	298
	17	**4014**
	18	**2904**

Features of Allele Number Nomenclature

- Compact enough for computer display
- Any lab using the same kit gets the same allele number
- Easy to vocalize

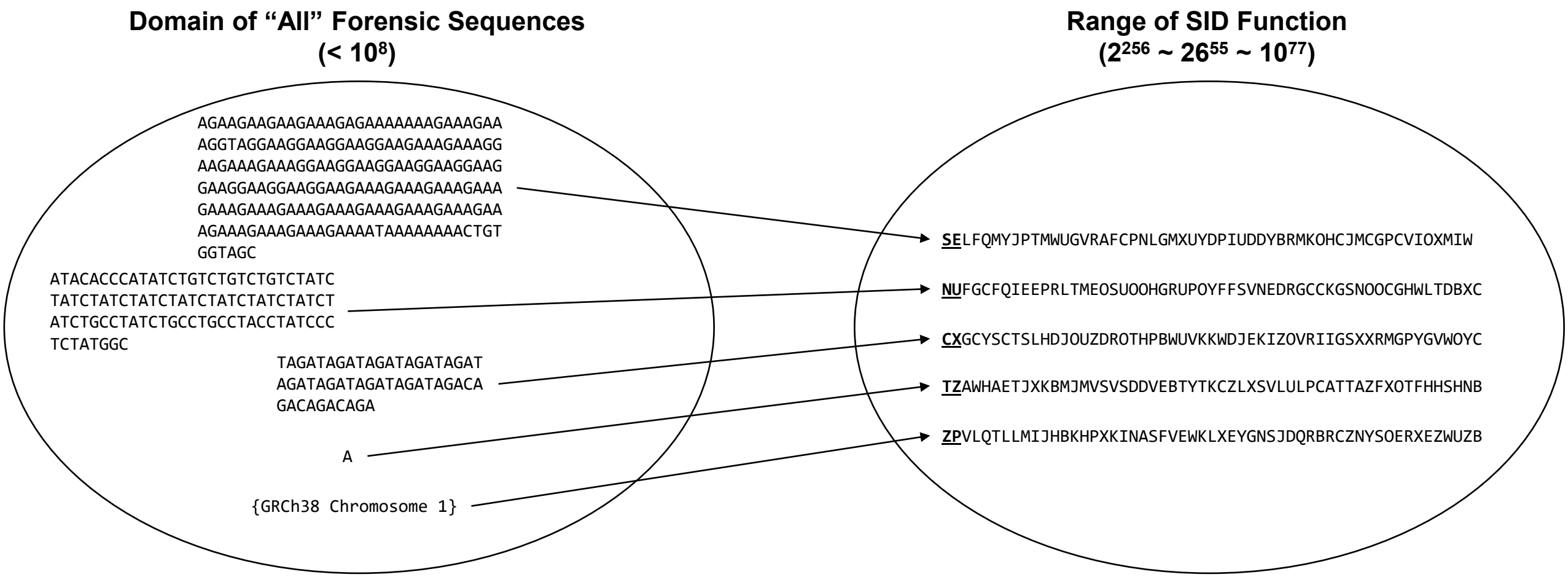
NicheVision Forensics, LLC

Outline

1. Motivation
2. The Challenge of MPS Artifacts in Mixed DNA Interpretation
3. Sequence Identifier (SID) Nomenclature
4. Why We Need SID: Labeling of Artifacts by ‘Phylogeny’
5. Mixture Interpretation of Sequence-Based MPS Data Using MixtureAce™ Plugin to ArmedXpert™ Software

SID: Shorthand Labels for Sequences

MixtureAce
. . .For Analysis of Mixed MPS Data



Any Sequence String → Mapping Function → 55-Character SID Label (Little-Endian) Digits Arranged Least- to Most-Significant

NicheVision Forensics, LLC

SIDs are Shorthand Labels for Sequences

MixtureAce
. . .For Analysis of Mixed MPS Data

Domain of "All" Forensic Sequences
($< 10^8$)

1,000 forensic loci
100 alleles each
1,000 trim sites each
$= 10^8$

Number of possible sequences of a 200-nucleotide amplicon
$= 4^{200} \sim 10^{120}$

Number of human chromosomes on earth
$\sim (8 \times 10^9) \times (2)$
$\sim 1.6 \times 10^{10}$

Number of different sequences observed in allele surveys
$<\sim 100$

Range of SID Function
($2^{256} \sim 26^{55} \sim 10^{77}$)

SELFQMYJPTMWUGVRAFCPNLGMXUYDPIUDDYBRMKOHCJMCGPCVIOXMIW

NUFGCFQIEEPRLTMEOSUOOHGRUPOYFFSVNEDRGCCKGSNOOCGHWLTDBXC

CXGCYSCTSLHDJOUZDROTHPBWUVKKWDJEKIZOVRIIGSXXRMGPYGVWOYC

TZAWHAETJXKBMJMVSVSDDVEBTYTKCZLXSVLULPCATTAZFXOTFHHSHNB

ZPVLQTLLMIJHBKHPXKINASFVEWKLXEYGNSJDQRBRCZNYSOERXEZWUZB

Any Sequence String → Mapping Function → 55-Character SID Label (Little-Endian)
Digits Arranged Least- to Most-Significant

NicheVision Forensics, LLC

SIDs are Shorthand Labels for Sequences

MixtureAce
. . .For Analysis of Mixed MPS Data

Domain of "All" Forensic Sequences
(< 10^8)

1,000 forensic loci
100 alleles each
1,000 trim sites each
1,000 artifacts each

= 10^{11}

Range of SID Function
(2^{256} ~ 26^{55} ~ 10^{77})

SELFQMYJPTMWUGVRAFCPNLGMXUYDPIUDDYBRMKOHCJMCGPCVIOXMIW

NUFGCFQIEEPRLTMEOSUOOHGRUPOYFFSVNEDRGCCKGSNOOCGHWLTDBXC

CXGCYSCTSLHDJOUZDROTHPBWUVKKWDJEKIZOVRIIGSXXRMGPYGVWOYC

TZAWHAETJXKBMJMVSVSDDVEBTYTKCZLXSVLULPCATTAZFXOTFHHSHNB

ZPVLQTLLMIJHBKHPXKINASFVEWKLXEYGNSJDQRBRCZNYSOERXEZWUZB

Any Sequence String → Mapping Function → 55-Character SID Label (Little-Endian)
Digits Arranged Least- to Most-Significant

NicheVision Forensics, LLC

Example Sequence
D2S441 [TCTA]12 including 30 nt upstream and 10 nt downstream

MixtureAce
. . .For Analysis of Mixed MPS Data

DNA Sequence (88 nt)

GCCAGGAACTGTGGCTCATCTATGAAAACTTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATATCATAACA

SHA-256 Hash in Hexadecimal (64 digits)

feb9d9c28bc82a468471f2b0a9afda5144a0d5805c6d2968f2bbc57a50ac0968

SHA-256 Hash in Hexavigesimal (55 digits)

4co4bmh75k48o4mjpjcbj5jealbhpboooii0hke5gg7ih1k33l46cba

Capitalized

4CO4BMH75K48O4MJPJCBJ5JEALBHPBOOOII0HKE5GG7IH1K33L46CBA

Reversed

A B C 6 4 L 3 3 K 1 H I 7 G G 5 E K H 0 I I O O O B P H B L A E J 5 J B C J P J M 4 O 8 4 K 5 7 H M B 4 O C 4

ASCII Codes

65 66 67 54 52 76 51 51 75 49 72 73 55 71 71 53 69 75 72 48 73 73 79 79 79 66 80 72 66 76 65 69 74 53 74 66 67 74 80 74 77 52 79 56 52 75 53 55 72 77 66 52 79 67 52

Offset

75 76 77 71 69 86 68 68 85 66 82 83 72 81 81 70 79 85 82 65 83 83 89 89 89 76 90 82 76 86 75 79 84 70 84 76 77 84 90 85 87 69 89 73 69 85 70 72 82 87 76 69 89 77 69

ASCII Letters

K L M G E V D D U B R S H Q Q F O U R A S S Y Y Y L Z R L V K O T F T L M T Z T W E Y I E U F H R W L E Y M E

Dynamic Allocation of Letters Within a Context — KL

Optionally Prepend with Allele Number — 12 KL

Optionally Prepend with Locus Name — D2S441 12 KL

Universal & Portable (Like Allele Numbers):

- The same sequence will yield the same SID regardless of laboratory generating the SID
- Independent of any external reference
- But ‘SID Dictionary’ is Trim Specific

https://nichevision.github.io/sid.js/

NicheVision Forensics, LLC

Algorithm is Highly Sensitive to Small Changes In Input Strings

MixtureAce
. . .For Analysis of Mixed MPS Data

In cryptography, the “avalanche effect” means that nearly all digits change in response to even a single change in the input string.

Letter Changes	Sequence String	SID Labels
Base Sequence	TAGATAGA	**EJ**ZUKYGKYTNLOLXDTVLVGOTSMYKZJVBDFKJZLWIXEVULUGXZZDFGFJD
Substitution	TAG**G**TAGA	**TX**SUCZXTZDJWUVOCTXTTLMPNJRMUHJRUDLRFFDGWFVQQCIKYHMVYFVB
Deletion	TAGATAG_	**IN**OGHEHOUNOGOBUYWZBCLPJOHQJMFLUJUYZFWTAMSIHWKZCALUGNUAC
Insertion	TAGATAGA**T**	**NP**UCXCAJHYLRPJWBFHTLJUCXXMYTKBAHZIUKJIBUFPJYFOEIOWLQCDB

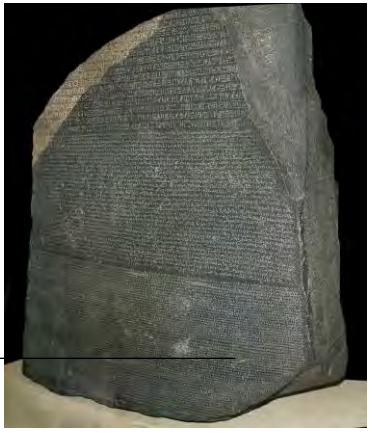
NicheVision Forensics, LLC

SID Rosetta Stone for Alleles at D1S1656

MixtureAce
. . .For Analysis of Mixed MPS Data

Dictionary of SID Labels and Sequence Strings
Genomic Extent = GRCh38 chr1:230769606…230769683

gi	gb	SID-3D	Length	Allele	String
1384961645	MH174866.1	GGN	19.3	[CA]5 CCTA [TCTA]14 TCA [TCTA]4	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCATCTATCTATCTATCTA
1384961644	MH174865.1	LCE	18.3	[CA]5 CCTA [TCTA]13 TCA [TCTA]4	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCATCTATCTATCTATCTA
1384961643	MH174864.1	DQQ	18	[CA]5 CCTA [TCTA]17	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961642	MH174863.1	IKU	17.3	[CA]5 [TCTA]13 TCA [TCTA]4	CACACACACATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCATCTATCTATCTATCTA
1384961640	MH174862.1	OGR	17.3	[CA]5 CCTA [TCTA]12 TCA TCTG [TCTA]3	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCATCTGTCTATCTATCTA
1384961638	MH174861.1	QIZ	17.3	[CA]5 CCTA [TCTA]12 TCA [TCTA]4	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCATCTATCTATCTATCTA
1384961635	MH174860.1	KGL	17	[CA]5 [TCTA]17	CACACACACATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961634	MH174859.1	SIA	17	[CA]5 CTTA [TCTA]16	CACACACACACTTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961633	MH174858.1	NEC	17	[CA]5 CCTA [TCTA]16	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961631	MH174857.1	ZDN	16.3	[CA]5 CCTA [TCTA]12 TCA [TCTA]3	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCATCTATCTATCTA
1384961630	MH174856.1	LQE	16.3	[CA]5 CCTA [TCTA]11 TCA [TCTA]4	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCATCTATCTATCTATCTA
1384961629	MH174855.1	BFZ	16	[CA]5 [TCTA]16	CACACACACATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961627	MH174854.1	FLL	16	[CA]5 CTTA [TCTA]15	CACACACACACTTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961626	MH174853.1	SCC	16	[CA]5 CCTA [TCTA]15	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961623	MH174852.1	CIG	15.3	[CA]5 CCTA [TCTA]11 TCA [TCTA]3	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCATCTATCTATCTA
1384961620	MH174851.1	QQI	15.3	[CA]5 CCTA [TCTA]10 TCA [TCTA]4	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCATCTATCTATCTATCTA
1384961618	MH174850.1	VZG	15	[CA]5 [TCTA]15	CACACACACATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961615	MH174849.1	QIQ	15	[CA]5 CTTA [TCTA]14	CACACACACACTTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961612	MH174848.1	OCB	15	[CA]5 CCTA [TCTA]14	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961611	MH174847.1	QUN	14.3	[CA]5 CCTA [TCTA]11 TCA [TCTA]2	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCATCTATCTA
1384961608	MH174846.1	QVO	14.3	[CA]5 CCTA [TCTA]9 TCA [TCTA]4	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCATCTATCTATCTATCTA
1384961606	MH174845.1	HNV	14	[CA]5 [TCTA]14	CACACACACATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961604	MH174844.1	JLZ	14	[CA]5 CCTA [TCTA]13	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961603	MH174843.1	---	14	[CA]5 CCTA [TCTA]13	Differs at rs1019813099
1384961602	MH174842.1	SVF	13	[CA]5 [TCTA]13	CACACACACATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961600	MH174841.1	AXL	13	[CA]5 TCTA GCTA [TCTA]11	CACACACACATCTAGCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961599	MH174840.1	YCZ	13	[CA]5 CCTA [TCTA]12	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961597	MH174839.1	FAE	12	[CA]5 [TCTA]12	CACACACACATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961595	MH174838.1	OGP	12	[CA]5 CCTA [TCTA]11	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961594	MH174837.1	CFP	11	[CA]5 [TCTA]11	CACACACACATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961593	MH174836.1	XAP	11	[CA]5 CCTA [TCTA]10	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961591	MH174835.1	IUW	10	[CA]5 [TCTA]10	CACACACACATCTATCTATCTATCTATCTATCTATCTATCTATCTA
1384961589	MH174834.1	BCM	10	[CA]5 CCTA [TCTA]9	CACACACACACCTATCTATCTATCTATCTATCTATCTATCTATCTA



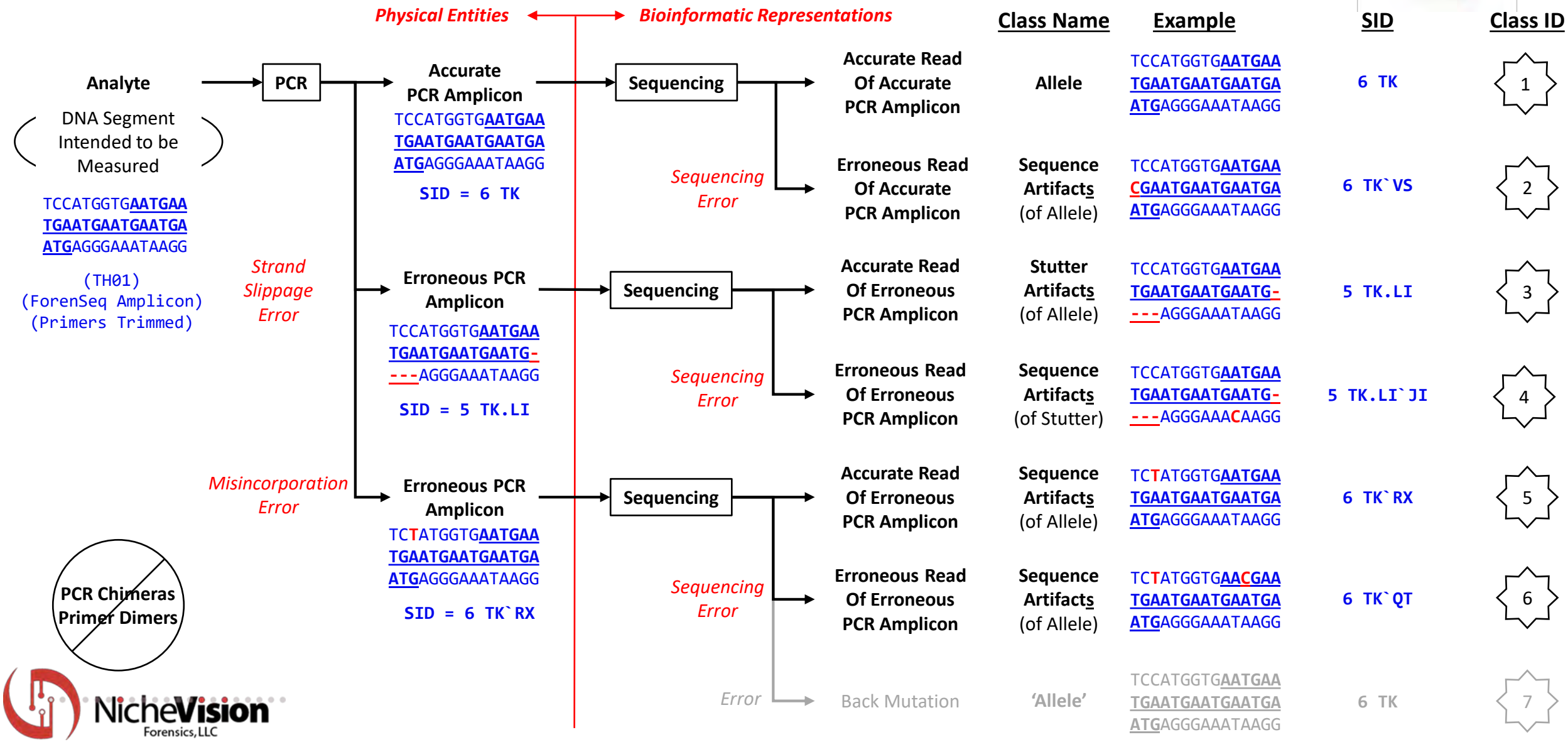
NicheVision Forensics, LLC

Outline

1. Motivation
2. The Challenge of MPS Artifacts in Mixed DNA Interpretation
3. Sequence Identifier (SID) Nomenclature
4. Why We Need SID: Labeling of Artifacts by ‘Phylogeny’
5. Mixture Interpretation of Sequence-Based MPS Data Using MixtureAce™ Plugin to ArmedXpert™ Software

‘Phylogenetic’ Model of the PCR-MPS Method

MixtureAce
. . .For Analysis of Mixed MPS Data



NicheVision Forensics, LLC

Artifact Assignment Rules

- Shared stutter artifacts can be apportioned to multiple parent alleles
- Ambiguous non-stutter artifacts remain as candidate alleles (potential minor)
- Categories of ambiguous non-stutter artifacts:
  1. Questioned sequences that are minimally close to > 1 parent
  2. Questioned sequences with minimal distances to any parent > δ

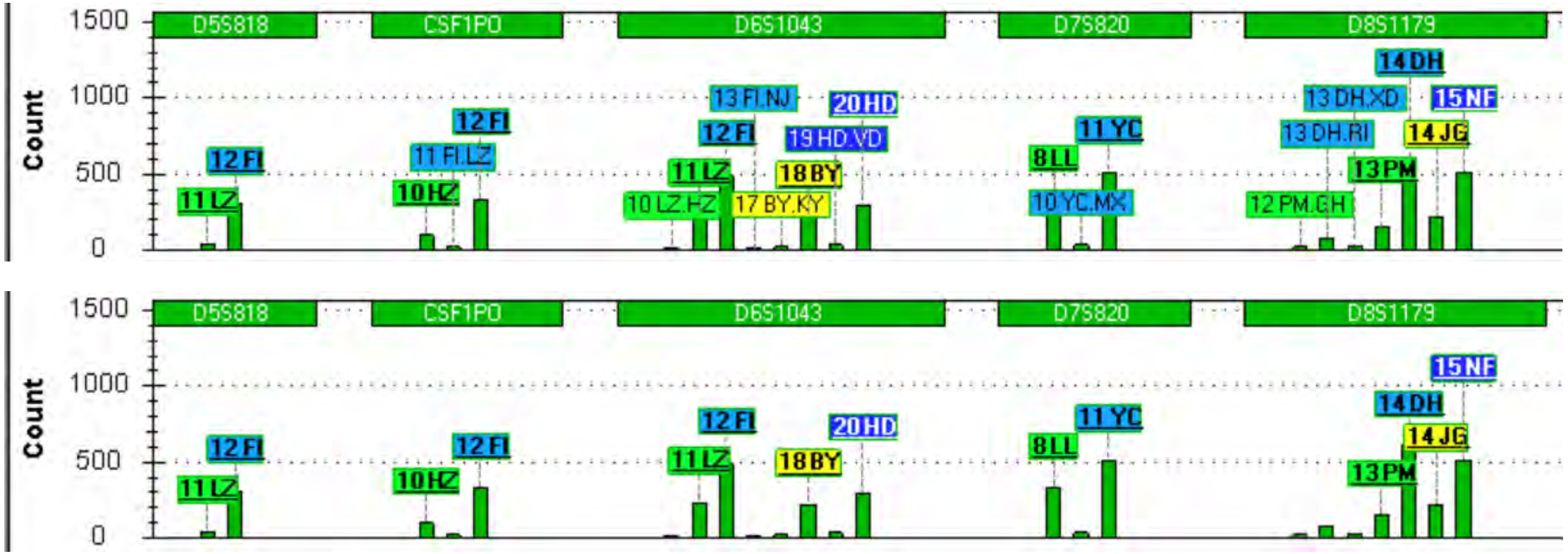
SID & Connector Conventions for MixtureAce Parent-Child Associations

PARENT.STUTTER	BT.NU	(dot)
PARENT`SEQARTIFACT	BT`HL	(tick)

Outline

1. Motivation
2. The Challenge of MPS Artifacts in Mixed DNA Interpretation
3. Sequence Identifier (SID) Nomenclature
4. Why We Need SID: Labeling of Artifacts by ‘Phylogeny’
5. Mixture Interpretation of Sequence-Based MPS Data Using MixtureAce™ Plugin to ArmedXpert™ Software

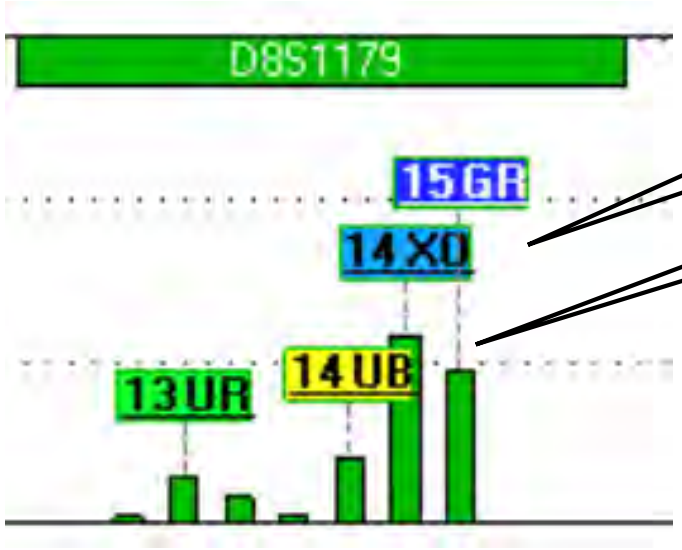
3:1 Mixture, AT = 10



Show Artifacts

Filter Artifacts

Isoalleles Are Displayed Separately

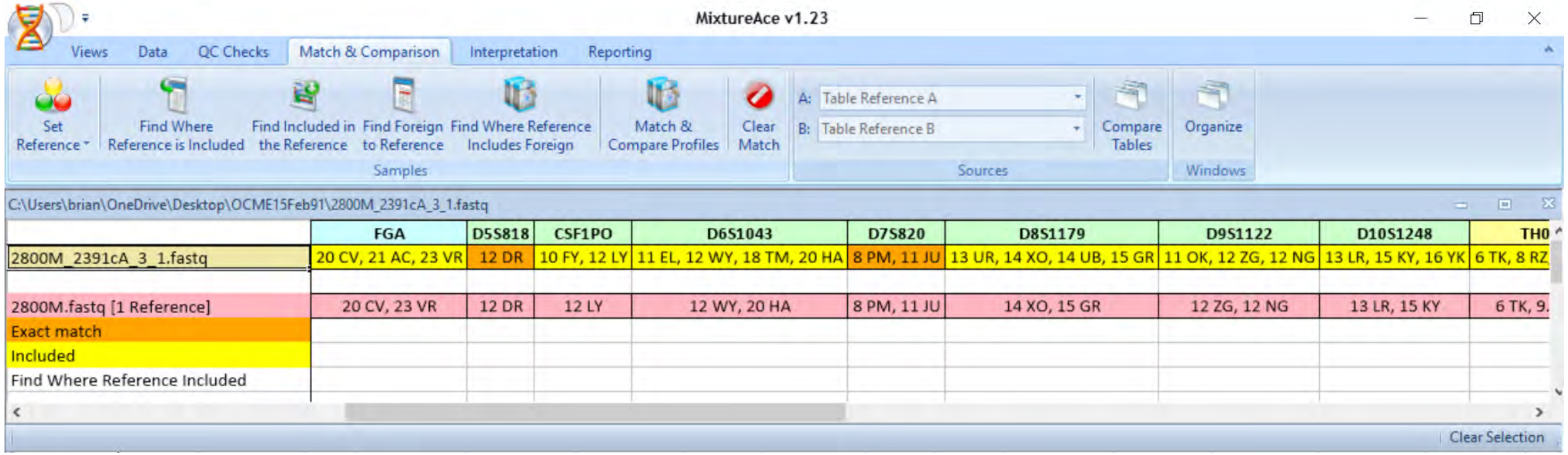


Length-14
Isoalleles From Different
Contributors

Contributor	Allele Length	Allele Sequence
2800M (Major)	14	[TCTA]1[TCTG]1[TCTA]12
2391c-A (Minor)	14	[TCTA]2[TCTG]1[TCTA]11

ArmedXpert Match & Comparison Tool With Sequence-Based Alleles

MixtureAce
. . .For Analysis of Mixed MPS Data

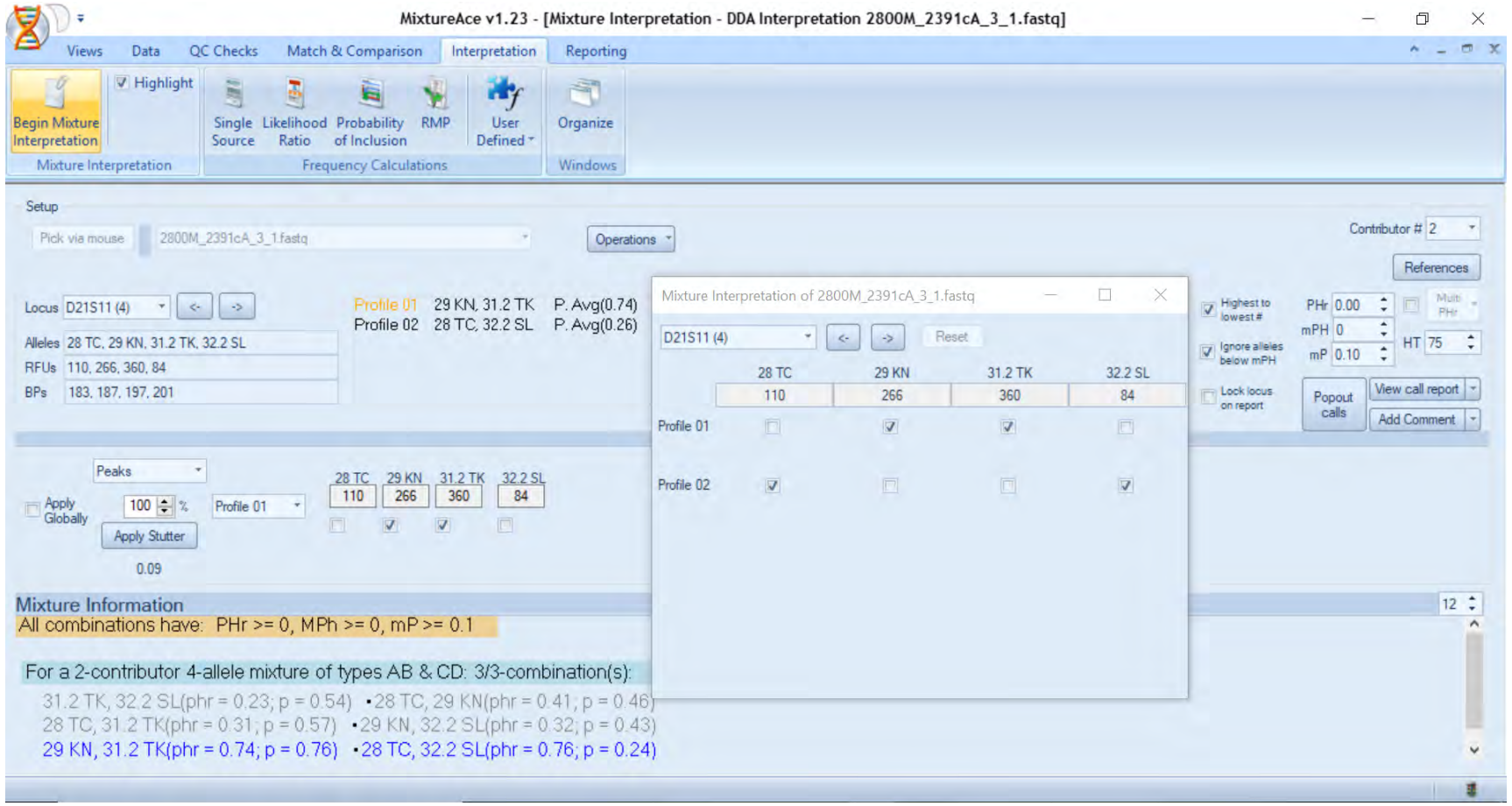


	FGA	D5S818	CSF1PO	D6S1043	D7S820	D8S1179	D9S1122	D10S1248	TH0
2800M_2391cA_3_1.fastq	20 CV, 21 AC, 23 VR	12 DR	10 FY, 12 LY	11 EL, 12 WY, 18 TM, 20 HA	8 PM, 11 JU	13 UR, 14 XO, 14 UB, 15 GR	11 OK, 12 ZG, 12 NG	13 LR, 15 KY, 16 YK	6 TK, 8 RZ
2800M.fastq [1 Reference]	20 CV, 23 VR	12 DR	12 LY	12 WY, 20 HA	8 PM, 11 JU	14 XO, 15 GR	12 ZG, 12 NG	13 LR, 15 KY	6 TK, 9.
Exact match									
Included									
Find Where Reference Included									

NicheVision Forensics, LLC

ArmedXpert Mixture Deconvolution With Sequence-Based Alleles

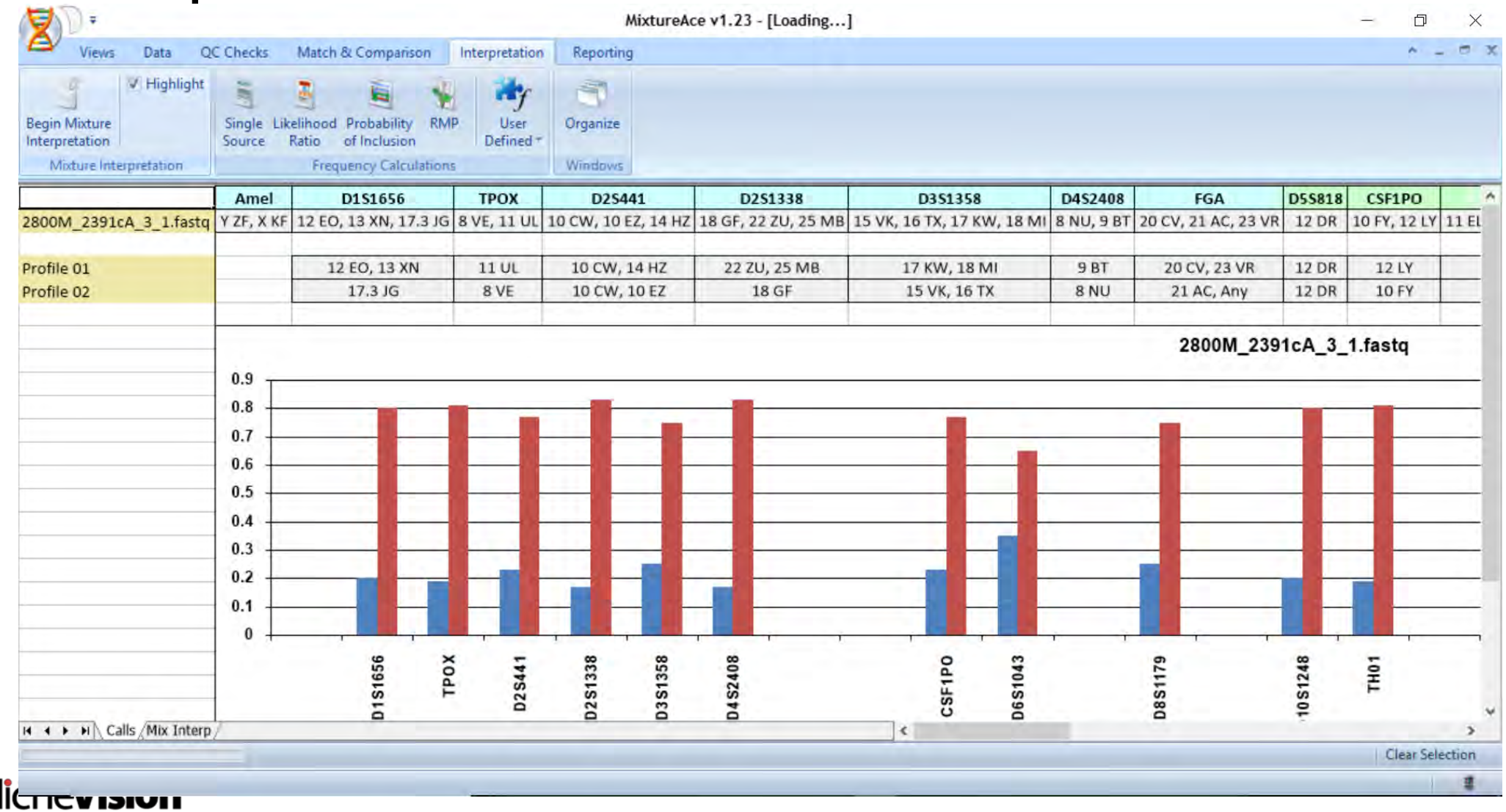
MixtureAce
. . .For Analysis of Mixed MPS Data



NicheVision Forensics, LLC

ArmedXpert Mixture Deconvolution Call Report

MixtureAce
. . .For Analysis of Mixed MPS Data



NicheVision Forensics, LLC

Audit Trail Of Sequence Type Assignments in CSV File

MixtureAce
. . .For Analysis of Mixed MPS Data

	C	D	E	F	H	I	J	K
1	Marker	SID	Length	Count	StutterOf	SEOf	SEQ1	SEQ2
85	D21S11	28 UT.ST	123	109	29 UT		[TCTA]4[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]10	TCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTC
86	D21S11	28 UT.OI	123	26	29 UT		[TCTA]4[TCTG]5[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11	TCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTC
87	D21S11	29 UT	127	1704			[TCTA]4[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11	TCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTC
88	D21S11	29 UT`IN	127	27		29 UT	[TCTA]4[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]1CCTA[TCTA]9	TCTATCTATCTATCTATCTGTCTGTCTGTCTGTCTGTC
89	D21S11	30.2 BZ.RZ	133	137	31.2 BZ		[TCTA]5[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]10TA[TCTA]1	TCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTC
90	D21S11	31.2 BZ	137	1714			[TCTA]5[TCTG]6[TCTA]3TA[TCTA]3TCA[TCTA]2TCCATA[TCTA]11TA[TCTA]1	TCTATCTATCTATCTATCTATCTGTCTGTCTGTCTGTC

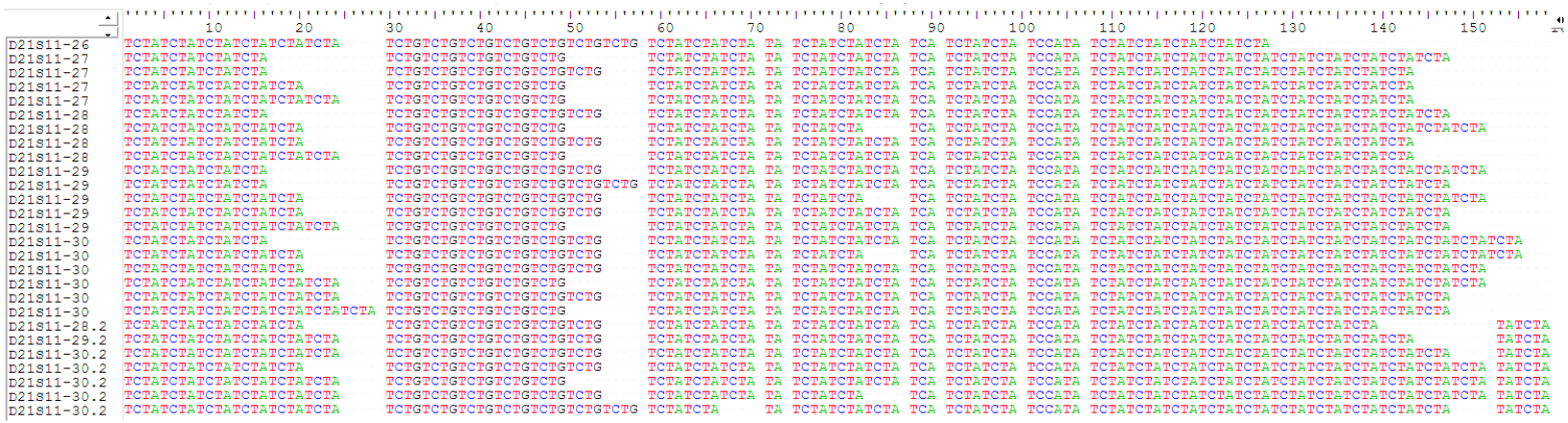
NicheVision Forensics, LLC

Acknowledgements

- SID Nomenclature
  - Nate Caldwell
  - Tom Faris
- Data
  - Dr. Elisa Wurmbach OCME
  - Dr. Michael Marciano, Syracuse U.
- Software Development
  - Luigi Armogida
  - Tom Faris
  - Abdul Alali

Leiden University
Medical Center

MPS –STR Allele Nomenclature Issues



Peter de Knijff, Dept. of Human Genetics, Leiden University Medical Center

MPS - STR Nomenclature

acknowledgments

- **Lab-staff:** **Kristiaan van der Gaag**
  Thirsa Kraaijenbrink
  Rick van Leeuwen
  Jerry Hoogenboom

MPS - STR Nomenclature

why a new nomenclature

Practical Issues:

- **We now have to describe sequence variation.**
- **We have a unique opportunity to repair forward – reverse mistakes.**
- **There will be "repeat" confusion (N-CE ≠ N-MPS).**
- **Have to deal with SNP variation.**
- **There is already a HGVS nomenclature in use.**
- **Describing MPS – STR variants in reports (articles) must be compatible with CE – STR results.**
- **Allele input in stat-progs and databases.**

MPS - STR Nomenclature

why a new nomenclature

Two choices:

- **A one size fits all solution (a code).**
- **A set of different solutions.**

Minimal criteria:

- **No online only repository / coding solution.**
- **Should be able to accommodate kit design variation.**
- **Accept FASTA – format.**

MPS - STR Nomenclature

nomenclature challenges

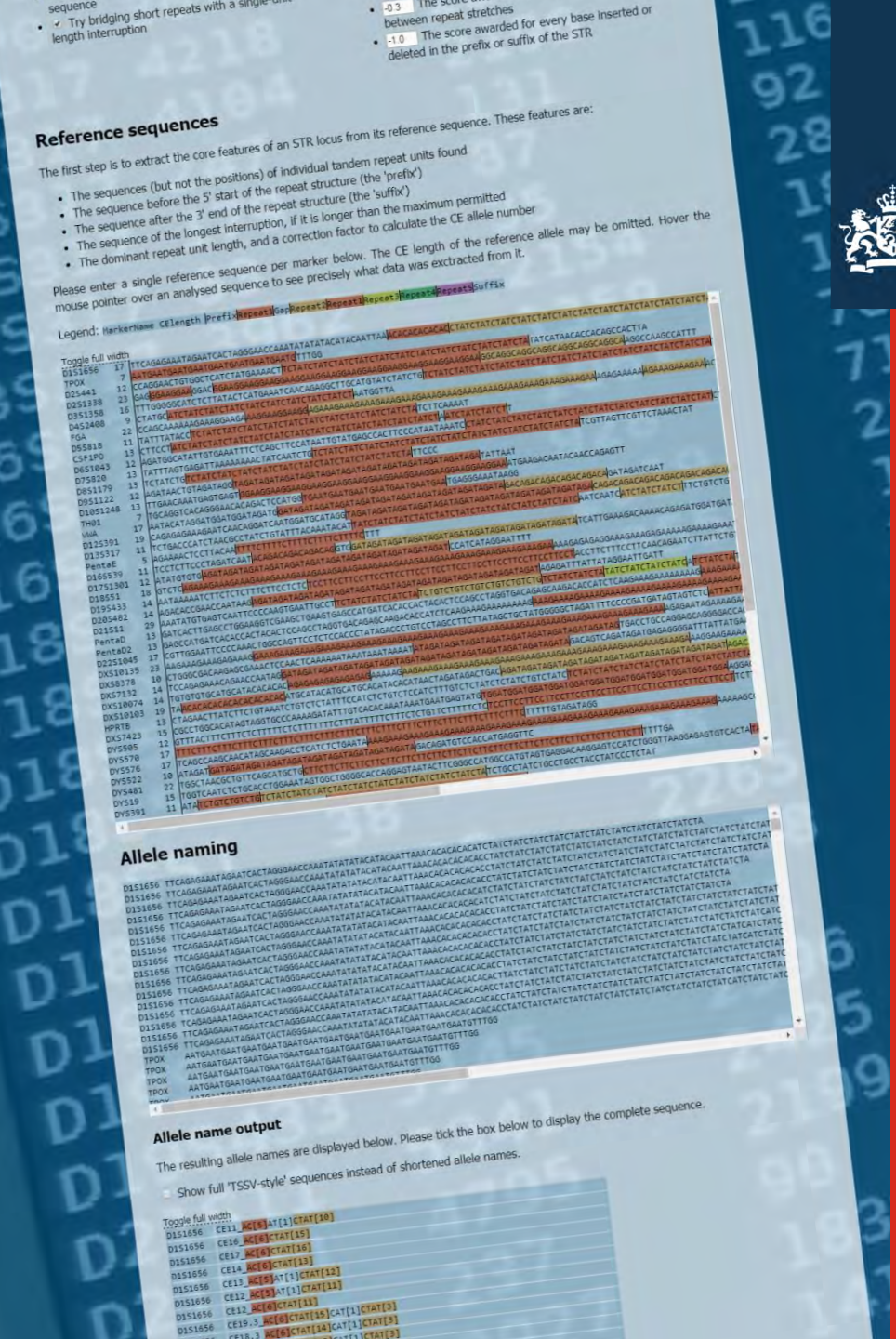
```
D13S317-11      TCTAACGCCT ATCTGTATTT ACAAATACAT -TATC-TATC -TATC-TATC
D13S317[CE12]   .....A.... .......... .......... -....-.... -....-....

D13S317-11      -TATC-TATC -TATC-TATC -TATC-TATC -TATC------ ------AATC
D13S317[CE12]   -....-.... -....-.... -....-.... -....-TATC -TATC-....

D13S317-11      -AATC-ATCT -ATCT-ATCT -TT
D13S317[CE12]   ------.... -....-.... -..
```

D13S317[CE11]-Chr13-GRCh37-g.82.722.160:82.722.223-TATC[11]AATC[2]ATCT[3]

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

D13S317[CE12] - TATC[13]AATC[1]ATCT[3] – 24 136G>A

D13S317[CE12] – 101



Netherlands Forensic Institute
Ministry of Justice

STRNaming

Standardised STR sequence allele naming

Jerry Hoogenboom & Kris van der Gaag

April 11th 2019

Different naming approaches for each situation

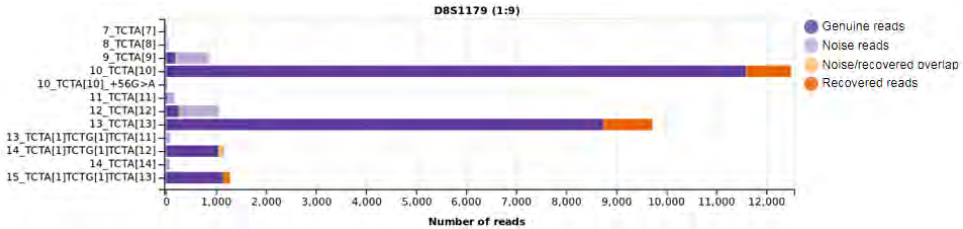
- The entire sequence
  - Exchange of data
  - Software processing and storage in database

- **Human-readable format; preserving sequence information**
  - **Publications, presentations, reports, etc.**
  - **Manual profile interpretation**
  - **Within-case manual comparisons**

- Very short code
  - Compact tables
  - Software that can’t handle complete sequences

Our focus: human-interpretable names

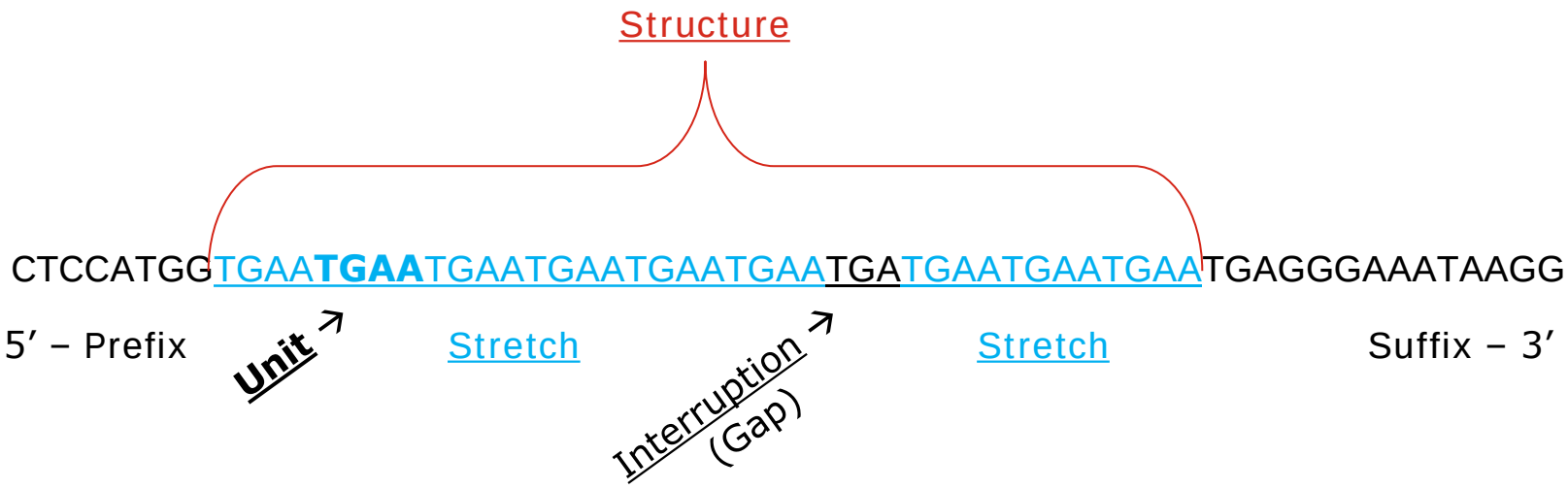
- One unique name describing one sequence
  - One-to-one relationship



- Sequence-descriptive
  - Small change in sequence → small change in name
  - The exact sequence change is obvious from the name
  - **Essential** in interpreting stutters/hybrids/artefacts by humans!

- Standardised and automated
  - Should work the same on any locus in current or future use
  - Don’t want to spend weeks working out the details of a locus
  - Publically available online & offline; open source algorithm

Definitions



Which name would you prefer?

CTCCATGGTGAATGAATGAATGAATGAATGAATGATGAATGAATGAATGAGGGAAATAAGG

CTCCATGGTGAATGAATGAATGAATGAATGATGATGAATGAATGAGGGAAATAAGG
CTCCATGGTGAATGAATGAATGAATGAATGAATGATGAATGAATGAGGGAAATAAGG
CTCCATGGTGAATGAATGAATGAATGAATGATGAATGAATGAATGAGGGAAATAAGG
CTCCATGGTGAATGAATGAATGAATGAATGAATGATGAATGAATGAGGGAAATAAGG

1. ATGA[6]TGA[1]ATGA[3] Nice! Two stretches of the same repeat unit!
2. TGAA[6]TGA[1]TGAA[3] Nice! Moved 5′ as far as possible!
3. ATGA[6]TGAA[3] Nice! No interruptions!
4. TGAA[6]TGA[2]ATGA[3] Nice! Only repeats, no interruptions, short prefix/suffix!

→ None of these is bad!

Also, which locus is this?

Which name would you prefer?

CTCCATGGTGAATGAATGAATGAATGAATGAATGATGAATGAATGAATGAGGGAAATAAGG 9.3

CTCCATGGTGAATGAATGAATGAATGAATGATGATGAATGAATGAGGGAAATAAGG 9.3
CTCCATGGTGAATGAATGAATGAATGAATGAATGATGAATGAATGAATGAGGGAAATAAGG 9.3
CTCCATGGTGAATGAATGAATGAATGAATGATGAATGAATGAATGAGGGAAATAAGG 9.3
CTCCATGGTGAATGAATGAATGAATGAATGAATGATGAATGAATGAGGGAAATAAGG 9.3

CE9.3 ATGA[6]TGA[1]ATGA[3]	Nice! Two stretches of the same repeat unit!
CE9.3 TGAA[6]TGA[1]TGAA[3]	Nice! Moved 5′ as far as possible!
CE9.3 ATGA[6]TGAA[3]	Nice? No interruptions!
CE9.3 TGAA[6]TGA[2]ATGA[3]	Nice? Only repeats, no interruptions, short prefix/suffix!

→ Do we care about the CE number?

→ What would allele 9 look like?

Which name would you prefer?

CTCCATGGTGAATGAATGAATGAATGAATGAATGAATGAATGAATGAGGGAAATAAGG 9

CTCCATGGTGAATGAATGAATGAATGAATGAATGAATGAATGAATGAGGGAAATAAGG 9
CTCCATGGTGAATGAATGAATGAATGAATGAATGAATGAATGAATGAGGGAAATAAGG 9
CTCCATGGTGAATGAATGAATGAATGAATGAATGAATGAATGAATGAGGGAAATAAGG 9
CTCCATGGTGAATGAATGAATGAATGAATGAATGAATGAATGAATGAGGGAAATAAGG 9

CE9.3	ATGA[6]TGA[1]ATGA[3]	CE9 ATGA[9]	Nice!
CE9.3	TGAA[6]TGA[1]TGAA[3]	CE9 TGAA[9]	Nice!
CE9.3	ATGA[6]TGAA[3]	CE9 ATGA[8]A[1]	Need to account for the extra A…
CE9.3	TGAA[6]TGA[2]ATGA[3]	CE9 TGAA[9]TGA[1]	Name has changed a lot?

→ We have 100s of markers more…

Our solution!

STRNaming

- Figures out what is repeated
- Figures out the nicest way of shortening that
  - The rules aren’t simple, but STRNaming does it for you
- Apply some pretty notation…
- Done!

Availability

- Free online & offline STRNaming tools
- Open-source the algorithm

But how?

The reference sequence

1. Find longest repeat stretch → *extract repeat unit only*
   - Repeat until no more non-overlapping stretches are found
2. Shift the extracted units (optional step)
   - E.g., CTAT → TATC, ATCT, TCTA
3. Find all stretches of all extracted units (ignoring overlap)
4. Select the nicest combination of repeat stretches (structure)

1. CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTTCA
   CTAT ATCT
2. CTAT TATC ATCT TCTA
3. CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTTCA
   CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTTCA
   CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTTCA
   CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTTCA
4. CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTTCA

The reference sequence

- From the reference, we store:
  - Which repeat **units** were used in its name → TCTA, ATCT
  - The prefix and suffix sequence → CTTCCTA, TCA
  - The sequence of the longest interruption (if very long) → n/a
  - Its length and CE allele number → 58bp = CE9

Sequence:

CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTTCA

Structure:

CTTCCTA(1)TCTA(9)ATCT(3)TCA(1)

Name:

CE9_TCTA[9]ATCT[3]

Naming other sequences

- From the reference, we had stored:
  - Which repeat **units** were used in its name → TCTA, ATCT
  - The prefix and suffix sequence → CTTCCTA, TCA
  - The sequence of the longest interruption (if very long) → n/a
  - Its length and CE allele number → 58bp = CE9

1. Find all stretches of TCTA and ATCT (ignoring overlap)

   CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTTCA

   CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTTCA
2. Select the nicest combination of repeat stretches (structure)

   CTTCCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAATCTATCTATCTTCA
3. Find and shorten repeat stretches in interruptions
   - Can use repeat units other than TCTA and ATCT
4. Convert the structure into an allele name
   - Simply a matter of notation, e.g.: CE9_TCTA[9]ATCT[3]

But what is nice???

Calculate a score for each STR structure

- Bonus points for:
  - Every base covered by a repeat
  - Every repeat of a unit
  - Every interruption that is exactly *x* repeat units long

- Penalty points for:
  - Every distinct repeat unit used
  - Every interruption introduced between stretches
  - Every base in an interruption
  - Every base inserted or deleted in the prefix or suffix

Same for all loci!

How many points? → Mostly common sense, expert opinion needed for the complex alleles

Some words on notation

- Variation in suffix

GGTAAACAGTATATTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTATTTG**A**AATGGA

CE10_TTTTC[10]

GGTAAACAGTATATTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTATTTG**C**AATGGA

CE10_TTTTC[10]_+7A>C

GGTAAACAGTATATTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTATTTGAATGGA

CE9.3_TTTTC[10]_+9A>-

GGTAAACAGTATATTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTATTTGAAA**A**TGGA

CE10.1_TTTTC[10]_+9.1->A

Some words on notation

- Variation in prefix

GGTA**A**ACAGTATATTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTATTTGAAATGGA

CE10_TTTTC[10]

10 9 8 7 6 5 4 3 2 1

GGTA**C**ACAGTATATTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTATTTGAAATGGA

CE10_TTTTC[10]_-9A>C

10 9 7 6 5 4 3 2 1

GGTAACAGTATATTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTATTTGAAATGGA

CE9.3_TTTTC[10]_-8A>-

10 9 8 7 6 5 4 3 2 1

GGTAAA**A**CAGTATATTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTTTTCTATTTGAAATGGA

CE10.1_TTTTC[10]_-8.1->A

Some words on notation

- Large interruptions (example: DYS389)

AGATTGATAGAGGGAGGGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGATACATAGATAATACAGA

DYS389I CE12_GATA[9]GACA[3]

AGATTGATAGAGGGAGGGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGA**TACATAGATAATACAGATGAGAGTTG**
GATACAGAAGTAGGTATAATGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGACAGACAC
ACACATAGATAATACAGA

DYS389II CE29_GATA[9]GACA[3]**[]**GATA[12]GACA[6]

AGATTGATAGAGGGAGGGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGATAGATAGATAATACAGATGAGAGTTG
GATACAGAAGTAGGTATAATGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGACAGACAGACAGACAGACAGACAC
ACACATAGATAATACAGA

DYS389II CE29_GATA[9]GACA[3][3C>G]GATA[12]GACA[6]

Eager to try it out?

Just drop me a line!

j.hoogenboom@nfi.nl

Final remarks

Ideas on the to-do list:

- Auto-convert to forward strand: ISFG reconsideration
- Analyse reference sequence with more upstream and downstream sequence content
  - Stabilises names across kits
  - Auto-determine analysed range per kit
  - Provide table of analysed range

→ We are happy to share what we've got!

→ We are open for suggestions and improvements!

Acknowledgements

LUMC (FLDO)

- Rick de Leeuw
- Peter de Knijff

King's College / Verogen

- Laurence Devesse

NFI

- Titia Sijen

STRAND Working Group

Some rough thoughts

Peter Gill, Oyvind Bleka

April 11, 2019
STR Nomenclature Meeting, London, UK

Mixtures

- Probabilistic genotyping open-source software EuroForMix
- Modules that can analyse SNPs and STRs using LUS
- We are designing software that will convert sequences into LUS primary, secondary, tertiary etc.
- When we analyse mixtures, we will encounter genotypes for which a reference sample is not available.
- Also we cannot tell if a sequence that is in an apparent stutter position of a major contributor, is indeed a stutter.
  - We cannot make the assumption of stutter – but we can filter low level products.

Mixtures

- When we go to court we report the strength of evidence
  - E.g.. the probability of the evidence if it has come from Mr X and an unknown individual(s) vs. the probability of the evidence if it has come from x unknown individuals.
- With numerous loci to contend with and large numbers of case-stains, it becomes very time-consuming to suggest that the operator should examine each locus separately
- Consequently, we are looking towards complete automation of the interpretation process where the output that the court is interested in is the likelihood ratio.
- Strictly we don’t need nomenclature to do this, we can use raw sequence.

Why do we need a nomenclature at all?

- It is feasible to use raw strings of sequenced bases for computers
- But for database searches there must be a way to compare strings with traditional nomenclature – i.e. back-compatibility..
- Sequence strings need to be available and tied to a nomenclature, whatever it may be.
- It is easy for a computer to turn a string into a designation using a look-up table.

Stutters

- In probabilistic genotyping it is a requirement to identify stutters.
- The LUS method will do this, but we also have to consider secondary and tertiary LUS because these will also stutter, albeit at lower levels.
- Given an allele, the expectation of a given stutter is probabilistic, and the probability of stutter, and its sequence, will be dependent upon the length of the parent allele.
- Many of the ‘exotic’ stutters will be very low level, and the evidential value would be low – they could be filtered. But we can anticipate that they may be important with future iterations of probabilistic genotyping. Consequently we should build this into the existing strategy in order to prevent having to revisit at a later time.

Examples using D21S11

- There are 5x type 37.2 variants

	RU	LUS1	LUS2	LUS3
$[TCTA]_7\ [TCTG]_{14}\ [TCTA]_3\ TCA\ [TCTA]_2\ TCCA\ TA\ [TCTA]_{12}$	39.2.3	14	12	7
$[TCTA]_9\ [TCTG]_{12}\ [TCTA]_3\ TCA\ [TCTA]_2\ TCCA\ TA\ [TCTA]_{12}$	39.2.3	12	12	9
$[TCTA]_9\ [TCTG]_{13}\ [TCTA]_3\ TCA\ [TCTA]_2\ TCCA\ TA\ [TCTA]_{11}$	39.2.3	13	11	9
$[TCTA]_{10}\ [TCTG]_{11}\ [TCTA]_3TCA\ [TCTA]_2\ TCCA\ TA\ [TCTA]_{12}$	39.2.3	12	11	10
$[TCTA]_{11}\ [TCTG]_{11}\ [TCTA]_3TCA\ [TCTA]_2\ TCCA\ TA\ [TCTA]_{11}$	39.2.3	11	11	11

All the same

35.2 variant

$[TCTA]_5\ [TCTG]_6\ [TCTA]_3\ TA\ [TCTA]_3\ TCA\ [TCTA]_2\ TCCA\ TA\ [TCTA]_{15}\ TA\ TCTA$ 36.2.2.2.3;15;6;5;3;3;2

Stutter

37.2 variant								LUS notation
$[TCTA]_7\ [TCTG]_{14}\ [TCTA]_3\ TCA\ [TCTA]_2\ TCCA\ TA\ [TCTA]_{12}$								39.2.3;14;12;7

RU	LUS1	LUS2	LUS3	
39.2.3	14	12	7	Parent allele
38.2.2	13	12	7	-1 stutter
38.2.2	14	11	7	-1 stutter
38.2.2	14	12	6	-1 stutter
37.2.2	13	12	6	-1 stutter at 2 positions

Repeat exercise for +1 stutters, -2 stutters and multiple stutters

Type 37.2 variant where, primary, secondary, tertiary LUS are the same length

$[TCTA]_{11}\ [TCTG]_{11}\ [TCTA]_3TCA\ [TCTA]_2\ TCCA\ TA\ [TCTA]_{11}$ 39.2.3;11;11;11

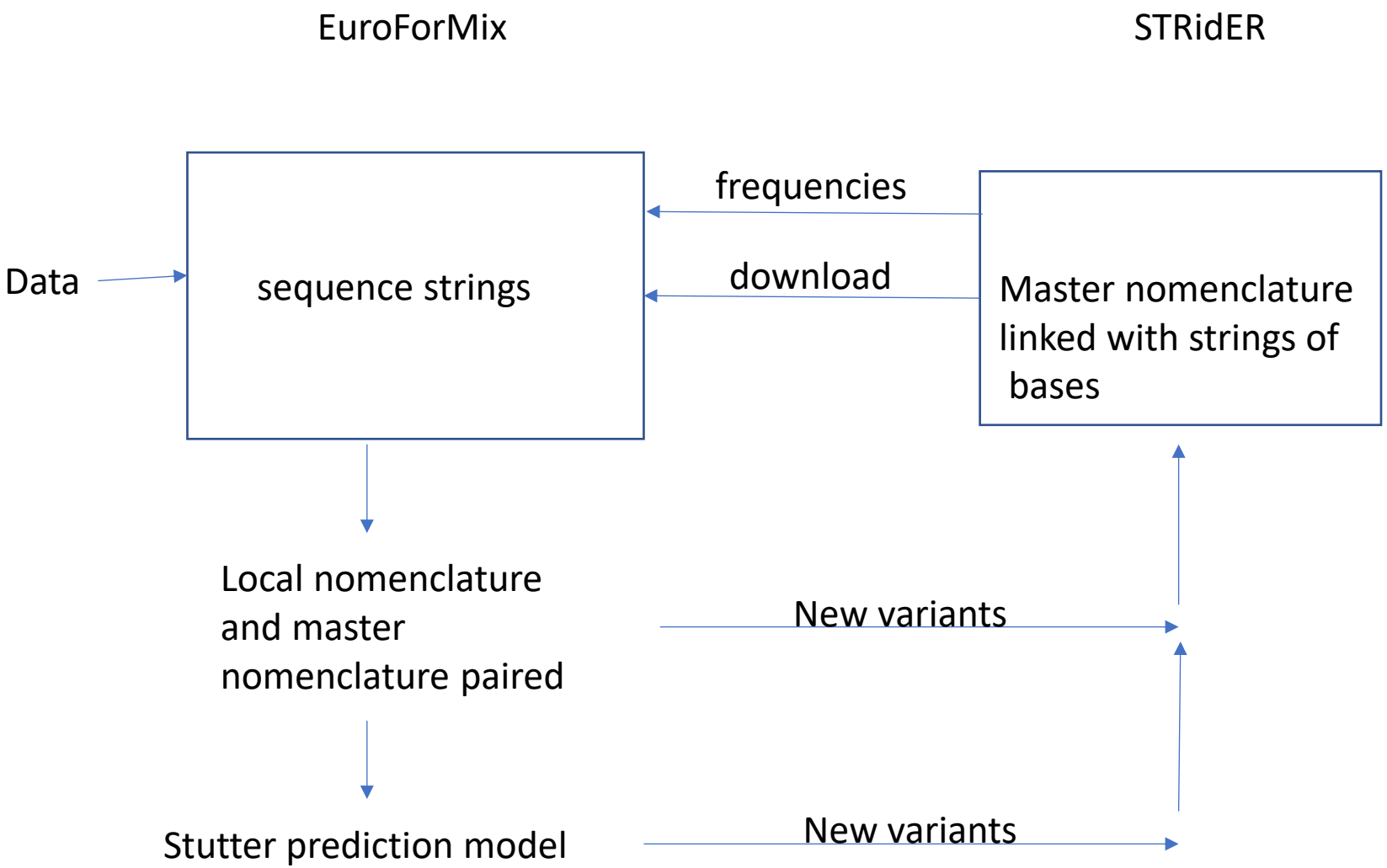
The notation does not tell us the order in the sequence, but this could be rectified by ordering the LUS variants as they appear e.g. 39;11;11;3;.3;2;1;.1;11

But this is a bit unwieldy

Conclusion

- Computers can take raw sequence and turn it into a local nomenclature that can be used for a particular software.
- LUS notation is probably sufficient if we use it on a per case basis – more difficult for a universal nomenclature.
- Within case a) we need to identify an allele b) we need to apply a frequency (use sequence comparison for this).
- A local nomenclature based on LUS is useful to assess stutters.
- But this can be local – per case.

Nomenclature



STRAND WG Meeting

LONDON
APRIL 11, 2018

TOASTR

A user-friendly web app for STR sequencing

labcon · · ·O·W·L·

Sebastian Ganschow, M.Sc
LABCON-OWL GmbH

LABCON-OWL

CORE BUSINESS

- clinical molecular diagnostics
- forensic DNA analysis

RESEARCH & DEVELOPMENT

- forensic multiplex kits for MPS
- bioinformatics solution for MPS-STR genotyping

Landscape of open-access genotyping tools

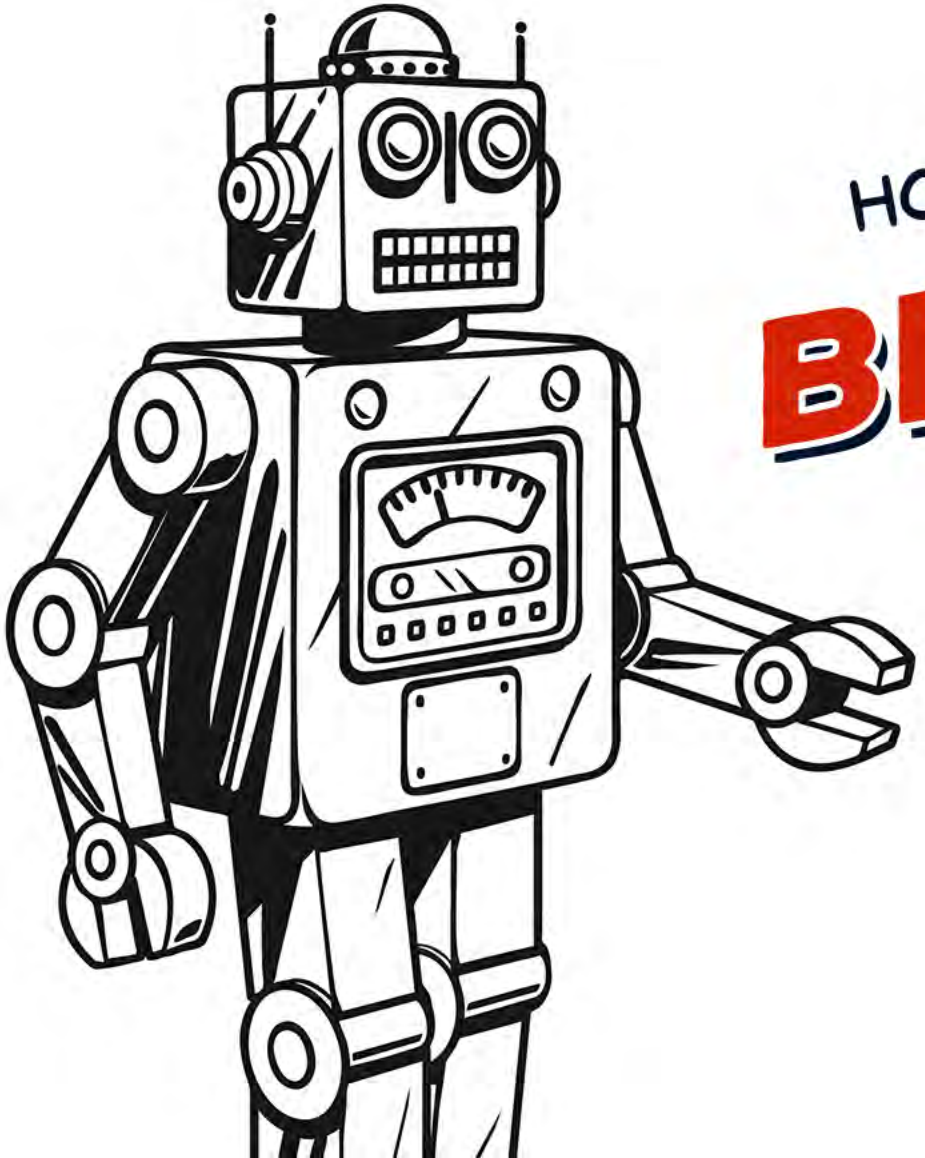
- gapped alignment problem:
  trade-off between runtime and accuracy
- bias due to incomplete reads
- do not report genotypes
- no stutter/noise modelling

- „selective“ alignment
- use whole-genome data
- legally and technically
  out of scope?

- anchor flanking sequences
- extract repeat region
- analyze length & sequence

Landscape of open-access genotyping tools

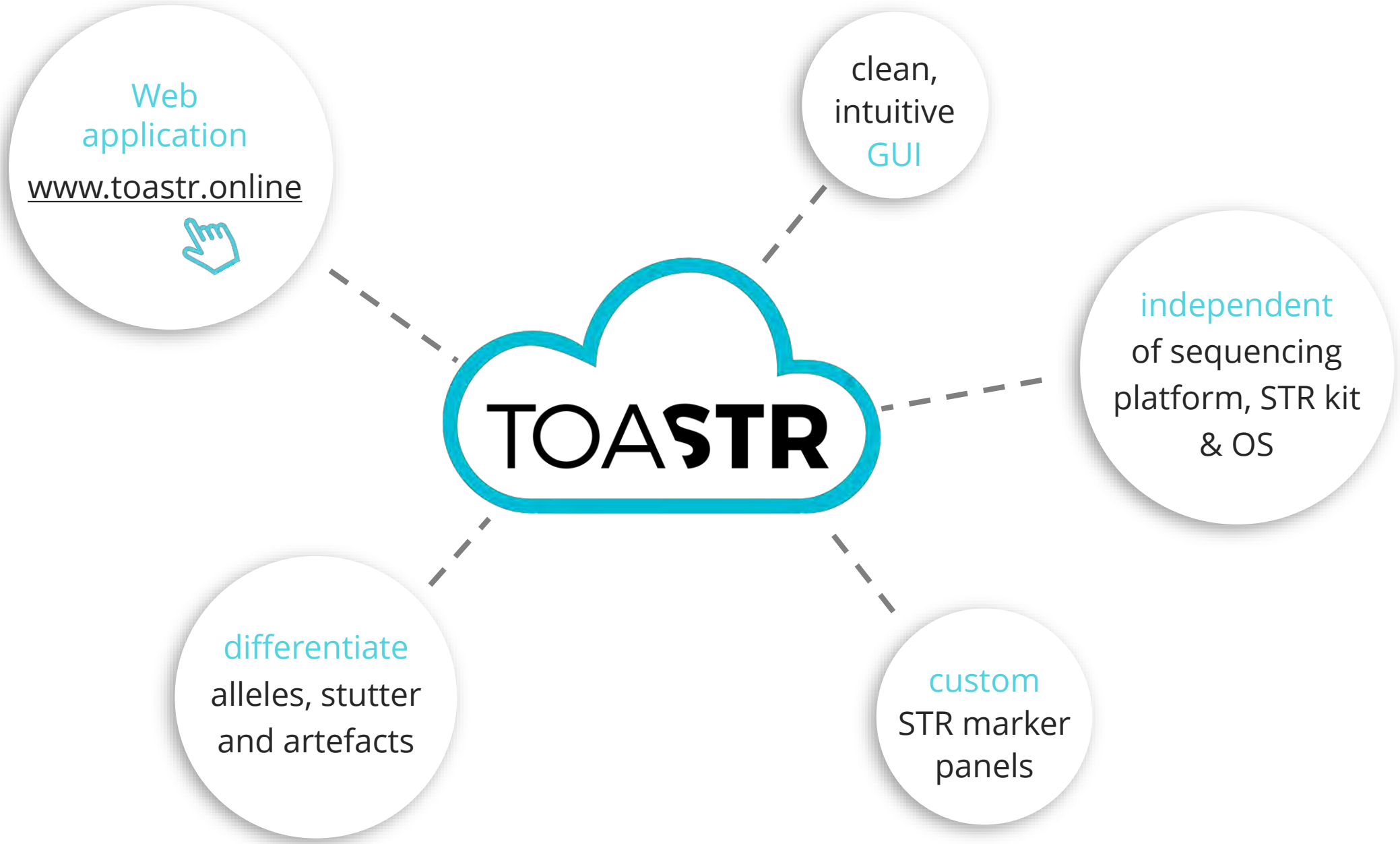
	Software	Programming language(s)	Platform	Input format	GUI	Parallel processing	Stutter modelling	Result visualization	ISFG nomenclature
genome-wide	**lobSTR** (Gymrek et al., 2012)	C/C++/R	**Unix**	FASTA, FASTQ, BAM	**no**	yes	yes	**no**	**no**
	RepeatSeq (Highnam et al., 2013)	C++/Python	**Unix**	BAM	**no**	yes	yes	**no**	**no**
	STR Viper (Cao et al., 2013)	Java	any	BAM, SAM	**no**	no	**no**	**no**	**no**
locus-centric	**STRait Razor v3** (Woerner et al., 2017)	C++	Windows, Unix	FASTQ	**no**	yes	**no**	yes (Excel workbook)	yes (known alleles)
	MyFLq (Van Neste et al., 2015)	Python	Web, Docker, Illumina BaseSpace	FASTA, FASTQ	yes	yes	**no**	yes	**no**
	FDSTools (Hoogenboom et al., 2017)	Python	**Unix**	FASTA	**no**	no	yes	yes	**no**
	STRinNGS (Friis et al., 2016)	Python/R	**Unix**	FASTQ, BAM	**no**	no	**no**	yes	**no**
	SEQ Mapper (Lee et al., 2016)	.NET	Windows, Web	FASTA, FASTQ	yes	no	**no**	**no**	**no**
	Altius (Bailey et al., 2017)	Python	Web	FASTQ	yes	yes	**no**	yes	yes
	toaSTR (Ganschow et al., 2018)	Perl	Web	FASTA, FASTQ	yes	yes	yes	yes	yes



HOW CAN WE MAKE

BIOINFORMATICS

easy?


Web application
www.toastr.online
clean, intuitive GUI
TOASTR
independent of sequencing platform, STR kit & OS
differentiate alleles, stutter and artefacts
custom STR marker panels

toaSTR algorithm

1 CALLING — 2 CLUSTERING — 3 FORMATTING — 4 MODELLING

```
>Seq_1
GTGGTGTGTATTCCCTGTGCCTTTGGGGTTTTTCTGTCGTTCACACGCATCTCCTTTTTGGCTGTGCCTTTGGGGGCATCTCTTATACTCATGAAATC
AACAGAGGCTTGCATGTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACTTTGTCTTTCTCTG
TCTCCCCTCTTTTACTTCCGTTTTTCTGTCGTTCACACTTCTCTTCTCCTTTTTCTGTCGTTCACACTTTTTCTGTCGTTCACATTATTCCCTGTGCCT

>Seq_2
GTGGTGTGTATTCCCTGTGCCTTTGGGGTTTTTCTGTCGTTCACACGCATCTCCTTTTTGGCTGTGCCTTTGGGGGCATCTCTTATACTCATGAAATC
AACAGAGGCTTGCATGTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACTTTGTCTTTCTCTGTCTCCCCT
CTTTTACTTCCGTTTTTCTGTCGTTCACACTTCTCTTCTCCTTTTTCTGTCGTTCACACTTTTTCTGTCGTTCACATTATTCCCTGTGCCT
```

recognition elements

-
-
-

toaSTR algorithm

1 CALLING — 2 CLUSTERING — 3 FORMATTING — 4 MODELLING

GAGGCTTGCATGTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACTTTGTCTTTC
GAGGCTTGCATGTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACTTTGTCTTTC
GAGGCTTGCATGTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACTTTGTCTTTC

5000x **TCTGTCTGTCTGTCTGTCATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA**

GAGGCTTGCATGTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACTTTGTCTTTC
GAGGCTTGCATGTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACTTTGTCTTTC
GAGGCTTGCATGTATCTGTCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATGAGACTTTGTCTTTC

3000x **TCTGTCTGTCTGTCTGTCATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA**

→ Observations

toaSTR algorithm

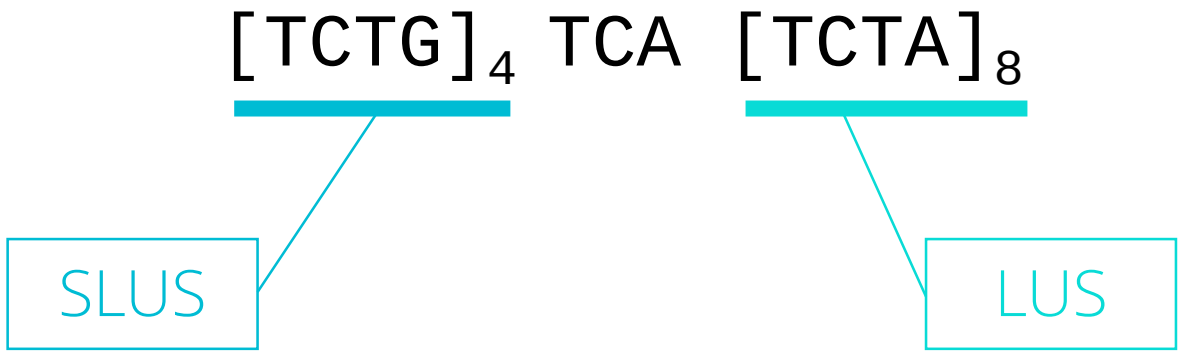
1 CALLING — 2 CLUSTERING — **3 FORMATTING** — 4 MODELLING

CE	Coverage	Sequence
12	5000x	$[TCTG]_4$ TCA $[TCTA]_8$
10	3000x	$[TCTG]_4$ TCA $[TCTA]_6$

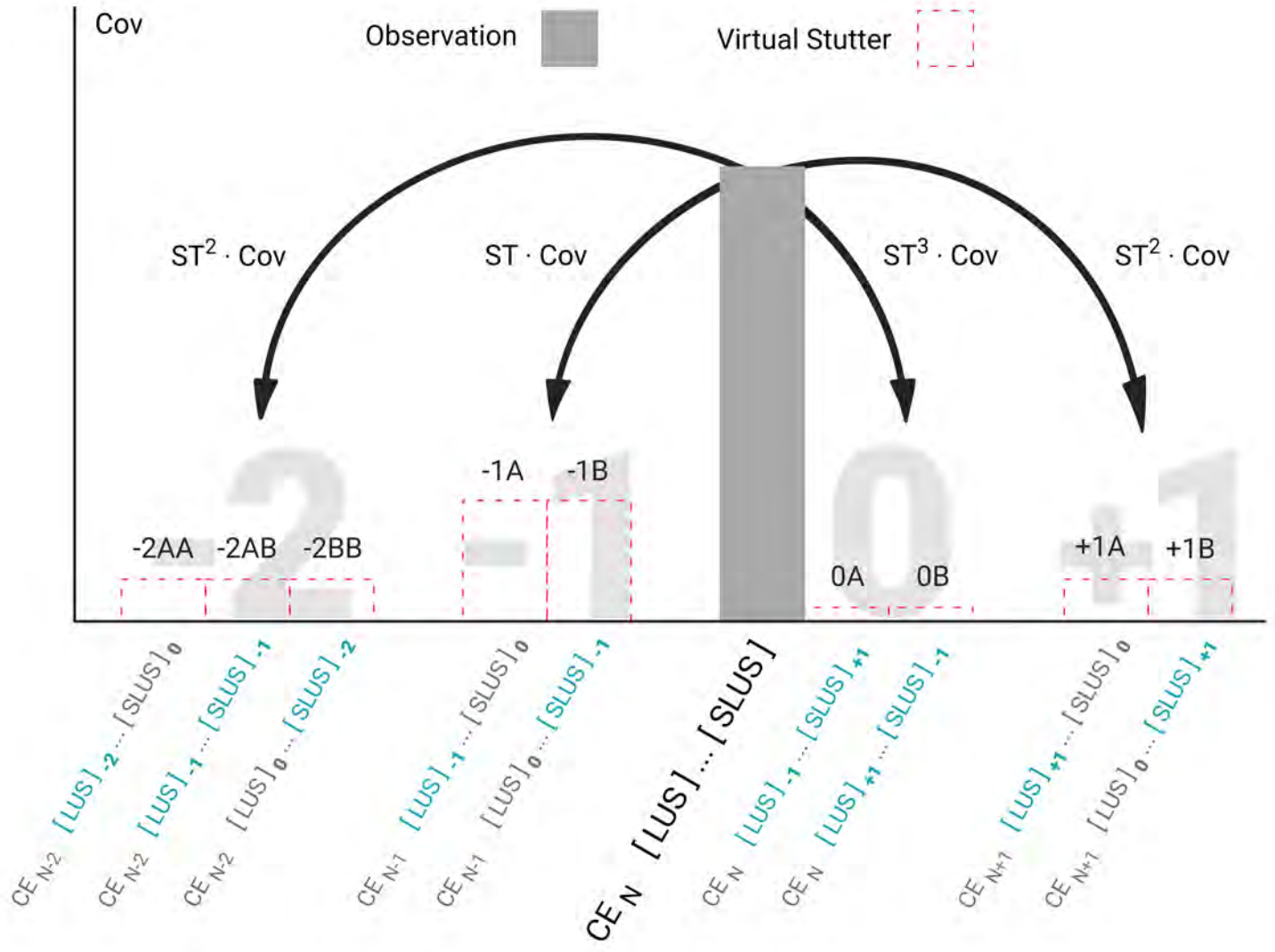
toaSTR algorithm

1 CALLING — 2 CLUSTERING — 3 FORMATTING — 4 MODELLING

- Classification of observations: alleles, stutter, artefact
- Sequence-based stutter prediction
- Assumption: stutter occurs most likely in the LUS and SLUS



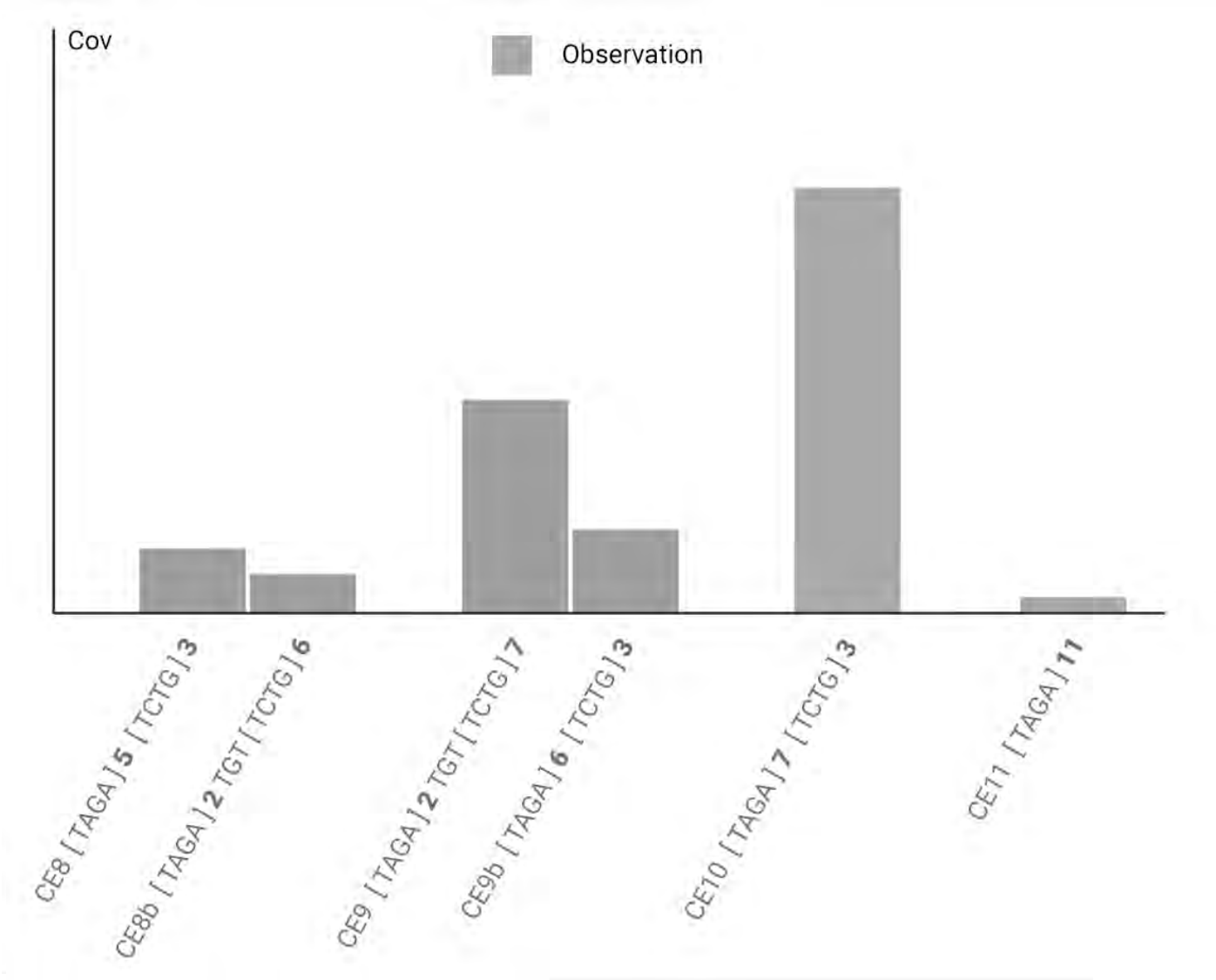
Stutter model



Classification

Stutter modelling enables automatic classification of observations:

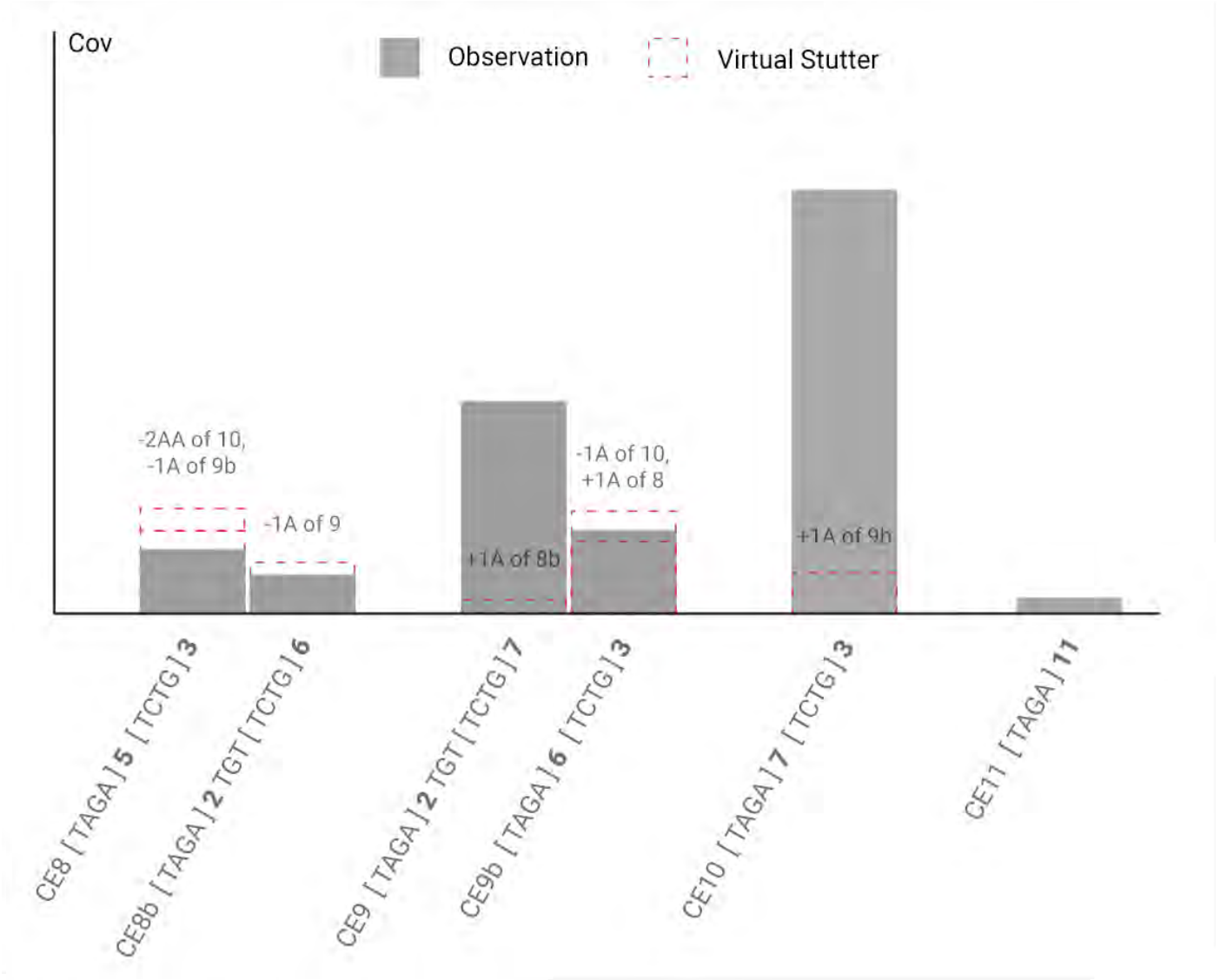
1. Calling of observations



Classification

Stutter modelling enables automatic classification of observations:

1. Calling of observations
2. Calculation of Virtual Stutter

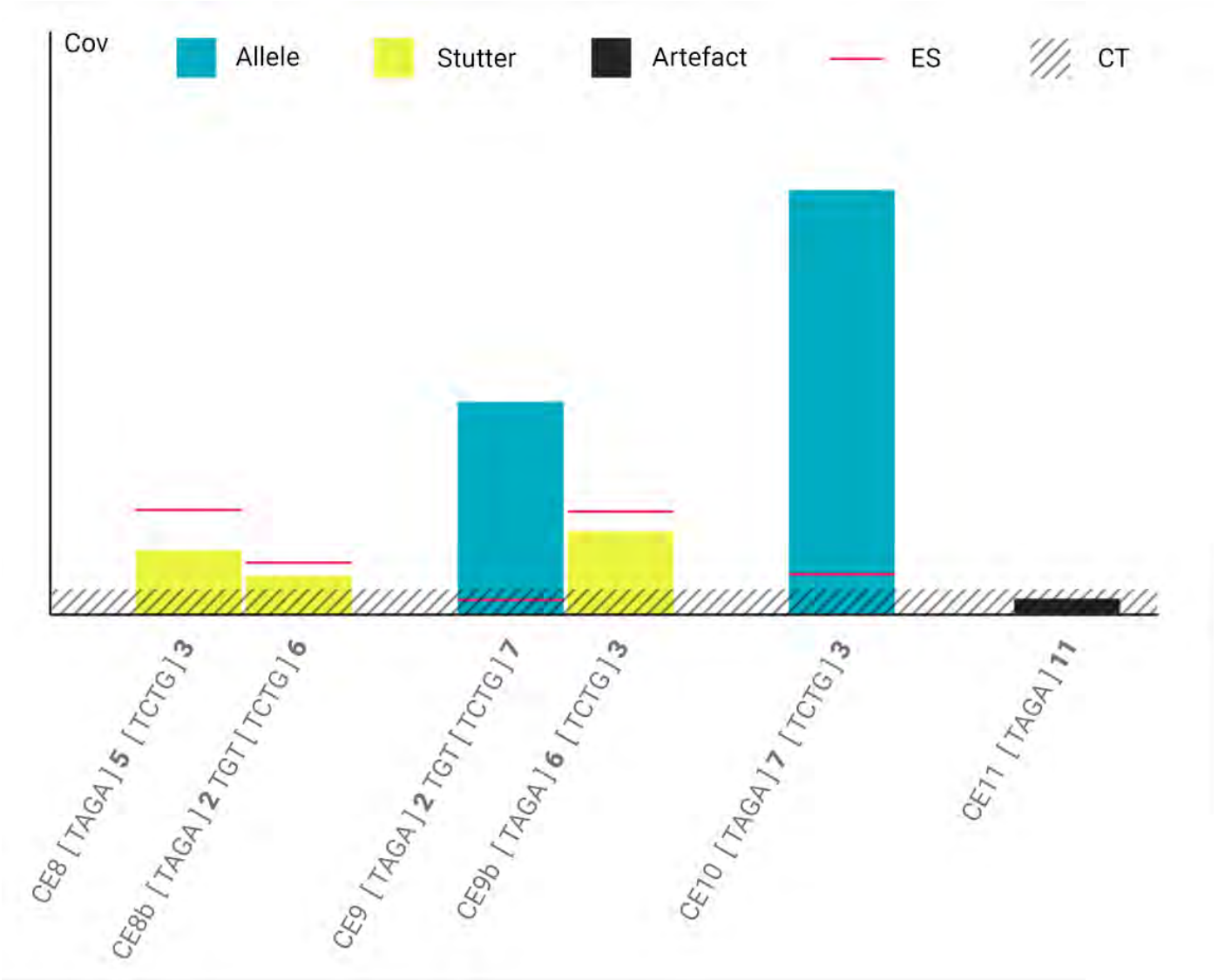


Classification

Stutter modelling enables automatic classification of observations:

1. Calling of observations
2. Calculation of Virtual Stutter
3. Classification

→ Supports the interpretation of mixed samples



Dashboard

Analyses

Start a new analysis, list finished and running analyses, show and edit results

GO

Panels

Create and view custom collections of STR markers and define stutter thresholds

VIEW

Manual

Read the comprehensive documentation of toaSTR features and algorithms

READ

Changelog

All notable changes to this project will be documented in this log.
The format is based on Keep a Changelog and this project adheres to Semantic Versioning.
The ⓘ icon indicates important changes that may affect your results.

1.0.0-beta.12 (2017-10-23)

Panels / New Panel

back to panels

Panel Name

0/30

Autosomal

D1S1656 %	D2S1338 %	D2S441 %	TPOX %
D3S1358 %	FGA %	CSF1PO %	D5S818 %
D6S1043 %	SE33 %	D7S820 %	D8S1179 %
D10S1248 %	TH01 %	D12S391 %	VWA %
D13S317 %	PENTA E %	D16S539 %	D18S51 %

FEEDBACK

Panels / New Panel

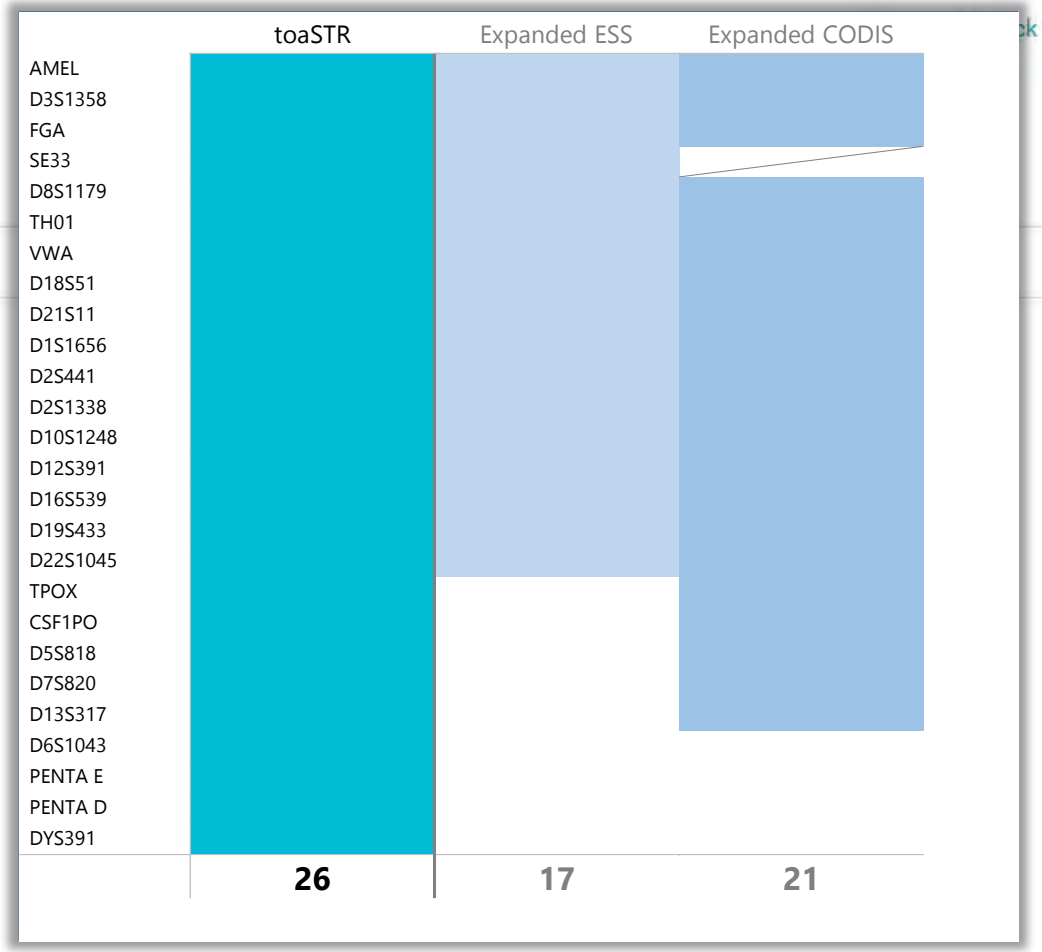
k to panels

Panel Name

21-plex in-house

Autosomal

D1S1656	15 %	D2S1338	20 %				
D3S1358	%	FGA	%				
D6S1043	%	SE33	%				
D10S1248	%	TH01	%				
D13S317	%	PENTA E	%	D16S539	%	D18S51	%



Dashboard

Analyses

Start a new analysis, list finished and running analyses, show and edit results

GO

Panels

Create and view custom collections of STR markers and define stutter thresholds

VIEW

Manual

Read the comprehensive documentation of toaSTR features and algorithms

READ

Changelog

All notable changes to this project will be documented in this log.
The format is based on Keep a Changelog and this project adheres to Semantic Versioning.
The ⓘ icon indicates important changes that may affect your results.

1.0.0-beta.12 (2017-10-23)

Analyses / New Analysis

Name

Notes

Sample type

Blood

Constellation

This is a single person sample [reference]

Panel

- Promega PowerSeq
- Illumina ForenSeq
- Thermo Early Access
- 21-plex in-house

Analyses / New Analysis

Name
2800M

Notes
Reference DNA

Sample type
Unknown

Constellation
This is a single person sample [reference]

Panel
- Promega PowerSeq
- Illumina ForenSeq
- Thermo Early Access
- 21-plex in-house

Constellation

This is a single person sample [reference]

Panel

- Promega PowerSeq
- Illumina ForenSeq
- Thermo Early Access
- 21-plex in-house

Analytical Threshold

10 reads (default)

Calling Threshold

2 % (default)

DATA FILE

Please refer to the manual for input recommendations.
Maximum file size 500MB. Supported formats:
FASTA (*.fasta, *.fa, *.fa.gz), FASTQ (*.fastq, *.fq, *.fq.gz)

START

FEEDBACK

INPUT READS

250000

CALLED READS

71568 (28%)

Quick edit

NAME

2800M

SAMPLE TYPE

Blood

CONSTELLATION

This is a single person sample [ref

NOTES

SAVE

Analyses / 2800M

ANALYTICAL THRESHOLD 10 reads

CALLING THRESHOLD 2 %

Summary

EXPORT

Promega PowerSeq Auto

SYSTEM	COVERAGE	ALLELES
D1S1656	4031	12 13
D2S1338	3049	22 25
D2S441	3220	10 14
TPOX	3584	11
D3S1358	2992	17 18
FGA	2511	20 23
CSF1PO	3662	12
D5S818	2982	12
D7S820	2514	8 11
D8S1179	3419	14 15
D10S1248	3044	13 15
TH01	3382	6 9.3
D12S391	3330	18 23
VWA	2195	16 19
D13S317	2396	9 11

INPUT READS

250000

CALLED READS

71568 (28%)

Quick edit

NAME

2800M

SAMPLE TYPE

Blood

CONSTELLATION

This is a single person sample [ref

NOTES

SAVE

EXPORT

D2S441

Called reads 3228
Stutter threshold 15 %

Results

CE	St	Cov	ES	Seq	Source
9	S	95	220	D2S441[CE9]-Chr02-GRCh38 68011947-68011994 [TCTA]9	-1A of 10, -2AA of 11,
10	A	1457	9	D2S441[CE10]-Chr02-GRCh38 68011947-68011994 [TCTA]10	+1A of 9, -1A of 11,
11	X	46	33	D2S441[CE11]-Chr02-GRCh38 68011947-68011994 [TCTA]11	+1A of 10,
13	S	58	227	D2S441[CE13]-Chr02-GRCh38 68011947-68011994 [TCTA]10 TTTA [TCTA]2	-1A of 14, -2AA of 15,
14	A	1506	6	D2S441[CE14]-Chr02-GRCh38 68011947-68011994 [TCTA]11 TTTA [TCTA]2	+1A of 13, -1A of 15,
14b	X	32	0	D2S441[CE14]-Chr02-GRCh38 68011947-68011994 [TCTA]10 TTTA TTTA [TCTA]2	
15	X	34	34	D2S441[CE15]-Chr02-GRCh38 68011947-68011994 [TCTA]12 TTTA [TCTA]2	+1A of 14,

CE = Capillary electrophoresis name, St = Status, A = Allele, S = Stutter, X = Artefact, (A) = manually excluded, Cov = Coverage, ES = Expected Stutter value, Seq = Comprehensive sequence name.

D12S391	3330	18	23
VWA	2195	16	19
D13S317	2396	9	11

CALLED READS
3330

STUTTER THRESHOLD
25 %

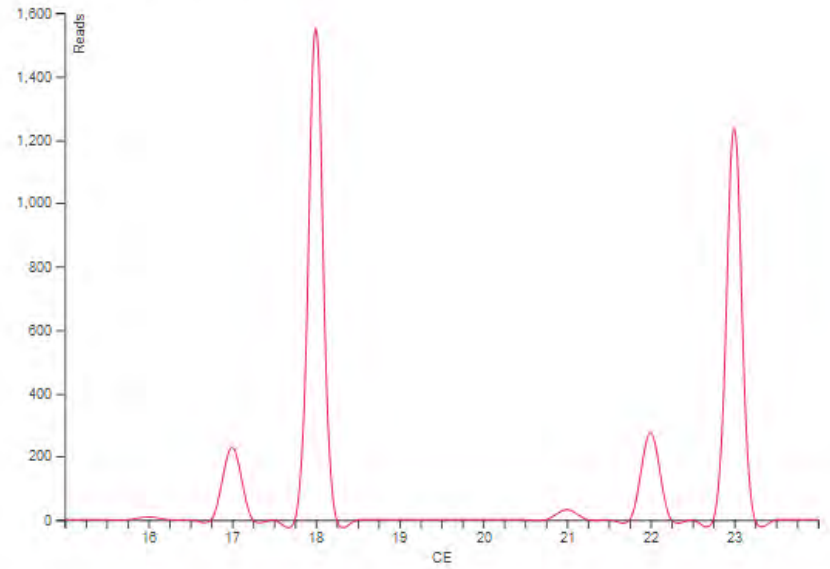
AUTOSOMAL
D1S1656
D2S1338
D2S441
TPOX
D3S1358
FGA
CSF1PO
D5S818
D7S820
D8S1179
D10S1248
TH01
D12S391
VWA
D13S317
PENTA E
D16S539
D18S51
D19S433

2800M / D12S391

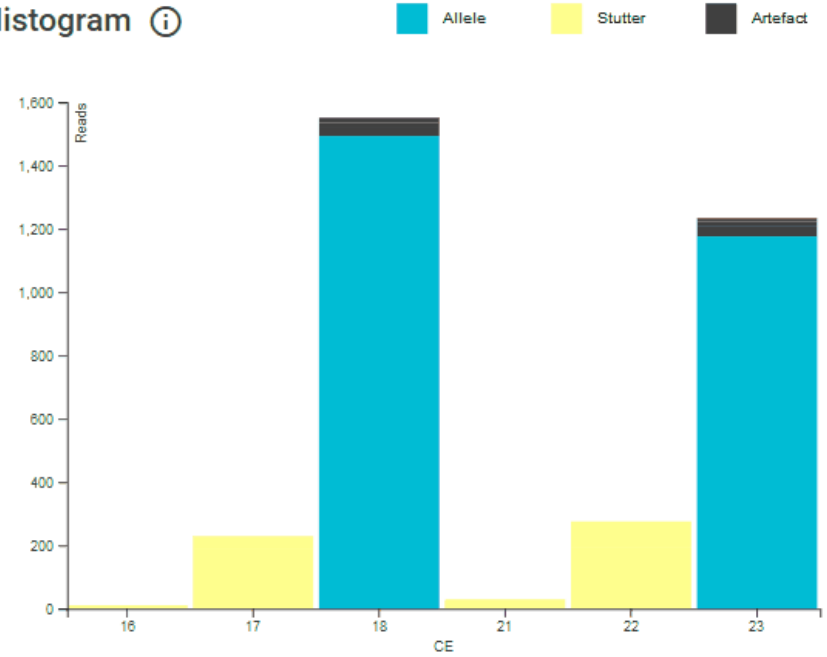
ANALYTICAL THRESHOLD 10 reads

CALLING THRESHOLD 2 % (66 reads)

Electropherogram



Histogram



Show alleles Show stutter Show artefacts Show source

REPORTED	CE	STATUS	COVERAGE	EXPECTED STUTTER	SEQUENCE
	16	Stutter	11	141	[AGAT]9 [AGAC]6 [AGAT]1
	17	Stutter	191	275	[AGAT]10 [AGAC]6 [AGAT]1

Dashboard

Analyses

Start a new analysis, list finished and running analyses, show and edit results

GO

Panels

Create and view custom collections of STR markers and define stutter thresholds

VIEW

Manual

Read the comprehensive documentation of toaSTR features and algorithms

READ

Changelog

All notable changes to this project will be documented in this log.
The format is based on Keep a Changelog and this project adheres to Semantic Versioning.
The ⓘ icon indicates important changes that may affect your results.

1.0.0-beta.12 (2017-10-23)

Content

Documentation

Introduction

In recent studies, massively parallel sequencing (MPS) has demonstrated its potential for the forensic analysis of short tandem repeats (STRs). In addition to nominal allele lengths, MPS can discover sequence variation in isoalleles (alleles that are identical by the number of repeats) and thus increase discriminatory power over conventional capillary electrophoresis (CE). However, considering currently available software, data analysis with routine use in mind turns out to be a cumbersome process, especially for laboratories with limited bioinformatical expertise.
We developed the web application toaSTR, a user-friendly tool for STR allele calling in MPS data independent of the instrument platform or the forensic kit used. toaSTR comes up with a clean, intuitive graphical user interface and well-documented parameter settings. Users have the ability to select from a wide range of STR markers to configure custom marker panels. This software supports both commercial and in-house multiplex PCR kits and various library preparation chemistries. Its sequence-based stutter-modelling algorithm automatically differentiates biological (iso-)alleles from stutter and artefacts to assist the interpretation of mixed samples.
toaSTR features a comprehensive data visualization with interactive diagrams and an adjustable tabular overview of sequence observations. Results are concordant with CE-based fragment analysis and can be exported for further analysis in biostatistical software or as an archivable/printable PDF document with sequence description in the ISFG-recommended nomenclature.

Citation:
Ganschow S, Wiegand P, Tiemann C (2017) toaSTR: A web-based forensic tool for the analysis of short tandem repeats in massively parallel sequencing data. http://dx.doi.org/10.1016/j.fsigss.2017.09.034

▲ top

Workflow recommendations

A common MPS genotyping workflow involves a multiplex PCR amplification of STR targets followed by library

Usage statistics

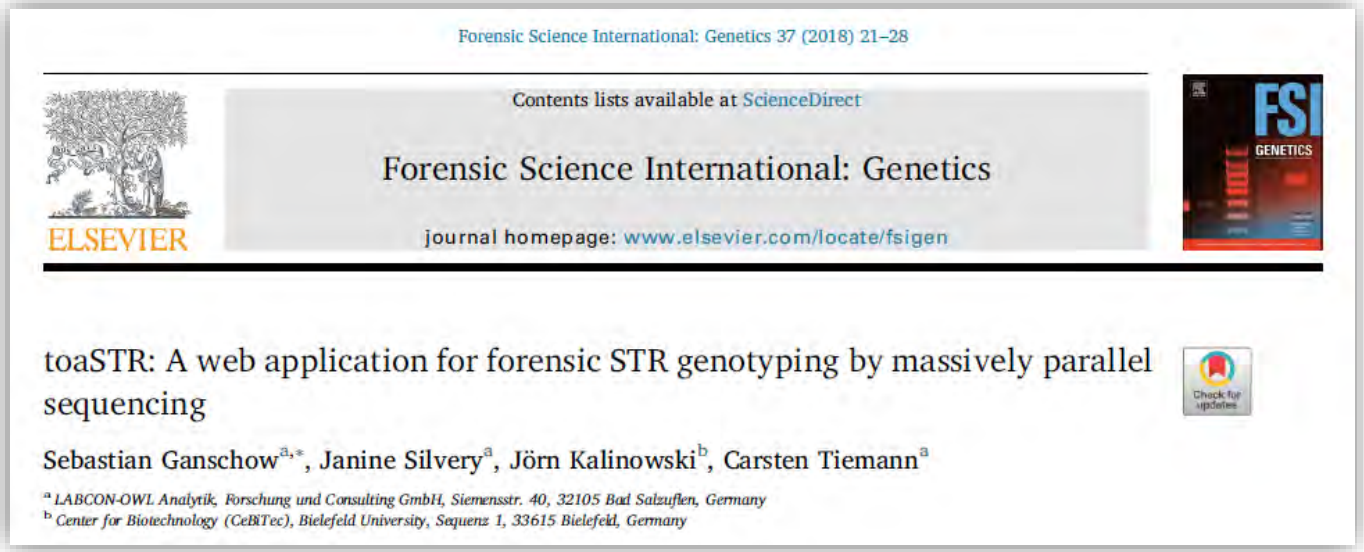
>60

registered users

>2,800

analyses performed

Future directions



Forensic Science International: Genetics 37 (2018) 21–28

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

toaSTR: A web application for forensic STR genotyping by massively parallel sequencing

Sebastian Ganschow[a,*], Janine Silvery[a], Jörn Kalinowski[b], Carsten Tiemann[a]

[a] LABCON-OWL Analytik, Forschung und Consulting GmbH, Siemensstr. 40, 32105 Bad Salzuflen, Germany
[b] Center for Biotechnology (CeBiTec), Bielefeld University, Sequenz 1, 33615 Bielefeld, Germany

- Refine the stutter model
- Capture flanking variation
- Interact with databases (e.g. NOMAUT)

Acknowledgement

ZIM
Zentrales Innovationsprogramm Mittelstand

Supported by:
Federal Ministry for Economic Affairs and Energy
on the basis of a decision by the German Bundestag

Ulm University Hospital
Institute for legal medicine
Division forensic genetics

LABCON-OWL
Analytik, Forschung und Consulting GmbH
Bad Salzuflen

Points of discussion

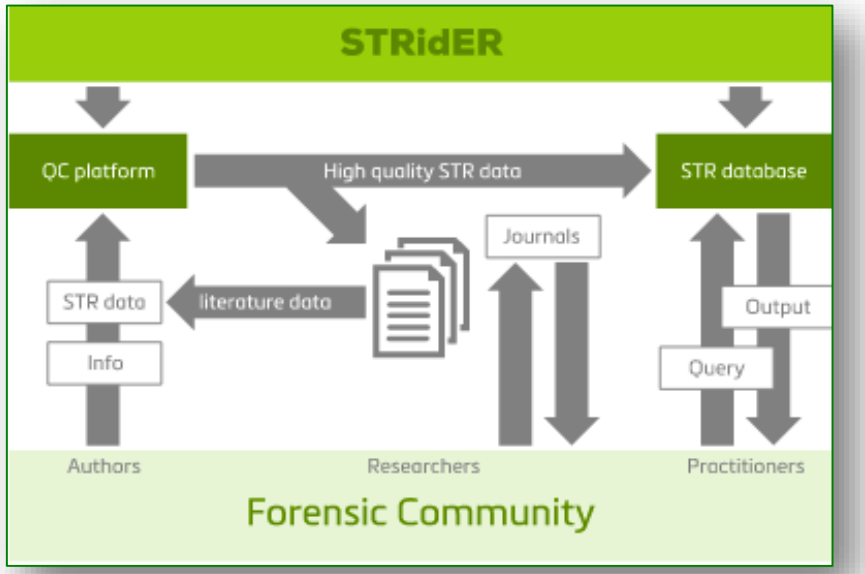
- Comprehensive **benchmarking** of MPS-STR genotyping tools on standardized data sets
- Standardization of **anchor** sequences (impacts the allelic detection capability):
  - locus specificity (mismatch tolerance)
  - kit compatibility
  - CE concordance
- Guideline for **quality control** of raw sequencing data
  - Quality filtering or quality trimming
  - How to measure quality? Which QC tools and quality thresholds?
  - Implement QM in genotyping software?
- Sequencing mode: single-end, paired-end, **merging** of paired end reads

STRAND working group
align | name | define

STR Databasing and Quality control

STRidER

shortened version for circulation



D8S1179

Allele	AUSTRIA	BELGIUM	BOSNIA AND HERZEGOWINA	CZECH REPUBLIC	DENMARK	FINLAND	FRANCE	GERMA
	222	206	171	200	200	230	208	
8	1.8018e-2	7.2816e-3	5.8480e-3	7.5000e-3	1.5000e-2	1.7391e-2	2.4039e-2	1.2840
9	1.8018e-2	1.2136e-2	8.7719e-3	5.0000e-3	1.0000e-2	8.6956e-3	9.6154e-3	1.2840
10	9.4595e-2	8.7379e-2	5.8479e-2	5.5000e-2	9.7500e-2	8.2609e-2	8.4135e-2	8.7613
11	1.0135e-1	9.7087e-2	3.2164e-2	1.0000e-1	8.0000e-2	1.3261e-1	8.8942e-2	7.7795
12	1.6216e-1	1.5049e-1	1.8713e-1	1.5250e-1	1.3000e-1	1.3261e-1	1.3462e-1	1.4199
13	2.9054e-1	3.1311e-1	3.4210e-1	3.5000e-1	3.4500e-1	3.5217e-1	3.1490e-1	3.1269
14	1.9144e-1	1.6990e-1	2.1637e-1	2.1250e-1	2.0750e-1	1.8478e-1	2.0433e-1	1.9864
15	1.0360e-1	1.2379e-1	1.1403e-1	9.7500e-2	8.5000e-2	5.6522e-2	1.0336e-1	1.1933
16	1.8018e-2	3.3981e-2	3.2164e-2	2.0000e-2	2.0000e-2	1.5217e-2	3.6058e-2	3.0211
17	2.2522e-3	4.8544e-3	2.9240e-3		5.0000e-3	1.0870e-2		6.0423

Martin Bodner, Walther Parson

STRAND WG meeting, London, 12 April 2019

Outline of presentation

1. (short) history of activities
2. STRidER and quality control
3. related activities

Early database considerations

ELSEVIER

Forensic Science International

Volume 131, Issues 2–3, 28 January 2003, Pages 184-196

A comparison of adjustment methods to test the robustness of an STR DNA database comprised of 24 European populations

Peter Gill [a], Lindsey Foreman [b], John S Buckleton [c], Christopher M Triggs [d], Heather Allen [a]

An aim of the European Network of Forensic Science Institutes (**ENFSI**) is to produce a **DNA database of second generation multiplex (SGM) STR profiles** that is representative of the resident cosmopolitan populations. To achieve this, data were collected from 24 different populations. All of the data were combined to form one database of 5,700 profiles from which allele proportions were calculated.

ENFSI STR database: STRbASE

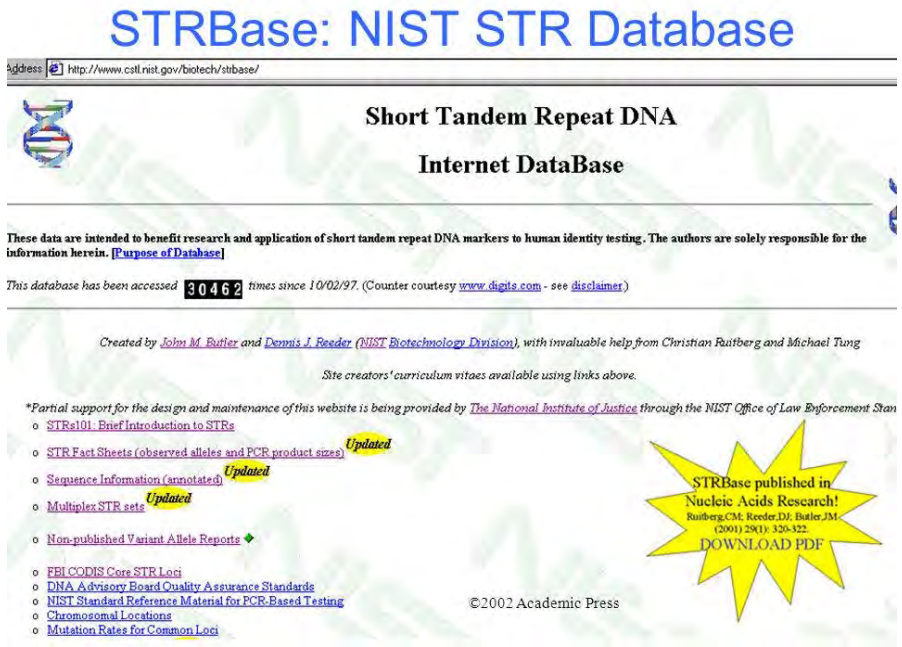
- established in 2004
- **high quality** autosomal STR database
- **19** European countries, **16** autosomal STR loci
- allele **frequency tables**, also for download and import into other software
- **query function:** single and batch query
- hosted by

Carefully enhancing STRbASE

- **STRbASE** became well-established
- **time for update:** more markers, more populations
- **new** elements?

- **name changed** to STRidER

- **maintaining the tried-and-tested**



How will STRidER be used

Survey sent out to all ENFSI DNA WG participants (2016)

Use of adjustment methods for genotype probabilities (only complete profiles)

- which formulae to present?
- which are the preferred methods?

27 labs returned answers

Survey on adjustment factors

- vast majority of users **does not use adjustments:**

  „*not necessary*“

  calculations using **third party software**

- majority of users prefers to receive **only unadjusted frequencies**
- exception: minor allele frequency **MAF (5/2n)**

- **data quality is the most important**

Formulae

Actual matching probability

$P_m = 2p_ip_j$	Heterozygotes
$P_m = p_i^2$	Homocygotes
$P_m = 2p_i - p_i^2$	Single alleles

Balding & Nichols (1994)

$P_m = \frac{2(\Theta + (1-\Theta)p_i)(\Theta + (1-\Theta)p_j)}{(1+\Theta)(1+2\Theta)}$	Balding-Nichols heterozygotes
$P_m = \frac{(2\Theta + (1-\Theta)p_i)(3\Theta + (1-\Theta)p_i)}{(1+\Theta)(1+2\Theta)}$	Balding-Nichols homocygotes
$P_m = \frac{(2\Theta + (1-\Theta)p_i)(3\Theta + (1-\Theta)p_i)}{(1+\Theta)(1+2\Theta)}$	Balding-Nichols single alleles

Balding size bias correction (1995)

$P_m = \frac{2(x_i+2)(x_j+2)}{(n+4)^2}$	heterozygotes
$P_m = \frac{(x_i+2)^2}{(n+4)^2}$	homocygotes

Confidence Intervals (NRC-Report 1996)

$Var(ln(2p_ip_j)) \approx \frac{p_i + p_j - 4p_ip_j}{2Np_ip_j}$	Confidence interval heterozygotes
$Var(ln(2p_i^2)) \approx \frac{2(1-p_i)}{Np_i}$	Confidence interval homocygotes
$Var(ln(2p_i - p_i^2)) \approx \frac{2(1-p_i)^3}{Np_i(2-p_i)^2}$	Confidence interval single alleles
$\Gamma = \sqrt{V_1 + V_2 + .. + V_k}$	
$Upperbound = log^{-1}(log_{10}(P_m)) + 1.96\Gamma$	

Formulae

Actual matching probability

$P_m = 2p_i p_j$	Heterozygotes
$P_m = p_i^2$	Homocygotes
$P_m = 2p_i - p_i^2$	Single alleles

Data quality is crucial

- **high quality data is the most important feature of** STRidER
- **<u>quality control</u>** is necessary towards better STR population data
- assistance to authors, reviewers and editors

- enormous experience at  from EMPOP mtDNA database.

Forensic Science International: Genetics 5 (2011) 138–141

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

observations indicating the need for STR data QC

Inspecting close maternal relatedness: Towards better mtDNA population samples in forensic databases

Martin Bodner [a], Jodi A. Irwin [b], Michael D. Coble [b,1], Walther Parson [a,*]

[a] Institute of Legal Medicine, Innsbruck Medical University, Müllerstr. 44, 6020 Innsbruck, Austria
[b] Armed Forces DNA Identification Laboratory, 1413 Research Blvd, Rockville, MD 20850, USA

Gomes et al. BMC Genomics (2015) 16:70
DOI 10.1186/s12864-014-1201-x

BMC Genomics

RESEARCH ARTICLE Open Access

Human settlement history between Sunda and Sahul: a focus on East Timor (Timor-Leste) and the Pleistocenic mtDNA diversity

Sibylle M Gomes[1†], Martin Bodner[2†], Luis Souto[1,3], Bettina Zimmermann[2], Gabriela Huber[2], Christina Strobl[2], Alexander W Röck[2], Alessandro Achilli[4,5], Anna Olivieri[4], Antonio Torroni[4], Francisco Côrte-Real[6] and Walther Parson[2,7*]

analysis and detection conditions [100,121-123]. After typing 15 autosomal STR loci and the amelogenin length polymorphism, pedigree construction, and likelihood ratio (LR) calculation using reported STR allele frequencies [7] [correcting the 10.2 allele frequency of D18S51 to 0.0 (L Souto, *pers. comm.*)], no donor pair revealed close maternal relatedness (*i.e.*, mother-child and sibling constellations) applying a cut-off LR of 1,000 [124,125]

STRidER

JOURNAL OF FORENSIC SCIENCES
Explore this journal >

Open Access

Erratum

Erratum

Tamyra R. Moretti Ph.D., Bruce Budowle Ph.D., John S. Buckleton Ph.D.

First published: 3 June 2015 Full publication history

DOI: 10.1111/1556-4029.12806 View/save citation

Cited by (CrossRef): 3 articles Check for updates Citation tools

Am score 2

View issue TOC
Volume 60, Issue 4
July 2015
Pages 1114–1116

This article corrects:

*Names of commercial manufacturers are provided for identification purposes only, and inclusion does not imply endorsement of the manufacturer or its products and services by the FBI. The views are those of the authors and do not necessarily reflect the official policy or position of the FBI or the US government.

Reference: Budowle B, Moretti TR, Baumstark AL, Defenbaugh DA, Keys KM. Population data on the thirteen CODIS core short tandem repeat loci in African Americans, US Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians. J Forensic Sci 1999;44(6):1277–86.

observations indicating the need for STR data QC

Re-typing of a widely applied population dataset after 16 years revealed a certain number of **clerical, technical, and data/sample processing errors**

Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER)

Martin Bodner[a], Ingo Bastisch[b], John M. Butler[c], Rolf Fimmers[d], Peter Gill[e,f], Leonor Gusmão[g,h,i], Niels Morling[j], Christopher Phillips[k], Mechthild Prinz[l], Peter M. Schneider[m], Walther Parson[a,n,*]

Content

I) Positioning **STRidER** relative to other existing databases (STRbase, ALFRED, popSTR, popAffiliator, ALLST*R); **important element of QC**

II) Rationale, concept and workflow of **QC** via **STRidER**

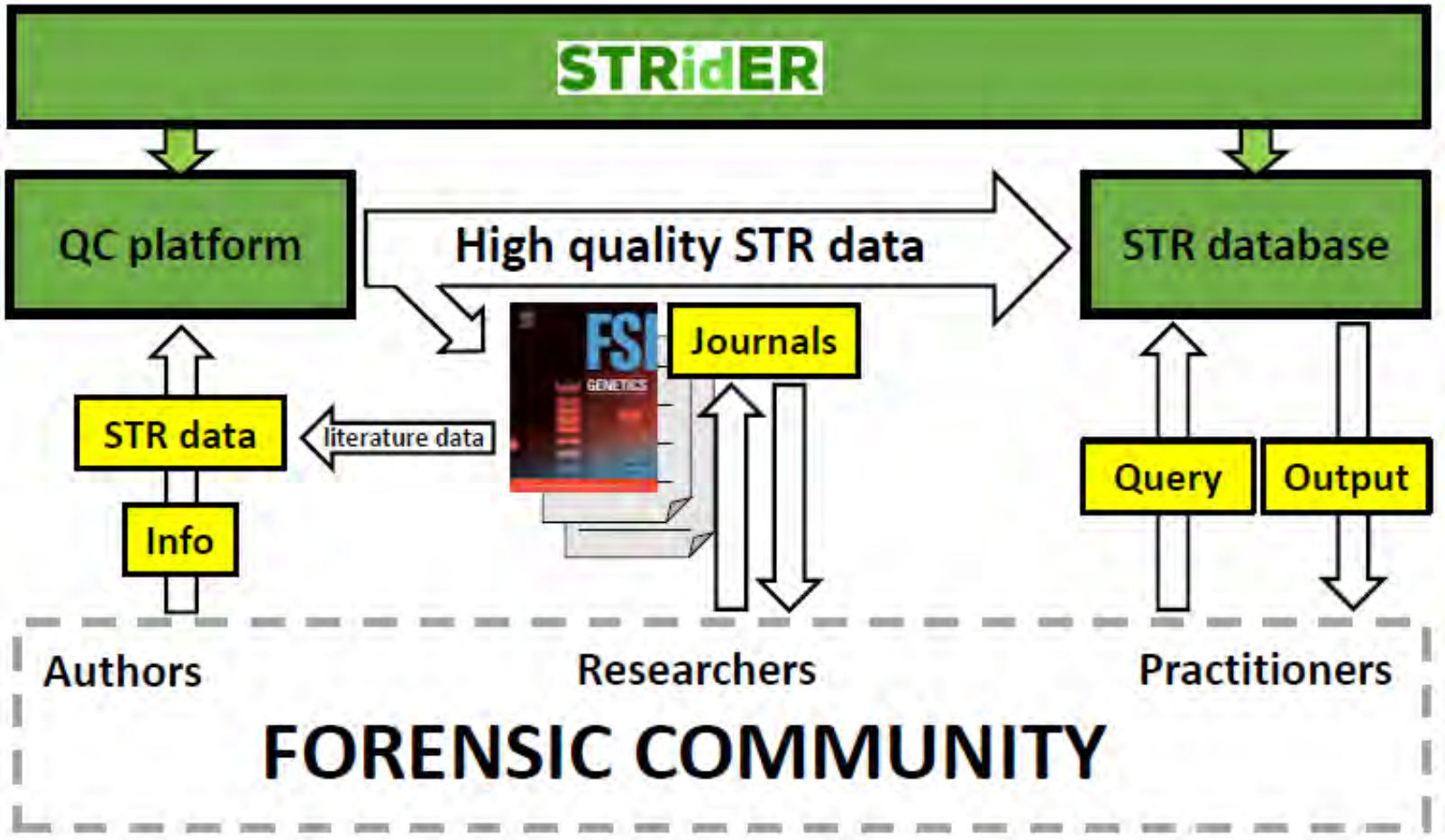
III) Benefits to forensic and other scientific community

IV) Transparency, traceability and protection of data

V) Outloook: **STR sequence data** in **STRidER** (MPS)

STRidER

STRs for identity ENFSI Reference database, v2



The ENFSI DNA Working Group provides an **updated version** to the previously published 'ENFSI DNA WG STR Population Database'

STRidER

STRs for identity ENFSI Reference database, v2

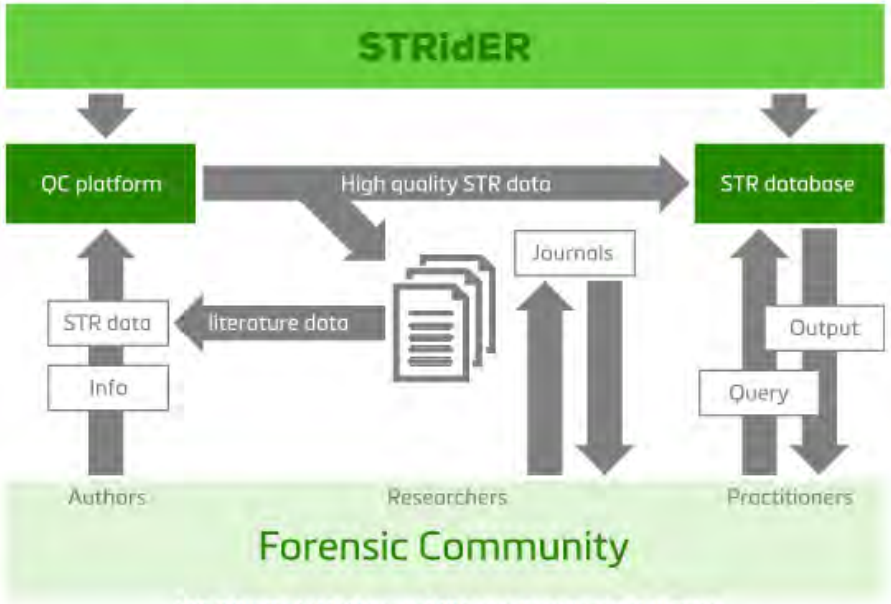
HOME | QUERY | BATCH QUERY | ABOUT | FREQUENCIES | FORMULAE | QUALITY CONTROL | STR SEQ NOMENCLATURE | CONTACT | TERMS OF USE

Welcome to STRidER!

STRidER (**STR**s for **I**dentity **E**NFSI **R**eference Database) is the expanded and enhanced version of the ENFSI STRbASE (2004-2016). This curated online high quality STR allele frequency population database enables scientifically reliable **STR genotype probability estimates** and provides **quality control** of autosomal STR data. A suite of software tools has been developed at the Institute of Legal Medicine, Medical University of Innsbruck to scrutinize STR population data and thus increase the quality of datasets to ensure reliable allele frequency estimates. STRidER acts as **frequency database and software platform** for the development of novel tools for STR data QC and other forensic analyses.

STRidER serves the STR community in forensics and beyond in inter-related ways:

- The high-quality autosomal STR allele frequency database can be directly queried
- Allele frequency tables of STR loci from diverse populations can be downloaded and used for third party software
- Centralized STR data quality control is offered prior to publication
- Accepted datasets will become rapidly available online and receive a unique and traceable STRidER accession number
- Allele frequencies and forensic/population genetic parameters are calculated from datasets
- Individual STR genotypes are not accessible on STRidER to comply with privacy regulations



STRidER in the field of forensic STR typing (from Bodner et al. 2016)

The **concept of STRidER** has been developed together with the DNA Commission of the ISFG and is outlined in *Bodner M, Bastisch I, Butler JM, Fimmers R, Gill P, Gusmão L, Morling N, Phillips C, Prinz M, Schneider PM, Parson W (2016) Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER). Forensic Sci Int Gen 24:97-102.*

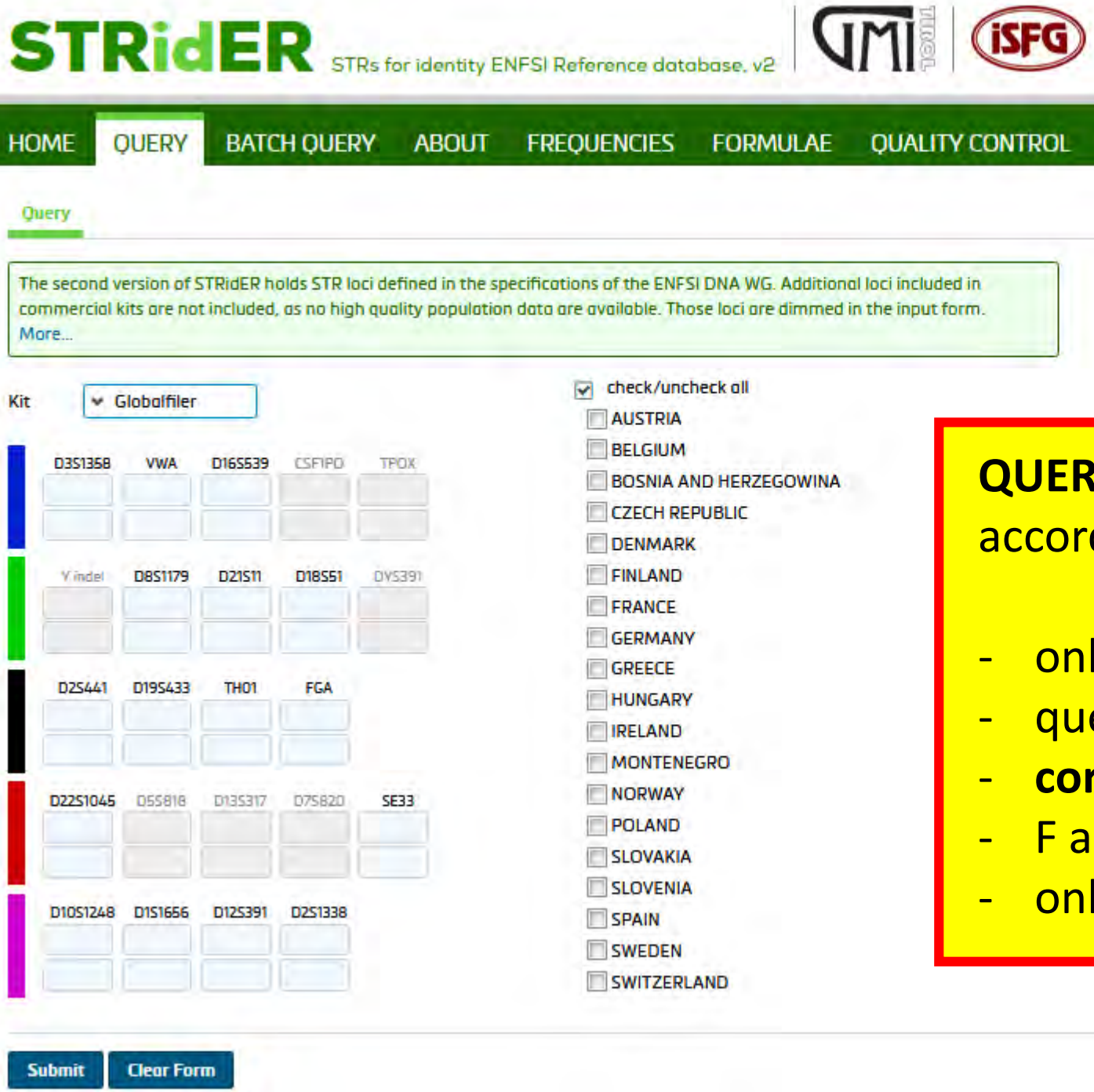
The STRidER online platform is work in progress. Additional datasets and features will continuously become available. To receive periodic news and stay updated about STRidER, register here for the STRidER newsletter.

Please consider **citing STRidER** [https://www.isfg.org/Publication;Bodner2016] when using it with your research.

https://strider.online/

ongoing addition of data and functions

STRidER Newsletter enabled (via HOME tab)



QUERY
according to questionnaire among ENFSI labs

- only **uncorrected AMP** is calculated
- query profile is not added to database
- **correction factors not offered** any longer
- F alleles no longer allowed
- only correction used is **MAF (5/2n)**

STRidER

STRs for identity ENFSI Reference database, v2

HOME | QUERY | BATCH QUERY | ABOUT | FREQUENCIES | FORMULAE | QUALITY CONTROL | STR SEQ NOMENCLATURE | CONTACT | TERMS OF USE

Query

The CSV file requires *commas (,)* as delimiters and *double quotes (")* as field enclosure characters.
Download a sample CSV file.

File format ◉ CSV ○ GeneMapper

CSV file Durchsuchen... Keine Datei ausgewählt.

- ☑ check/uncheck all
- ☐ AUSTRIA
- ☐ BELGIUM
- ☐ BOSNIA AND HERZEGOWINA
- ☐ CZECH REPUBLIC
- ☐ DENMARK
- ☐ FINLAND
- ☐ FRANCE
- ☐ GERMANY
- ☐ GREECE
- ☐ HUNGARY
- ☐ IRELAND
- ☐ MONTENEGRO
- ☐ NORWAY
- ☐ POLAND
- ☐ SLOVAKIA
- ☐ SLOVENIA
- ☐ SPAIN
- ☐ SWEDEN
- ☐ SWITZERLAND

BATCH QUERY

Submit

Frequencies

These tables include allele frequencies and number of samples (n) from the most recent database release sorted by marker and country.
In these tables, „1" represents all rare alleles shorter than the accepted allele categories. The value „99" represents all rare alleles longer than the accepted categories.

This data can be downloaded as XML file.

VWA

Allele	AUSTRIA	BELGIUM	BOSNIA AND HERZEGOWINA	CZECH REPUBLIC	DENMARK	FINLAND	FRANCE	GERMANY	GREECE	HUNGARY	IRELAND	MONTENEGRO	NORWAY	POLAND	SLOVAKIA	SLOVENIA	SPAIN	SWEDEN	SWITZERLAND	Europe	Entire Database
	222	206	171	200	200	230	208	662	208	224	304	200	202	206	247	207	449	424	402	5172	5172
11								7.5529e-4												9.6674e-5	9.6674e-5
12									4.8077e-3											1.9335e-4	1.9335e-4
13			1.1696e-2					2.2659e-3	2.4038e-3	2.2321e-3			2.4753e-3		2.0243e-3	2.4155e-3	6.6815e-3	1.1792e-3		1.8368e-3	1.8368e-3
14	1.0586e-1	1.0680e-1	1.1111e-1	1.0000e-1	7.0000e-2	1.3043e-1	8.6539e-2	9.7432e-2	9.3750e-2	1.1161e-1	1.1349e-1	1.4500e-1	8.6634e-2	7.7670e-2	1.1943e-1	1.0145e-1	1.1024e-1	9.4340e-2	1.0448e-1	1.0335e-1	1.0335e-1
15	9.2342e-2	1.2136e-1	1.2573e-1	9.7500e-2	9.7500e-2	5.2174e-2	1.2740e-1	1.0347e-1	7.9327e-2	1.1384e-1	1.0197e-1	9.0000e-2	9.9010e-2	8.4951e-2	1.1943e-1	1.2077e-1	1.2361e-1	8.9623e-2	1.0697e-1	1.0296e-1	1.0296e-1
16	1.7568e-1	1.9903e-1	2.0468e-1	1.7500e-1	2.6000e-1	1.7609e-1	2.4038e-1	2.2130e-1	1.6827e-1	2.0536e-1	2.1875e-1	1.7500e-1	2.2277e-1	2.2330e-1	1.9231e-1	1.8599e-1	2.4276e-1	2.0991e-1	2.0647e-1	2.0872e-1	2.0872e-1
17	2.8604e-1	2.7185e-1	2.3977e-1	3.1250e-1	2.3000e-1	2.7174e-1	2.3317e-1	2.5453e-1	3.1731e-1	3.0134e-1	2.7138e-1	2.8750e-1	2.8960e-1	2.7670e-1	2.7530e-1	2.8985e-1	2.7171e-1	2.6533e-1	2.7736e-1	2.7291e-1	2.7291e-1
18	2.5901e-1	2.0146e-1	2.1053e-1	2.2750e-1	2.4000e-1	2.0435e-1	2.1154e-1	2.2054e-1	2.4279e-1	1.7634e-1	1.9243e-1	2.1250e-1	1.9802e-1	2.4757e-1	2.0445e-1	2.1739e-1	1.7038e-1	2.4174e-1	2.0896e-1	2.1384e-1	2.1384e-1
19	7.2072e-2	8.0097e-2	9.0643e-2	7.2500e-2	8.2500e-2	1.3696e-1	8.6539e-2	8.6103e-2	7.4519e-2	7.1429e-2	9.3750e-2	7.2500e-2	8.6634e-2	8.0097e-2	7.6923e-2	5.5556e-2	6.1247e-2	7.9009e-2	8.2090e-2	8.0917e-2	8.0917e-2
20	9.0090e-3	1.9418e-2	5.8480e-3	1.5000e-2	1.7500e-2	2.1739e-2	1.4423e-2	1.2840e-2	1.4423e-2	1.5625e-2	8.2237e-3	1.7500e-2	1.4852e-2	9.7087e-3	1.0122e-2	2.1739e-2	1.3363e-2	1.6509e-2	1.3682e-2	1.4115e-2	1.4115e-2
21					2.5000e-3	6.5217e-3		7.5529e-4	2.4038e-3	2.2321e-3						4.8309e-3		2.3585e-3		1.0634e-3	1.0634e-3

TH01

Allele	AUSTRIA	BELGIUM	BOSNIA AND HERZEGOWINA	CZECH REPUBLIC	DENMARK	FINLAND	FRANCE	GERMANY	GREECE	HUNGARY	IRELAND	MONTENEGRO	NORWAY	POLAND	SLOVAKIA	SLOVENIA	SPAIN	SWEDEN	SWITZERLAND	Europe	Entire Database
	222	206	171	200	200	230	208	662	208	224	304	200	202	206	247	207	454	425	402	5178	5178
5	2.2522e-3	2.4272e-3						1.5106e-3		2.2321e-3			2.4753e-3					1.1765e-3	1.2438e-3	7.7250e-4	7.7250e-4
6	2.0946e-1	2.1359e-1	2.7778e-1	2.3750e-1	2.3750e-1	1.9565e-1	2.4038e-1	2.2659e-1	2.6923e-1	2.2098e-1	2.3191e-1	3.3500e-1	2.2030e-1	2.0631e-1	2.3077e-1	2.3430e-1	2.5330e-1	1.9529e-1	2.1890e-1	2.3165e-1	2.3165e-1
7	1.2613e-1	1.8689e-1	1.3158e-1	1.6000e-1	1.5250e-1	2.1522e-1	1.5865e-1	1.4804e-1	1.2019e-1	1.6071e-1	2.0230e-1	1.0750e-1	2.3267e-1	1.5291e-1	1.4372e-1	1.3043e-1	1.7181e-1	1.9059e-1	1.7537e-1	1.6348e-1	1.6348e-1
8	1.1486e-1	1.2136e-1	1.3158e-1	1.0000e-1	1.1250e-1	1.1522e-1	1.2019e-1	1.3897e-1	1.3462e-1	1.1384e-1	9.2105e-2	1.4500e-1	8.6634e-2	8.4951e-2	1.2348e-1	1.2560e-1	1.0573e-1	9.1765e-2	1.2189e-1	1.1529e-1	1.1529e-1
8.3	2.2522e-3		2.9240e-3					7.5529e-4							2.0243e-3				1.2438e-3	4.8281e-4	4.8281e-4
9	1.6441e-1	1.5049e-1	1.9883e-1	1.7750e-1	1.5250e-1	1.8913e-1	1.8269e-1	1.7523e-1	2.1635e-1	2.1205e-1	1.3651e-1	1.7750e-1	1.3366e-1	2.1359e-1	2.0040e-1	1.9565e-1	1.9053e-1	1.5765e-1	1.7413e-1	1.7748e-1	1.7748e-1
9.3	3.6712e-1	3.1796e-1	2.4561e-1	3.2250e-1	3.4000e-1	2.7826e-1	2.8365e-1	2.9305e-1	2.3077e-1	2.7679e-1	3.2895e-1	2.2250e-1	3.1436e-1	3.3010e-1	2.9555e-1	3.0193e-1	2.6982e-1	3.5647e-1	2.9353e-1	2.9973e-1	2.9973e-1
10	1.1261e-2	7.2816e-3	1.1696e-2	2.5000e-3	5.0000e-3	6.5217e-3	1.4423e-2	1.5861e-2	2.8846e-2	1.3393e-2	8.2237e-3	1.2500e-2	9.9010e-3	1.2136e-2	4.0486e-3	1.2077e-2	8.8106e-3	7.0588e-3	1.3682e-2	1.1008e-2	1.1008e-2
10.3	2.2522e-3																			9.6562e-5	9.6562e-5

Formulae

Actual matching probability

$P_m = 2p_i p_j$ Heterozygotes

$P_m = p_i^2$ Homozygotes

A minimum allele frequency of 5/2n [1] is used for calculations.

[1] National Research Council. (1996) The evaluation of forensic DNA evidence. National Academy Press, Washington D.C.

MPS STR Nomenclature: ISFG considerations

Forensic Science International: Genetics 22 (2016) 54–63

Contents lists available at ScienceDirect

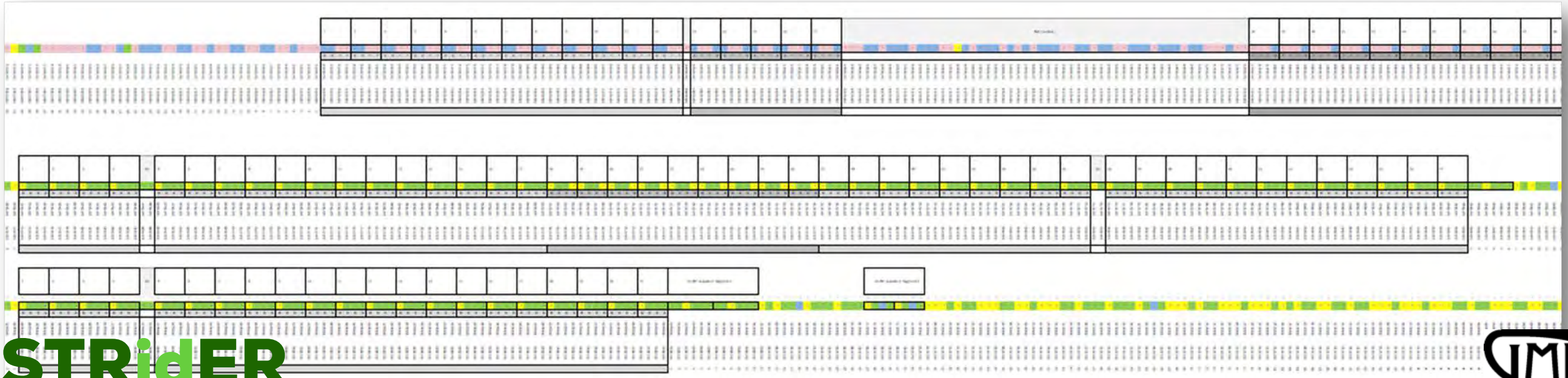
Forensic Science International: Genetics

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

Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements

Walther Parson[a,b,*], David Ballard[c], Bruce Budowle[d,e], John M. Butler[f], Katherine B. Gettings[f], Peter Gill[g,h], Leonor Gusmão[i,j,k], Douglas R. Hares[l], Jodi A. Irwin[l], Jonathan L. King[d], Peter de Knijff[m], Niels Morling[n], Mechthild Prinz[o], Peter M. Schneider[p], Christophe Van Neste[q], Sascha Willuweit[r], Christopher Phillips[s]

ESM1 file showing STR sequence structure + flanking region



STRidER

STRidER STRs for identity ENFSI Reference database, v2

HOME | QUERY | BATCH QUERY | ABOUT | FREQUENCIES | FORMULAE | QUALITY CONTROL | STR SEQ NOMENCLATURE | CONTACT | TERMS OF USE

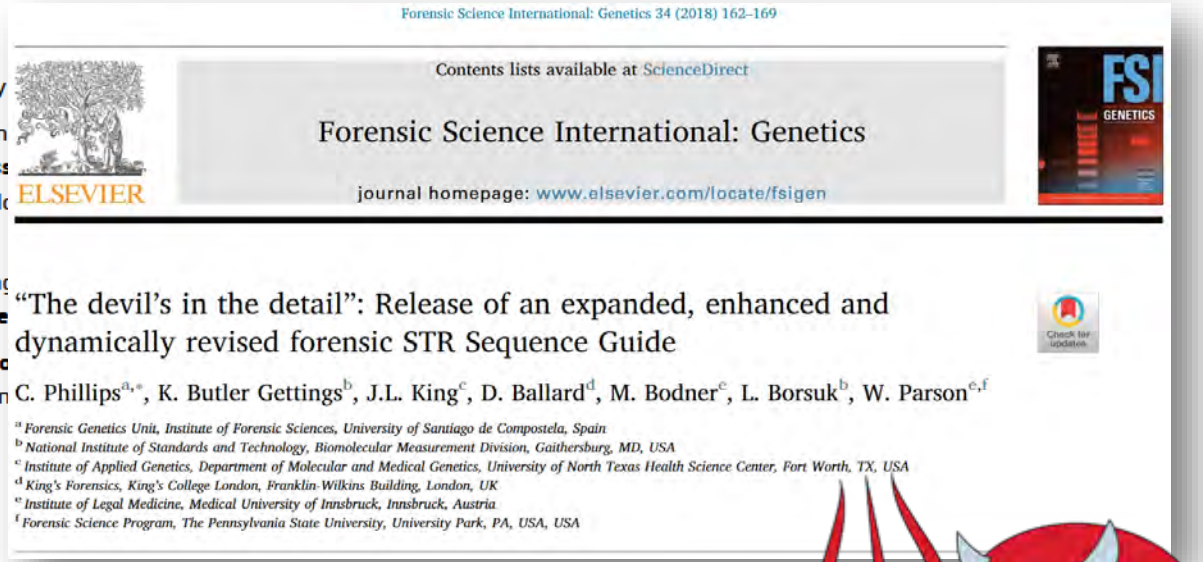
STR Sequence Nomenclature

The **'Forensic STR Sequence Structure' file** is an updated set of forensic STR sequences that was originally

Parson W, Ballard D, Budowle B, Butler JM, Gettings KB, Gill P, Gusmão L, Hares DR, Irwin JA, King JL, de Kn
Willuweit S, Phillips C: **Massively parallel sequencing of forensic STRs: Considerations of the DNA commiss
(ISFG) on minimal nomenclature requirements.** Forensic Science International Genetics 2016, 22: 54-63 (do
http://www.isfg.org/Publication;Parson2016).

The original file has been expanded, enhanced and revised as described in the publication Phillips C, Getting
"The devil's in the detail": Release of an expanded, enhanced and dynamically revised forensic STR Seque

The most recent version of this permanently curated and updated Forensic STR sequence structure file co
here. The updates since the last version are reported in a change log contained in the file. To receive inform
structure file and to stay updated about STRidER, register here for the STRidER newsletter.


Forensic Science International: Genetics 34 (2018) 162–169

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

"The devil's in the detail": Release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide

C. Phillips[a,*], K. Butler Gettings[b], J.L. King[c], D. Ballard[d], M. Bodner[e], L. Borsuk[b], W. Parson[e,f]

[a] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain
[b] National Institute of Standards and Technology, Biomolecular Measurement Division, Gaithersburg, MD, USA
[c] Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, Fort Worth, TX, USA
[d] King's Forensics, King's College London, Franklin-Wilkins Building, London, UK
[e] Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria
[f] Forensic Science Program, The Pennsylvania State University, University Park, PA, USA, USA

Updated **Forensic STR Sequence Structure Guide and change log** is available on **STRidER**

STRAND working group
align | name | define

v4 online, v5 in prep.
STRAND WG + user feedback

STRidER STRs for identity ENFSI Reference database, v2

HOME | QUERY | BATCH QUERY | ABOUT | FREQUENCIES | FORMULAE | QUALITY CONTROL | STR SEQ NOMENCLATURE | CONTACT | TERMS OF USE

Quality Control

STRidER provides quality control of autosomal STR data. STRidER is accepting datasets from diverse worldwide populations and forensically relevant autosomal STR markers that comply with ethical standards. Minimum requirements of journals might apply when datasets are intended for peer reviewed publication. A suite of software tools has been developed to scrutinize STR population data and thus increase the quality of datasets to ensure reliable allele frequency estimates. The board of the International Society of Forensic Genetics (ISFG) and the editors of Forensic Science International: Genetics invited STRidER to logistically organize and perform quality control (QC) of autosomal STR population data in the course of manuscript preparations for the journal [1]. Before STR population papers are put forward to the editors for review, the authors are requested to submit the data to STRidER. After positive evaluation, the authors will be contacted with the respective STRidER accession numbers that serve as indicator of successful QC for the editors and reviewers. The necessary steps for submission of CE-based STR data to STRidER are outlined below. Please contact STRidER in case you want to submit STR sequence data.

Step 1

Prepare your **STR data file** as shown in the example file that can be downloaded and used as template. It is a tab delimited text file that can be created using standard text software or MS Excel (then, save file under .txt format). The minimum requirements for population datasets for *Forensic Science International: Genetics* [1] are 15 autosomal STR loci typed in 500 samples (for exceptional populations, the latter number can be smaller, please contact STRidER before submission).

The initial lines (identified using the "#" symbol) specify details of the dataset and origin of the samples. Line 1 must contain a description of population(s) reported (e.g., the title of the study), number of samples, geographic origin, and the number of STR loci. Line 2 must indicate the contact author's name with email address. Further text lines marked with "#" can be included for comments or description of the detailed geographic background and the appropriate metapopulation affiliation of the genotypes. Lines below these text lines list the original STR genotypes. Allele nomenclature criteria are applied as described in the "About" tab of this website. The order of loci does not matter. Alleles for the same locus have to be reported in adjacent columns. Loci names must not contain spaces. Report both alleles for homozygous loci. Use "." instead of "," for incomplete alleles, e.g. "9.3" not "9,3". Note that only complete genotypes are accepted. It is imperative that STR genotypes are reported individually and unshuffled using a unique identifier for each genotype in the dataset. The names are necessary for correspondence.

Also prepare an **accompanying STR information file** per population containing additional information on the dataset as outlined in the example information file. This information might be necessary for evaluation of the dataset. Keep raw data files available for any later inquiries. Please also send the allele frequency table and forensic parameters you have calculated from the dataset (no special format required).

Step 2

Submit your files to STRidER by email (see contact). The genotype data should be submitted as a file containing the following notation: Author_country_number of samples.txt (e.g. Parson_AUT_573.txt), the accompanying file should be named Author_country_number of samples_Info.xls or .xlsx (e.g. Parson_AUT_573_Info.xls). The data will be quality checked as outlined in [2] using in-house software.

Step 3

After STRidER evaluation, communication with respect to individual genotypes may follow. Once your data passed QC you will receive the STRidER accession number(s) for your data together with allele frequencies and forensic/population genetic parameters calculated from the dataset(s). Please provide accession number(s) to the journal editor and cite STRidER [2] in your manuscript.

Step 4

Data that successfully passed QC will be uploaded onto the STRidER database.

References

[1] Gusmão L, Butler JM, Linacre A, Parson W, Roewer L, Schneider PM, Carracedo A (2017) Revised guidelines for the publication of genetic population data; Forensic Sci Int Gen 30:160-163

[2] Bodner M, Bastisch I, Butler JM, Fimmers R, Gill P, Gusmão L, Morling N, Phillips C, Prinz M, Schneider PM, Parson W (2016) Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER); Forensic Sci Int Gen 24:97-102

QUALITY CONTROL instructions

FSI:G Guidelines 2017

Forensic Science International: Genetics xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Editorial

Revised guidelines for the publication of genetic population data

Since 2007, when the journal was launched, the number of submissions of manuscripts reporting population genetic data to FSI: Genetics has continuously increased. This type of data is very welcome, considering the importance of having accurate estimates

1. Quality control of population DNA databases

To improve the quality of the submitted to the journal, in 2010 [1],

STRidER QC
is now **mandatory**

Gusmao et al., 2017

To also improve the quality of the data generated from autosomal STRs, the ISFG executive board and the editors of FSI: Genetics have now invited STRidER (http://strider.online), a publicly available, centrally curated online allele frequency database and quality control platform for autosomal STRs [5], to logistically organize and perform quality control before autosomal STR manuscripts are put forward for review. Upon successful QC, STRidER accession numbers will be assigned to the submitted population data that serve as indicators of successful QC for the editors and reviewers. The necessary steps for submission of autosomal STR genotypes to STRidER are outlined below.

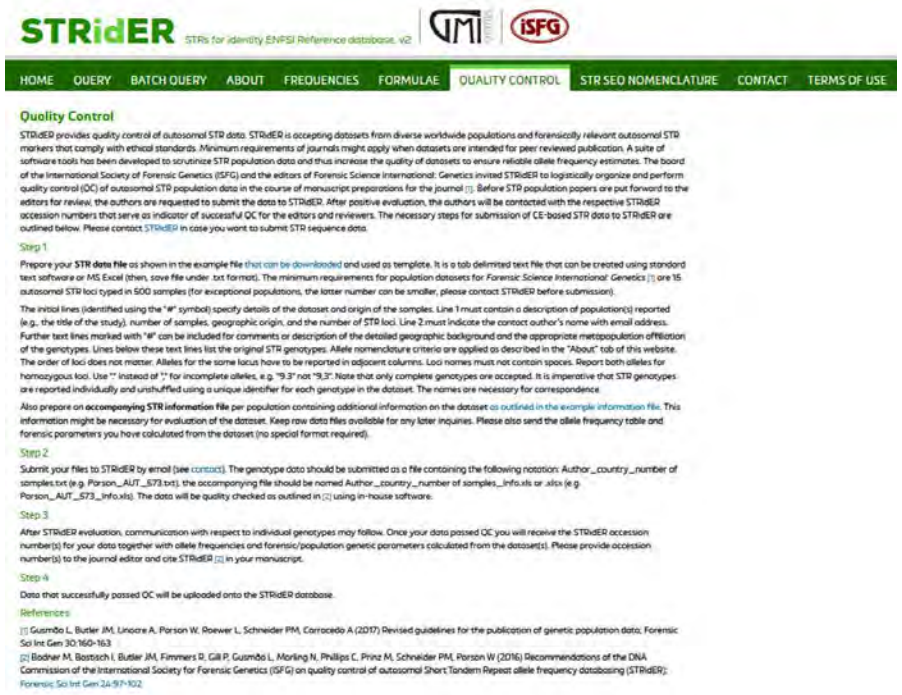
STRidER

Submission of STR data for QC to STRidER

- **instructions** on data preparation and submission in publication and on website
- **template files** on website

Please provide this information in as much detail as possible. It can be helpful for evaluating your STR genotype data.
Please name this file "Author_country_number of samples_Info".xls

accompanying information file for STR dataset	Parson_AUT_527.txt
submitting lab	Institute of Legal Medicine,Innsbruck, Austria
contact person name	Walther Parson
contact person e-mail	walther.parson@i-med.ac.at
lab accreditation status if any	ISO 17025
intention for manuscript publication	yes or no
manuscript running title	STR population data for Austria
intended journal	FSI Genetics
informed consent/ethics approval/data generation according to national laws	not applicable / confirmed
type of sample set	tribe / admixed urban population
exclusion/inclusion criteria	including all residents
unrelatedness	four generations, assessed by interviews
geographic origin country/region/city	Austria/Tyrol (province)
metapopulation	unspecified
subpopulations	none
total number of individuals	527
published data from overlapping samples	yes, sample 1 published in XYZ
concordance if applicable	yes
type of specimen	buccal swabs
DNA extraction method/direct amp.	Chelex-100
CE length based alleleles/allele sequencing	CE length based
STR typing kit version/ homemade	insert name of commercial kit
allelic ladder	ladder included in kit or other
detection platform	ABI3100
detection chemistry (polymer)	POP5, POP6
data analysis software	insert name of software
data analysis software settings	standard settings
peak detection thresholds	insert rfu thresholds
positive control(s) - pass/fail/none	6 of 6 passed
raw data available	yes (if not: please contact STRidER before submission)
data transfer mode (manual/automated)	manual
suspected null alleles	none
observed discordances from different chemistries	n.a.
additional comments concerning the dataset	none




```
# 527 unshuffled genotypes from Austria, Tyrol province, at 22 autosomal STR loci using kit XYZ
# submitted by Walther Parson, Institute of Legal Medicine, Medical University of Innsbruck, walther.parson@i-med.ac.at
# admixed urban population, random unrelated sample
# this file should be named "Author_country_number of samples.txt" (e.g., "Parson_AUT_527.txt")
# further lines marked with "#" can be included for comments or description of the detailed geographic background and the appropriate m
Sample ID       AMEL    AMEL    D3S1358 D3S1358 D1S1656 D1S1656 D2S441  D2S441  D10S1248        D10S1248        D13S317 D13S317 PENTA_E
sample1 X       Y       15      18      12      17.3    11.3    14      14      17      12      13      7       23      10      12
sample2 …       …       …       …       …       …       …       …       …       …       …       …       …       …       …       …
sample3 …       …       …       …       …       …       …       …       …       …       …       …       …       …       …       …
sample4 …       …       …       …       …       …       …       …       …       …       …       …       …       …       …       …
…       …       …       …       …       …       …       …       …       …       …       …       …       …       …       …       …
```

QC of STR data on STRidER

Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER)

Martin Bodner[a], Ingo Bastisch[b], John M. Butler[c], Rolf Fimmers[d], Peter Gill[e,f], Leonor Gusmão[g,h,i], Niels Morling[j], Christopher Phillips[k], Mechthild Prinz[l], Peter M. Schneider[m], Walther Parson[a,n,*]

- assessment of **non-DNA** information
- assessment of reported **genotypes** and allele frequencies in dataset
- EPGs / **raw data** might be requested for QC: **true variants or errors**?

- **optimized procedure** for the detection of common data idiosyncrasies
- not an independent evaluation of all raw data

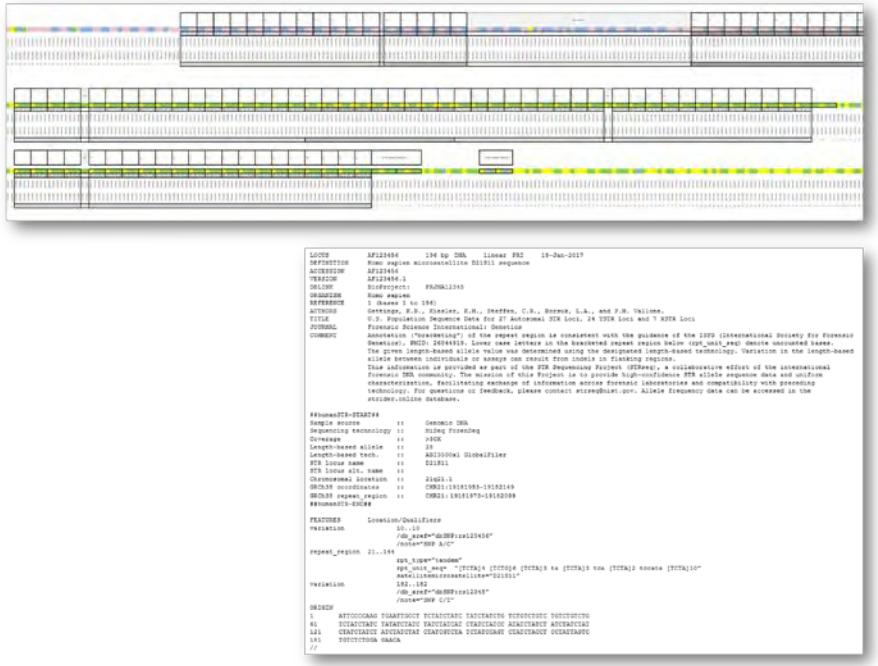
- **communication** during entire QC: discussion of all findings

QC of STR sequence data on STRidER

- submission of FASTA-like strings per locus per individual
- alignment and comparison to references:

  Forensic STR Sequence Structure Guide

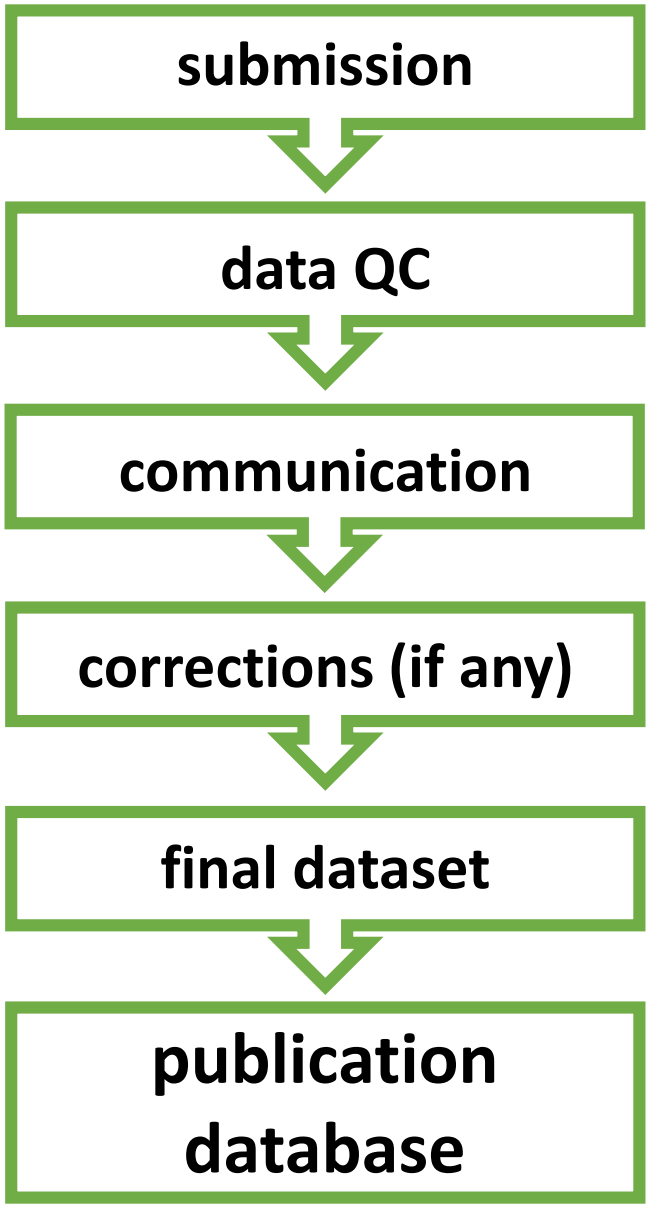
  STRSeq catalogue



- translation into CE allele
- inspection of flanking region <u>where available</u>

- **no STR sequence nomenclature** required, assessed or returned in reports
- database will host STR sequence data in future

Flowchart of QC on STRidER



Report on QC of STR data on STRidER

Information on **QC process**

Detailed **QC report**

- submitted data
- **corrections** made to dataset

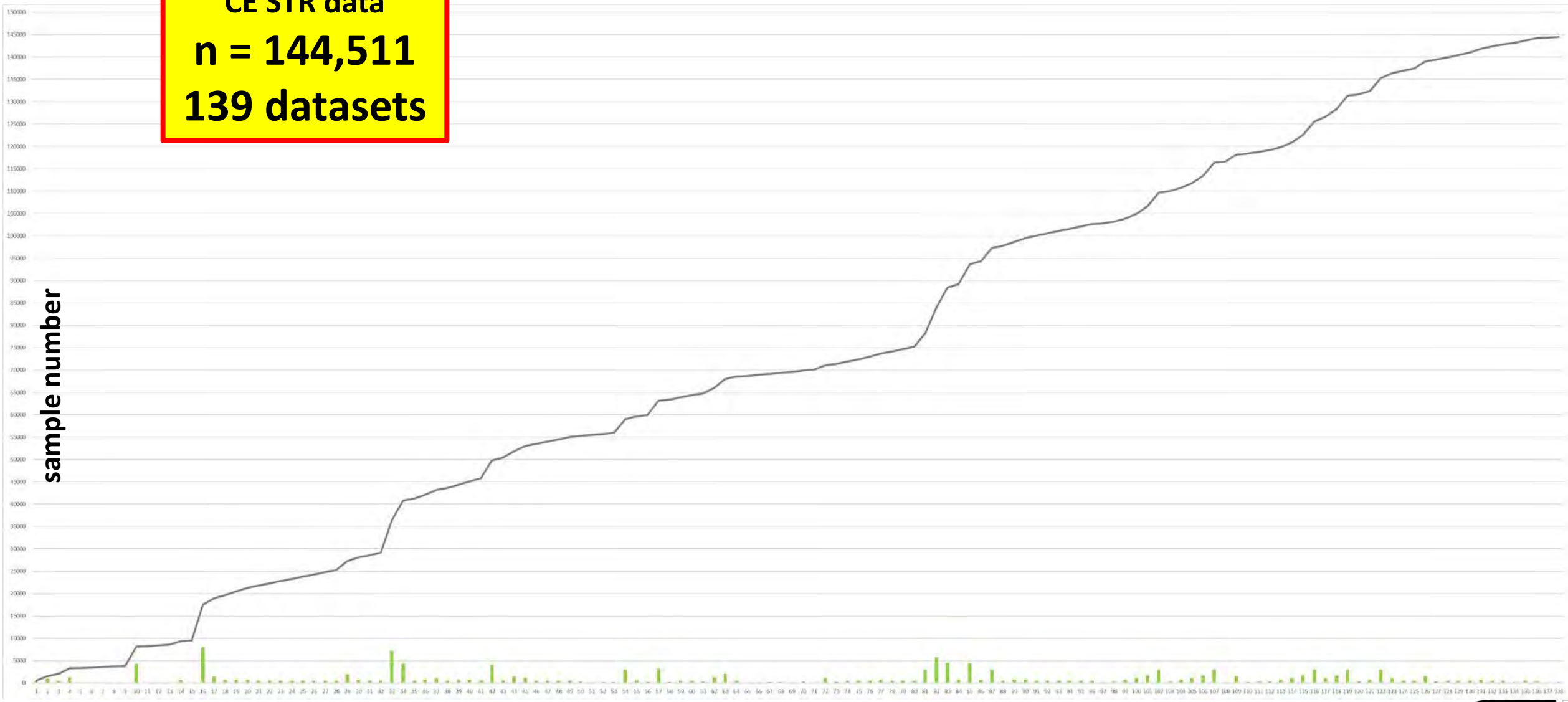
FIGURE REMOVED

STRidER accession number (STR.....) for publication

(corrected) **dataset +** allele frequency table

Submissions to STRidER Aug 2017 – March 2019

CE STR data
n = 144,511
139 datasets



Online submission tool for QC

- **form** to enter required non-genetic information (drop-down/free text)
- upload of **genotype table** in required format
- **initial plausibility tests** performed during submission process
- submission completed: e-mail notification to submitter and STRidER
- **external testing** ongoing dna.bases

back to more basic checks (completeness etc.)

Related activities

STRSeq

Seq for STRs

DNA SeqEx
STR Sequencing & Exchange

STRAND working group
align | name | define

dna.bases

STRidER

GMI DNA

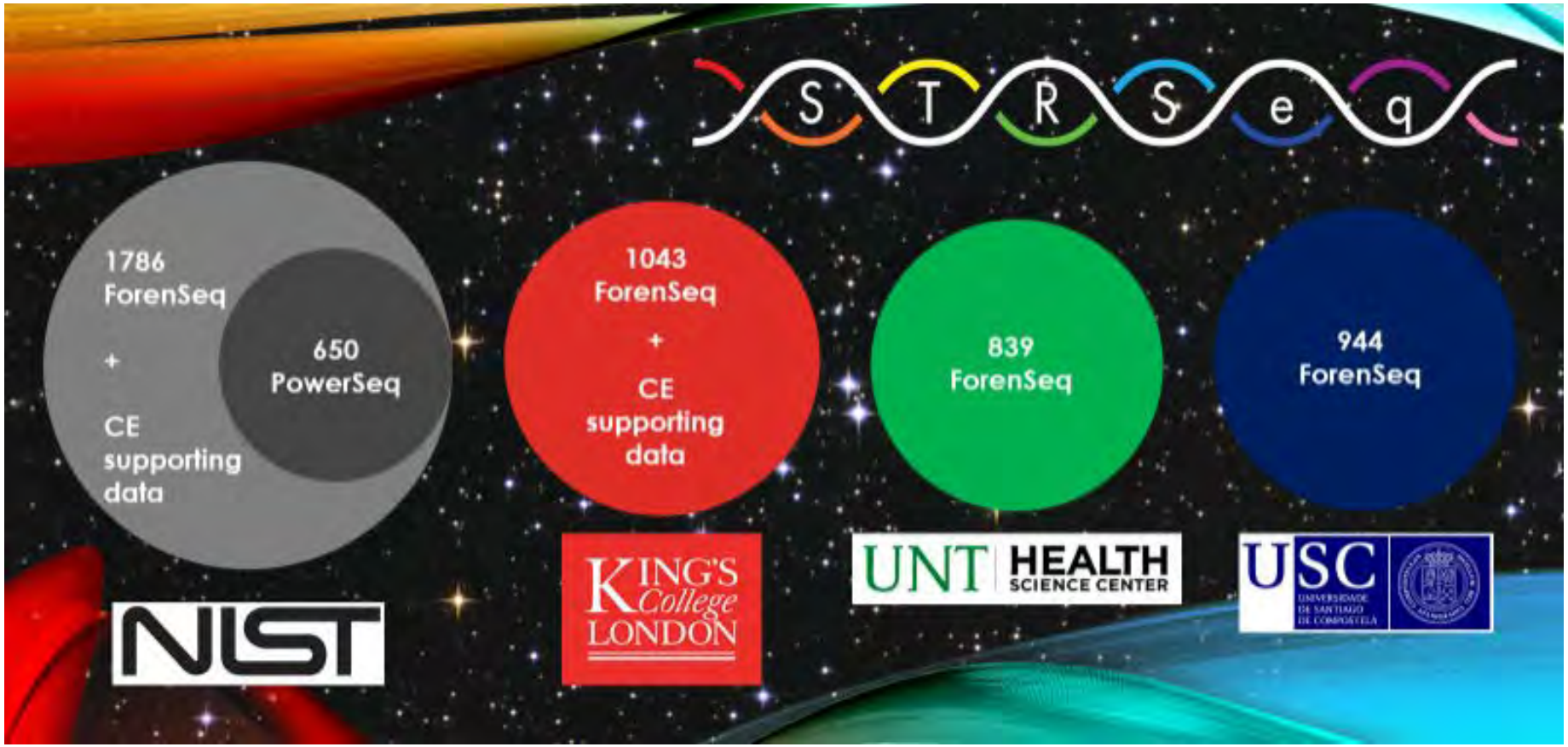
(1) STRSeq NCBI BioProject

Mission: To provide **high-confidence STR allele sequence records** with uniform annotation, facilitating exchange of information across forensic laboratories.

- collaborators with large datasets “seed” the BioProject
- NIST evaluates raw sequence data with agnostic bioinformatic pipeline
- GenBank record for **all unique sequences**
- BioProject searchable by string (BLAST), locus, allele…

STRAND working group

align | name | define



www.ncbi.nlm.nih.gov/bioproject/PRJNA380127

Forensic Science International: Genetics 31 (2017) 111–117

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci

Katherine Butler Gettings[a,*], Lisa A. Borsuk[a], David Ballard[b], Martin Bodner[c], Bruce Budowle[d,e], Laurence Devesse[b], Jonathan King[d], Walther Parson[c,f], Christopher Phillips[g], Peter M. Vallone[a]

NCBI BioProject-STRseq and **STRidER**

Collaboration in QC and exchange of data

(2) DNASEQEX

DNA-STR Massive Sequencing & International Information Exchange
(HOME/2014/ISFP/AG/LAWX/4000007135)

DNASeqEx - DNA-STR Massive Sequencing & International Information Exchange

2 years (2016-2018)

Project

DNASEQEX

Lutz Roewer · Walther Parson · Antonio Alonso · Sascha Willuweit · Bruce Budowle

Institutions: Charité Universitätsmedizin Berlin, Medizinische Universität Innsbruck, Instituto Nacional de Toxicología y Ciencias Forenses

Goal: DNASEQEX is an EU-funded project called "DNA-STR Massive Sequencing & International Information Exchange"; includes the validation of a global 23 STR & 27 Y-STR profiling system by Massively Parallel Sequencing (MPS); testing of 50 marker…

Objectives

- Promote the implementation of MPS technology for improved STR profiling and international data exchange
  → Inter-laboratory evaluation studies

- Evaluate the impact of STR sequencing on National DNA databases (EU Prüm)
  → Alonso et al. 2017 ***FSIG***

- Facilitate and standardize forensic STR sequence allele nomenclature
  → **NOMAUT** - lead Berlin

< Previous Article | July 2017 Volume 29, Pages e23–e25 | Next Article >

European survey on forensic applications of massively parallel sequencing

Antonio Alonso
National Institute of Toxicology and Forensic Sciences, Madrid Department, Spain
Petra Müller
Institute of Legal Medicine, Medical University of Innsbruck, Austria
Lutz Roewer, Sascha Willuweit
Institute of Legal Medicine and Forensic Sciences, Charité–Universitätsmedizin Berlin, Germany
Bruce Budowle
Walther Parson

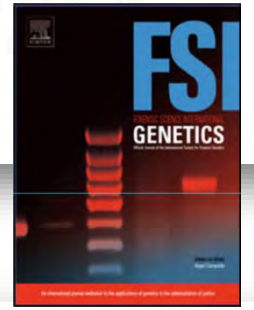
DOI: https://doi.org/10.1016/j.fsigen.2017.04.017

Research paper

Inter-laboratory validation study of the ForenSeq™ DNA Signature Prep Kit

Steffi Köcher[a,*,1], Petra Müller[b,1], Burkhard Berger[b], Martin Bodner[b], Walther Parson[b,c], Lutz Roewer[a], Sascha Willuweit[a], The DNASeqEx Consortium

[a] *Institute of Legal Medicine and Forensic Sciences, Charité – Universitätsmedizin Berlin, Germany*
[b] *Institute of Legal Medicine, Medical University of Innsbruck, Austria*
[c] *Forensic Science Program, The Pennsylvania State University, PA, USA*



Research paper

Systematic evaluation of the early access applied biosystems precision ID Globalfiler mixture ID and Globalfiler NGS STR panels for the ion S5 system

Petra Müller[a], Antonio Alonso[b], Pedro A. Barrio[b], Burkhard Berger[a], Martin Bodner[a], Pablo Martin[b], Walther Parson[a,c,*], The DNASEQEX Consortium

[a] *Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria*
[b] *National Institute of Toxicology and Forensic Sciences, Madrid Department, Las Rozas de Madrid, Spain*
[c] *Forensic Science Program, The Pennsylvania State University, PA, USA*

ELECTROPHORESIS

Review

Current state-of-art of STR sequencing in forensic genetics

Antonio Alonso, Pedro Alberto Barrio, Petra Müller, Steffi Köcher, Burkhard Berger, Pablo Martin, Martin Bodner, Sascha Willuweit, Walther Parson, Lutz Roewer, Bruce Budowle

DNA SeqEx
STR Sequencing & Exchange

(3) dna.bases

MONOPOLY 2016 - STEFA - WP G7

Empowering forensic genetic DNA databases for the interpretation of next generation sequencing profiles (**dna.bases**)

STRidER & EmPOP
Jan 2018 - Dec 2019

Sequence alignments
Increase sample size
Increase markers/regions
Further develop QC tools
User-friendly access

STRidER dna.bases EMPOP

Objectives of the work package

1) Develop a **new database engine concept** that enables event-based query of unaligned nucleotide sequence strings. This is relevant for the determination of DNA profiles with multiple sequence differences that are caused by single mutational events.

Collaborate to use and further develop existing STR sequence alignment tools

Objectives of the work package

1) Develop a **new database engine concept** that enables event-based query of unaligned nucleotide sequence strings. This is relevant for the determination of DNA profiles with multiple sequence differences that are caused by single mutational events.

2) Extension of the number of **STR markers** and quality-controlled DNA profiles from global populations, including full mtGenome data and **STR sequences** generated by NGS.

Objectives of the work package

1) Develop a **new database engine concept** that enables event-based query of unaligned nucleotide sequence strings. This is relevant for the determination of DNA profiles with multiple sequence differences that are caused by single mutational events.

2) Extension of the number of STR markers and quality-controlled DNA profiles from global populations, including full mtGenome data and STR sequences generated by NGS.

3) Update of existing and development of new tools for **quality control** of relevant genetic data. **Development and deployment of QC tools in database.**

Objectives of the work package

1) Develop a **new database engine concept** that enables event-based query of unaligned nucleotide sequence strings. This is relevant for the determination of DNA profiles with multiple sequence differences that are caused by single mutational events.

2) Extension of the number of STR markers and quality-controlled DNA profiles from global populations, including full mtGenome data and STR sequences generated by NGS.

3) Update of existing and development of new tools for **quality control** of relevant genetic data.

4) Enable **user-friendly access** to the databases from diverse platforms including mobile devices and establish links to existing software for data interpretation of DNA evidence including Forensim and LRmix studio.

(4) SeqForSTRs

Sequencing of **For**ensic **STRs**

3 years (2017-2020)

Tasks

Population study STRs as data basis to evaluate MPS - STRidER
Concordance between CE and MPS
Validation Study
Cost - benefit study
Feedback to European Laboratories (ENFSI)

Consortium

Federal crime lab **Wiesbaden** (Consortial leader)

State crime labs: Berlin, Bavaria, Rheinland-Pfalz

Legal Medicine labs: Berlin, Cologne, Innsbruck, Ulm

Acknowledgements

Richard Scheithauer, Daniela Niederwieser,
Martin Pircher, Stefan Troger

Monopoly 2010

Monopoly 2014

vXWeb

ISFG Commission on MPS of STRs
ISFG Commission on STRidER

ENFSI laboratories

Peter Gill

Co-funded by the Internal Security Fund of the European Union

EU 779485 — STEFA — ISFP-2016-AG-IBA-ENFSI

Thank you to all attendees!

STRAND *working group*

align | name | define

David Ballard
Martin Bodner
Lisa Borsuk
Katherine Gettings
Jonathan King
Walther Parson
Christopher Phillips