AUTOSOMAL STR GENOMICS 101:
Sequence Variation and Nomenclature

1

AGENDA

Core Sets and Technology

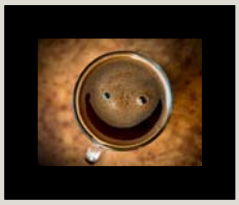
Genomic Characterization of STRs

EXERCISE 1

Historical Nomenclature

STRBase and STRSeq

EXERCISE 2



New Ideas in Nomenclature

SID, STRNaming

EXERCISE 3

STRAND and Wrap Up

2

STR CORE SETS

3

Early Development of PCR-STRs for Identification

TH01	Edwards, A., et al.	1991	Am. J. Hum. Genet. 49:746-756
TH01	Polymeropoulos, M.H., et al.	1991	Nucleic Acids Res. 19:3753
ACTBP2 (SE33)	Polymeropoulos, M.H., et al.	1992	Nucleic Acids Res. 20:1432
FGA (FIBRA)	Mills, K.A., et al.	1992	Hum. Mol. Genet. 1:779
TPOX	Anker, R., et al.	1992	Hum. Mol. Genet. 1:137
VWA (vWF)	Kimpton, C.P., et al.	1992	Hum. Mol. Genet. 1:287
D21S11	Sharma, V. and Litt, M.	1992	Hum. Mol. Genet. 1:67
Amelogenin	Sullivan, K., et al.	1993	BioTechniques 15:636-641
D3S1358	Li, H., et al.	1993	Hum. Mol. Genet. 2:1327
D18S51	Staub, R.E., et al.	1993	Genomics 15:48-56
D12S391	Lareu, M.V., et al.	1996	Gene 182:151-153

4

Early Development of U.S. Core STR Sets

1989, 1991, 1995	TWGDAM issued guidelines for QA in DNA Analysis
1996	FBI Lab sponsors a meeting for interlaboratory validation of STR loci
1997	13 CODIS Core Loci agreed upon at STR Project Meeting
1998	NDIS implemented
1998	QAS for Forensic Testing Laboratories approved
1999	QAS for Convicted Offender DNA Databasing Laboratories approved
2011	QAS update
2015	Expanded 20 CODIS core selected
2020	QAS update

5

Early Development of European Core STR Sets

1991	EDNAP formalizes as a WG of ISFH
1992	EDNAP agrees on the use of STRs
1993	FSS publishes three multiplexes, 14 STR loci (inc. vWA, TH01, D21S11)
1995	FSS publishes SGM 6-plex: TH01, vWA, FGA, D8S1179, D18S51, and D21S11
1996	EDNAP interlab determines TH01 and vWA would be used
1998	Interpol establishes ESS: TH01, vWA, FGA, D21S11
1999	Interpol expands ESS: D3S1358, D8S1179, D18S51
2005	ENSFI and EDNAP discuss extension of ESS, propose 2 sets of 3X miniSTRs
2006-2008	Prüm allows comparisons across European databases, adventitious matches
2008	ENSFI meeting agreement to add: D1S1656, D2S441, D10S1248, D12S391, D22S1045

6

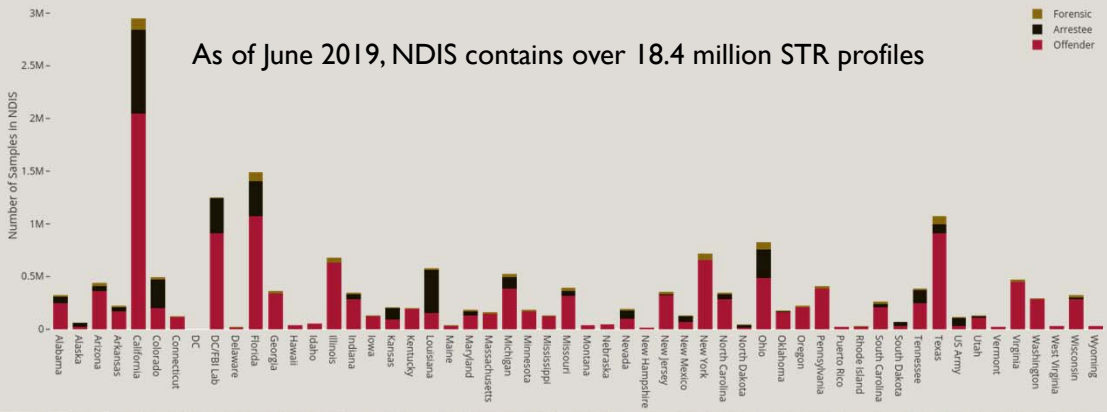
DNA Databases Worldwide

http://dnapolicyinitiative.org/

7

Advances in DNA Databases

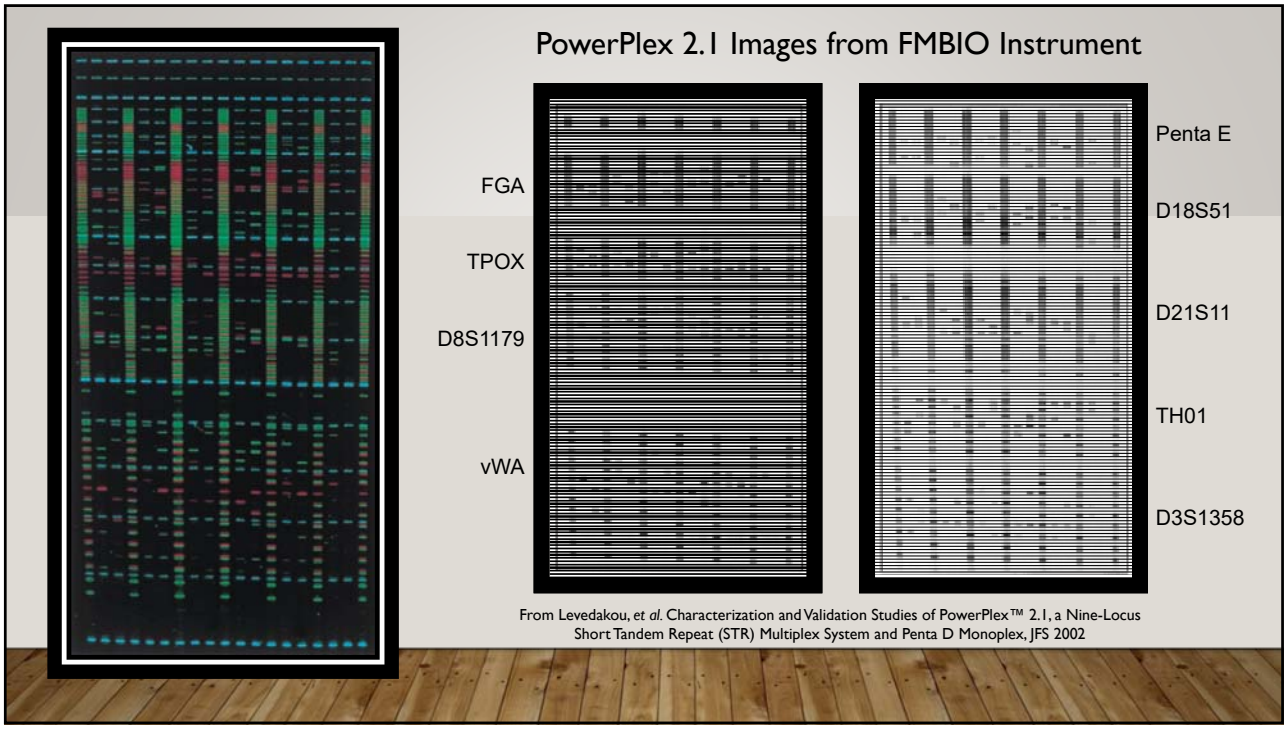
As of June 2019, NDIS contains over 18.4 million STR profiles



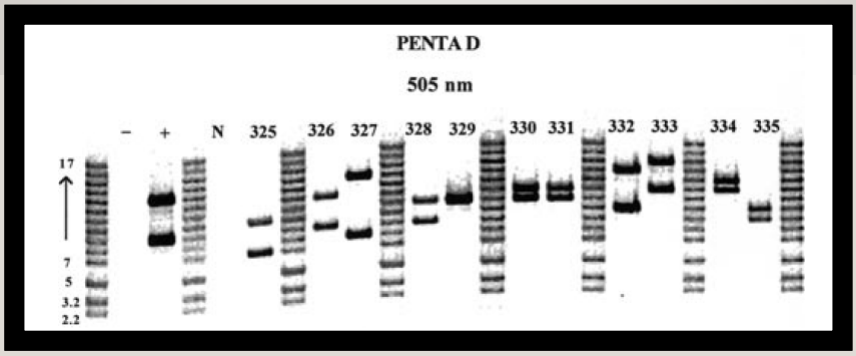
8

STR TECHNOLOGY

9



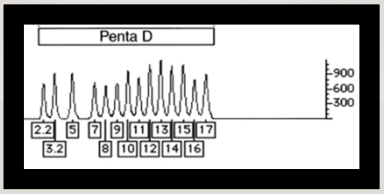
10

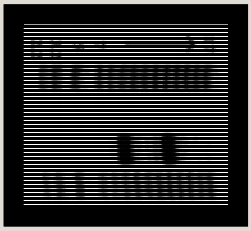


From Levedakou, *et al.* Characterization and Validation Studies of PowerPlex™ 2.1, a Nine-Locus Short Tandem Repeat (STR) Multiplex System and Penta D Monoplex, JFS 2002

11

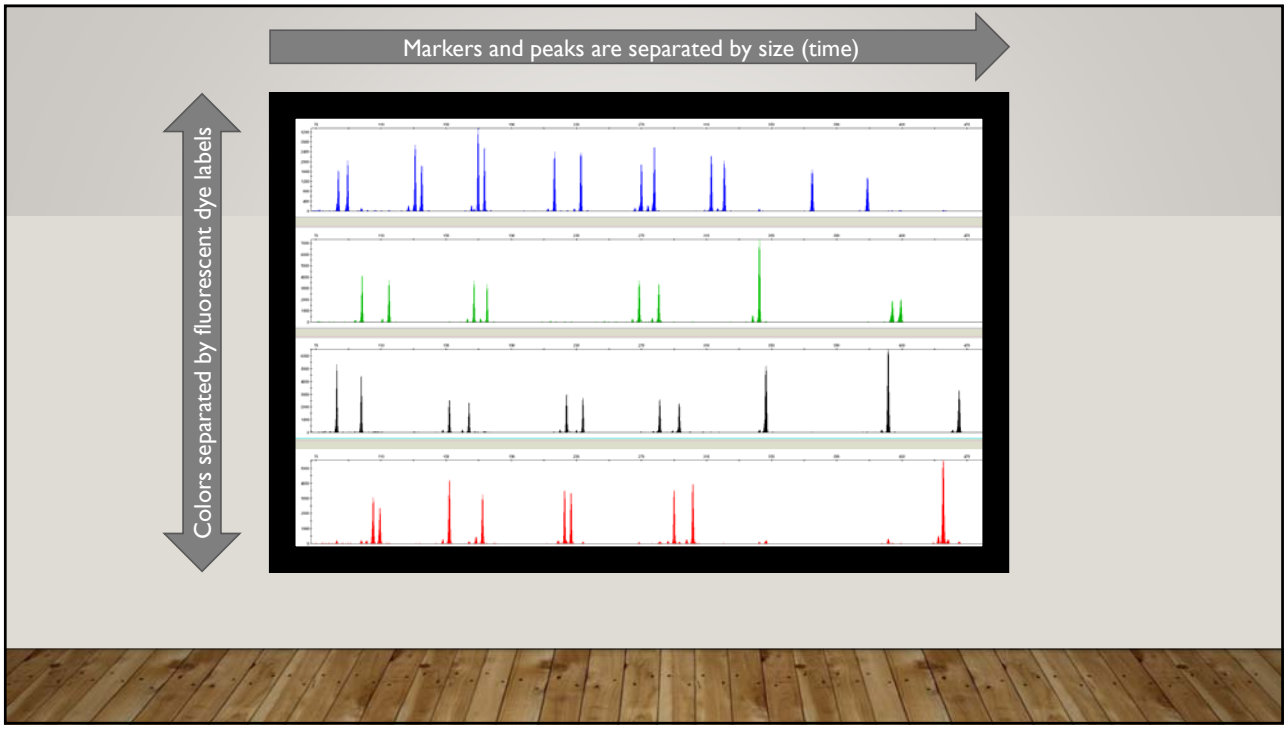
Penta D Locus by 310
(Capillary Electrophoresis)



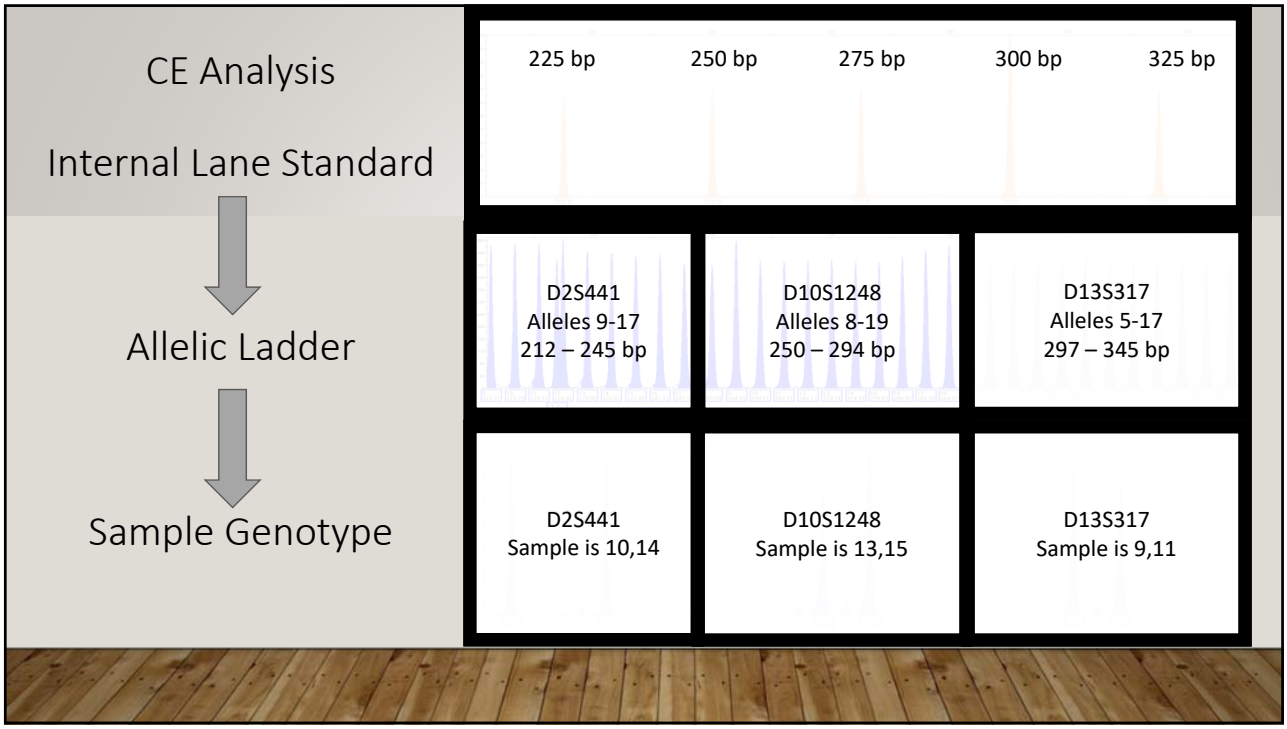


Penta D Locus by FMBIO
(Gel Electrophoresis)

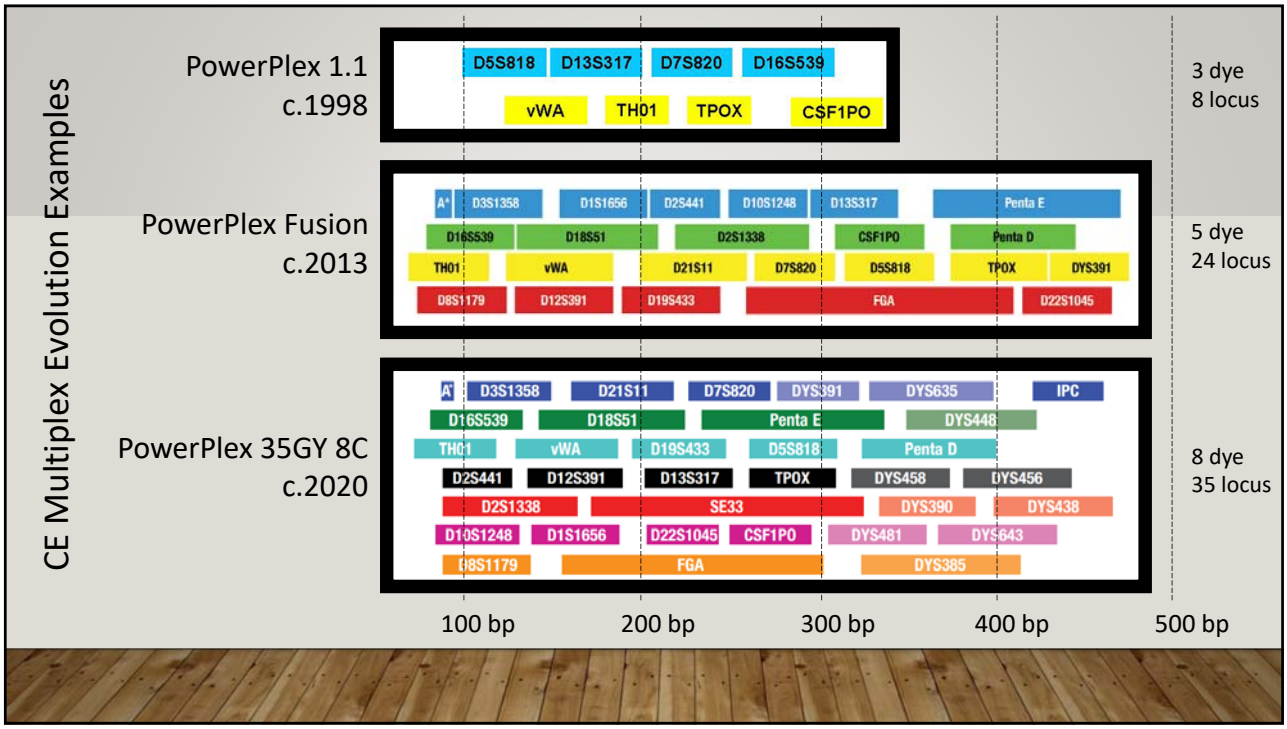
12


Markers and peaks are separated by size (time)
Colors separated by fluorescent dye labels

13

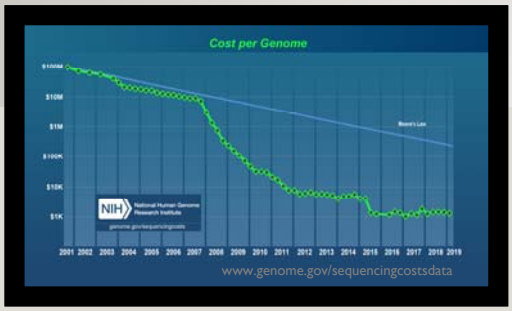

CE Analysis
Internal Lane Standard
Allelic Ladder
Sample Genotype
225 bp
250 bp
275 bp
300 bp
325 bp
D2S441
Alleles 9-17
212 – 245 bp
D10S1248
Alleles 8-19
250 – 294 bp
D13S317
Alleles 5-17
297 – 345 bp
D2S441
Sample is 10,14
D10S1248
Sample is 13,15
D13S317
Sample is 9,11

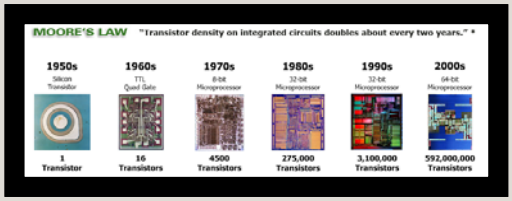
14

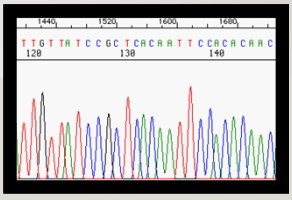


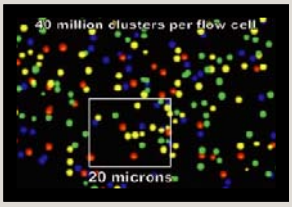
15

ADVANCES IN DNA SEQUENCING







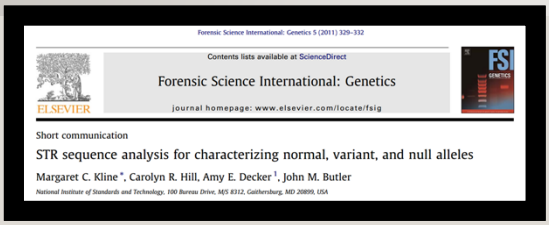


2001:
First human genome published, requiring 15 years of effort at a cost of 3 billion dollars

2014:
One instrument can sequence 45 human genomes in one day for $1000 each

16

SEQUENCING FORENSIC STRS


Forensic Science International: Genetics 5 (2011) 329–332

Contents lists available at ScienceDirect

Forensic Science International: Genetics

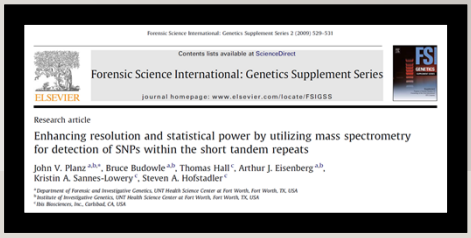
journal homepage: www.elsevier.com/locate/fsig

Short communication

STR sequence analysis for characterizing normal, variant, and null alleles

Margaret C. Kline *, Carolyn R. Hill, Amy E. Decker [1], John M. Butler

National Institute of Standards and Technology, 100 Bureau Drive, M/S 8312, Gaithersburg, MD 20899, USA


Forensic Science International: Genetics Supplement Series 2 (2009) 529–531

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

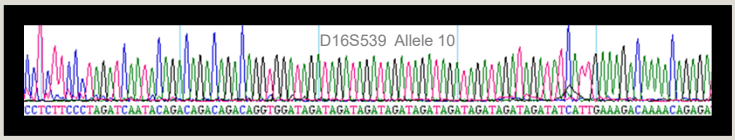
journal homepage: www.elsevier.com/locate/FSIGSS

Research article

Enhancing resolution and statistical power by utilizing mass spectrometry for detection of SNPs within the short tandem repeats

John V. Planz [a,b,*], Bruce Budowle [a,b], Thomas Hall [c], Arthur J. Eisenberg [a,b], Kristin A. Sannes-Lowery [c], Steven A. Hofstadler [c]

[a] Department of Forensic and Investigative Genetics, UNT Health Science Center at Fort Worth, Fort Worth, TX, USA
[b] Institute of Investigative Genetics, UNT Health Science Center at Fort Worth, Fort Worth, TX, USA
[c] Ibis Biosciences, Inc., Carlsbad, CA, USA



Determine the base composition of a PCR product containing STRs

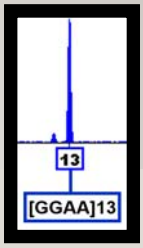
$A_{10}G_{20}C_{12}T_4 > A_{10}G_{20}C_{11}T_5$

Provides Content not Context

17

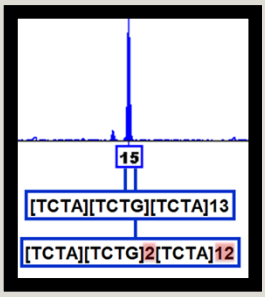
SEQUENCING FORENSIC STRS

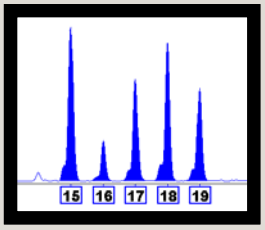
Targeted sequencing may distinguish same length alleles



Greater degree of multiplexing

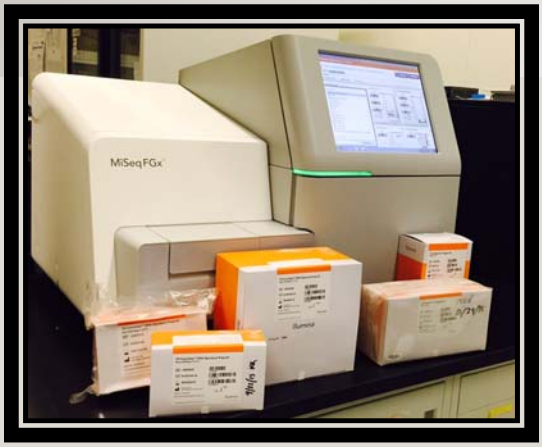
- smaller amplicons
- SNPs and other loci





18

FORENSIC NGS KITS FOR STR TYPING



Verogen ForenSeq and FGx

Applied Biosystems
Precision ID
GlobalFiler NGS
STR Panel

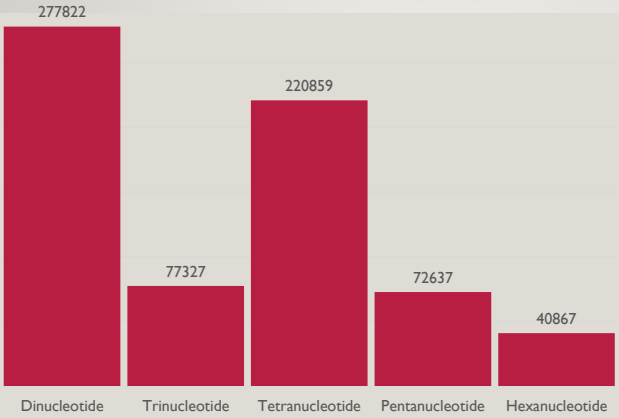
Promega
PowerSeq 46GY

19

GENOMIC CHARACTERIZATION OF STRs

20

STR Distribution Across Human Genome



Period Length Thresholds

Di-	11bp	≥ [AT]6
Tri-	14bp	≥ [AAT]5
Tetra-	14bp	≥ [AAAT]4
Penta-	16bp	≥ [AAAAT]4
Hexa-	17bp	≥ [AAAAAT]3

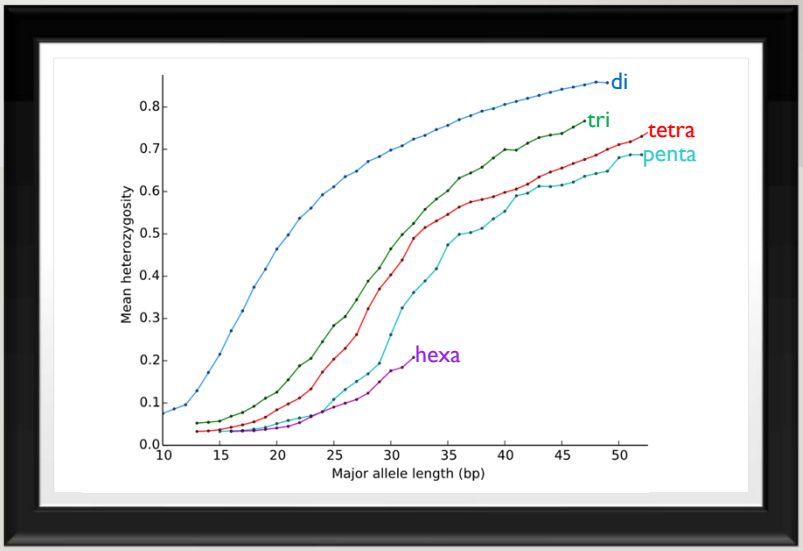
Total = 689,512 STRs

adapted from:
Willems T, Gymrek M, Highnam G; 1000 Genomes Project Consortium, Mittelman D, Erlich Y. The landscape of human STR variation. Genome Res. 2014 Nov; 24(11):1894-904.

21

Heterozygosity vs Major Allele Length by Period

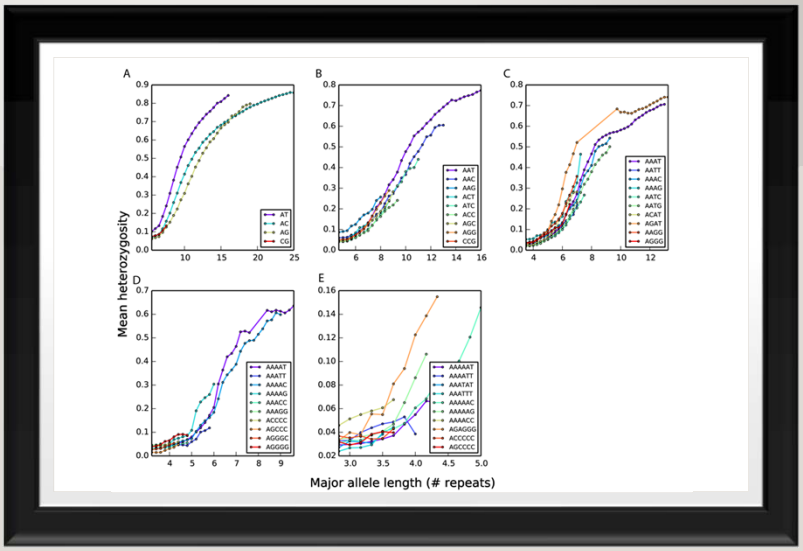
adapted from:
Willems T, Gymrek M, Highnam G; 1000 Genomes Project Consortium, Mittelman D, Erlich Y. The landscape of human STR variation. Genome Res. 2014 Nov; 24(11):1894-904.



22

Heterozygosity vs Allele Length by Period by Motif

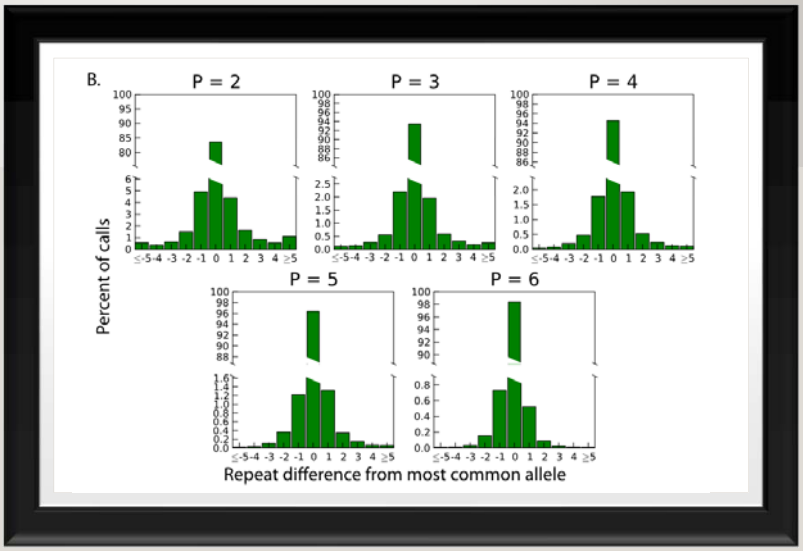
adapted from:
Willems T, Gymrek M, Highnam G; 1000 Genomes Project Consortium, Mittelman D, Erlich Y. The landscape of human STR variation. Genome Res. 2014 Nov; 24(11):1894-904.



23

Repeat Difference from Most Common Allele by Period

adapted from:
Willems T, Gymrek M, Highnam G; 1000 Genomes Project Consortium, Mittelman D, Erlich Y. The landscape of human STR variation. Genome Res. 2014 Nov; 24(11):1894-904.



24

REPEAT MOTIF CATEGORIES

SIMPLE consists of one repeat sequence

- [TCTA]n
- May have nonconsensus alleles: CCTA [TCTA]n

COMPOUND consists of two or more different repeat sequences

- [AGAT]n [AGAC]n
- May have nonconsensus alleles: [AGAT]n [AGAC]n AGAT

COMPLEX contains interspersed elements of varying period(s)

- [TCTA]n [TCTG]n [TCTA]n ta [TCTA]n tca [TCTA]2 tccata [TCTA]n

25

SIMPLE REPEATS

Consists of one repeat sequence

May have nonconsensus alleles

Locus	Allele Range	Motif	Freq	Type
D1S1656	10 to 18	CCTA [TCTA]n	54%	Simple
	10 to 17	[TCTA]n	20%	Simple
	14.3 to 19.3	CCTA [TCTA]n TCA [TCTA]n	24%	
D2S441	8 to 13	[TCTA]n	47%	Simple
	12 to 17	[TCTA]n TTTA [TCTA]2	26%	Simple
	10 to 13	[TCTA]n TCTG TCTA	18%	Simple
	11.3 to 14.3	[TCTA]n TCA [TCTA]n	5%	
D5S818	7 to 15	[ATCT]n	75%	Simple
	8 to 14	[ATCT]n	24%	Simple
	13 to 15	[ATCT]3 ATGT [ATCT]n	2%	
CSF1PO	7 to 15	[ATCT]n	99%	Simple
D6S1043	8 to 15	[ATCT]n	63%	Simple
	15 to 22	[ATCT]n ATGT [ATCT]n	34%	Simple
	23 to 26	[ATCT]n ATGT [ATCT]4 ATGT [ATCT]n	2%	
	18.3 to 23.3	[ATCT]n ATGT [ATCT]2 ATC [ATCT]n	4%	
D7S820	6 to 14	[TATC]n	100%	Simple
D8S1179	8 to 16	[TCTA]n	38%	Simple
	11 to 17	TCTA TCTG [TCTA]n	37%	Simple
	11 to 18	[TCTA]2 TCTG [TCTA]n	24%	Simple
D10S1248	8 to 19	[GGAA]n	99%	Simple
TH01	5 to 11	[AATG]n	89%	Simple
	9.3	[AATG]6 ATG [AATG]3	21%	
D13S317	8 to 15	[TATC]n	99%	Simple
Penta E	5 to 25	[TCTTT]n	99%	Simple
D16S539	8 to 15	[GATA]n	100%	Simple
D18S51	9 to 24, 28	[AGAA]n	99%	Simple
	14, 15	AGAA AGCA [AGAA]n	1%	
D19S433	9 to 17	[CCTT]n CCTA CCTT CTTT CCTT	78%	Simple
	12.2 to 18.2	[CCTT]n CCTA CCTT TT CCTT	21%	
Penta D	2, 3.2, 5 to 17	[AAAGA]n	100%	Simple
D22S1045	8 to 19	[ATT]n ACT [ATT]2	100%	Simple

26

COMPOUND REPEATS

Consists of two or more different repeat sequences

May have nonconsensus alleles

Locus	Allele Range	Motif	Freq	Type
D2S1338	15 to 24	[GGAA]n [GGCA]n	56%	Compound
	19 to 27	[GGAA]2 GGAC [GGAA]n [GGCA]n	42%	Compound
	18 to 20	[GGAA]n GAAA [GGAA]2 [GGCA]7	4%	Compound
D3S1358	11 to 20	TCTA [TCTG]2 [TCTA]n	56%	Compound
	14 to 20	TCTA [TCTG]3 [TCTA]n	28%	Compound
	12 to 18	TCTA TCTG [TCTA]n	15%	
FGA	17 to 29	[GGAA]2 GGAG [AAAG]n AGAA AAAA [GAAA]3	96%	Compound
	22 to 30	[GGAA]2 GGAG [AAAG]5 AAGG [AAAG]n AGAA AAAA [GAAA]3	4%	Compound
	16.2 to 25.2	[GGAA]2 GGAG [AAAG]n --AA AAAA [GAAA]3	2%	
vWA	11 to 21	[TAGA]n [CAGA]3-6 TAGA	90%	Compound
	14 to 15	[TGGA]0,1 [TAGA]3 TGGA [TAGA]3 [CAGA]4 TAGA CAGA TAGA	10%	Compound
D12S391	14 to 27	[AGAT]n [AGAC]n AGAT	80%	Compound
	18 to 27	[AGAT]n [AGAC]n	16%	Compound
	17.3 to 19.3	AGAT GAT [AGAT]n [AGAC]7 AGAT	3%	

27

COMPLEX REPEATS

Contain interspersed elements of varying period(s) and nonconsensus alleles

Locus	Allele Range	Motif	Freq	Type
D21S11	26 to 39	[TCTA]n [TCTG]n [TCTA]n TA [TCTA]n TCA [TCTA]2 TCCATA [TCTA]n	76%	Complex
	28.2 to 34.2	[TCTA]n [TCTG]n [TCTA]3 TA [TCTA]n TCA [TCTA]2 TCCATA [TCTA]n TA TCTA	23%	Complex
SE33	7, 11 to 23	CT [CTTT]2-3 c [CTTT]n CT [CTTT]3 CT [CTTT]2	47%	Complex
	19.2 to 33.2	CT [CTTT]2 [CCTT]1-3 C [CTTT]n TT [CTTT]n CT [CTTT]3 CT [CTTT]1-2	39%	Complex

28

CATEGORIES OF FLANKING REGIONS

1. No polymorphisms
2. Polymorphisms Associated with a Sequence Variant
3. Rare polymorphisms
4. Population or Allele Specific polymorphisms
5. "Old" polymorphisms (not population or allele specific)
6. Multiple polymorphisms in Haplotype

29

Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
No SNPs	D3S1358								
	FGA								
	D12S391								
	Penta E								
	D19S433								
	D21S11								
SNPs Associated with Repeat Region Sequence Variant	D1S1656	rs4847015	0.191	0.336	0.311	5	5	4	0
	vWA	rs11063971	0.044	0.086	0.078	1	1	1	0
		rs11063970	0.044	0.086	0.078	1	1	1	
		rs11063969	0.044	0.086	0.078	1	1	1	
		rs75219269	0.044	0.086	0.078	1	1	1	
		rs199970098	0.029	-	0.011	2	-	1	
Rare SNPs	D2S441	rs74640515	0.015	0.014	0.067	1	1	2	2
	CSF1PO	C-T 36bp 5'	-	-	0.011	-	-	1	1
	D8S1179	rs138862078	0.007	-	-	1	-	-	1
	D10S1248	T-G 2bp 3'	-	0.007	-	-	1	-	1
	D18S51	rs535833682	0.029	-	-	3	-	-	1
	Penta D	rs7279663	0.022	0.014	0.011	1	2	1	2
	D22S1045	rs190864081	0.007	-	-	1	-	-	1
Population / Allele Specific SNPs	TPOX	rs13422969	0.135	-	0.011	2	-	1	4
		rs115644759	0.022	-	-	2	-	-	
		rs149212737	-	0.007	-	-	1	-	
	TH01	rs79373318	0.148	-	-	3	-	-	3
Single "Old" SNP (not population or allele specific)	D16S539	rs1728369	0.404	0.171	0.278	4	4	4	5
		G-A 94bp 5'	-	-	0.011	-	-	1	
	D2S1338	rs6736691	0.221	0.264	0.300	8	6	4	3
Multiple SNPs in Haplotype	D5S818	G-T 4bp 3'	0.316	0.257	0.189	7	5	6	17
		rs25768	0.228	0.257	0.156	5	5	5	
		rs146841551	0.029	-	-	1	-	-	
		rs541272009	0.007	-	-	1	-	-	
	D13S317	rs9546005	0.309	0.286	0.333	5	4	7	16
		rs73525369	0.066	-	-	3	-	-	
		rs73250432	-	0.014	0.011	-	2	1	
		rs146621667	0.015	-	-	2	-	-	
		4bp del 8bp 3'	-	-	0.022	-	-	2	
		4bp del 21bp 3'	0.007	-	0.022	1	-	1	
	D7S820	rs16887642	0.177	0.050	0.022	4	2	2	15
		rs7789995	0.015	0.164	0.122	2	4	4	
		rs7786079	0.176	0.021	0.011	5	2	1	
		1bp del 21bp 3'	-	-	0.011	-	-	1	

30

Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
No SNPs	D3S1358								
	FGA								
	D12S391								
	Penta E								
	D19S433								
	D21S11								

No SNPs were identified
in the PowerSeq sequenced flanking regions of:

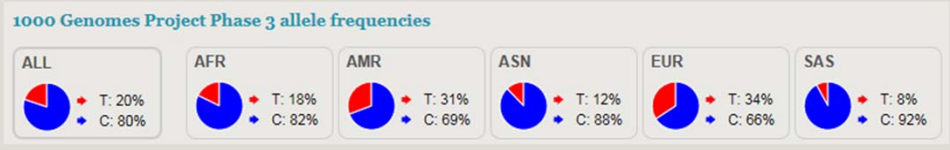
D3S1358	Penta E
FGA	D19S433
D12S391	D21S11

31

Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
SNPs Associated with Repeat Region Sequence Variant	D1S1656	rs4847015	0.191	0.336	0.311	5	5	4	0

D1S1656 has one SNP, rs4847015

- Well distributed across populations…



32

Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
SNPs Associated with Repeat Region Sequence Variant	D1S1656	rs4847015	0.191	0.336	0.311	5	5	4	0

D1S1656 has one SNP, rs4847015

- Well distributed across populations
- Well distributed across STR alleles
- Does NOT increase the number of alleles…

33

Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
SNPs Associated with Repeat Region Sequence Variant	D1S1656	rs4847015	0.191	0.336	0.311	5	5	4	0

D1S1656 has one SNP, rs4847015
and it is always associated with x.3 alleles:

15.3
16.3
17.3
18.3
19.3

→ rs4847015 = T

34

Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
SNPs Associated with Repeat Region Sequence Variant	vWA	rs11063971	0.044	0.086	0.078	1	1	1	0
		rs11063970	0.044	0.086	0.078	1	1	1	
		rs11063969	0.044	0.086	0.078	1	1	1	
		rs75219269	0.044	0.086	0.078	1	1	1	
		rs199970098	0.029	-	0.011	2	-	1	

vWA has five SNPs

- Four are coinherited
- All four associated with a sequence variant
- Fifth SNP also associated with a sequence variant

35

Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
Rare SNPs	D2S441	rs74640515	0.015	0.014	0.067	1	1	2	2
	CSF1PO	C-T 36bp 5'	-	-	0.011	-	-	1	1
	D8S1179	rs138862078	0.007	-	-	1	-	-	1
	D10S1248	T-G 2bp 3'	-	0.007	-	-	1	-	1
	D18S51	rs535833682	0.029	-	-	3	-	-	1
	Penta D	rs7279663	0.022	0.014	0.011	1	2	1	2
	D22S1045	rs190864081	0.007	-	-	1	-	-	1

Rare SNPs were identified
in the **PowerSeq** sequenced flanking regions of:

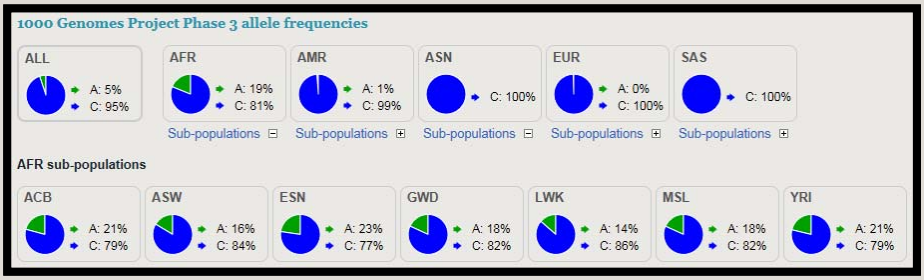
D2S441
CSF1PO
D8S1179
D10S1248
D18S51
Penta D
D22S1045

36

Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles		
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic
Population / Allele Specific SNPs	TPOX	rs13422969	0.135	-	0.011	2	-	1

TPOX: rs13422969

Well distributed across Africa, rare elsewhere



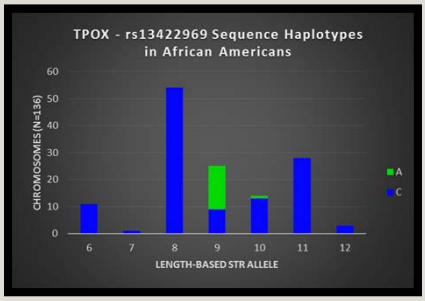
37

Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles		
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic
Population / Allele Specific SNPs	TPOX	rs13422969	0.135	-	0.011	2	-	1

TPOX: rs13422969

Well distributed across Africa, rare elsewhere

Associated with “9” allele

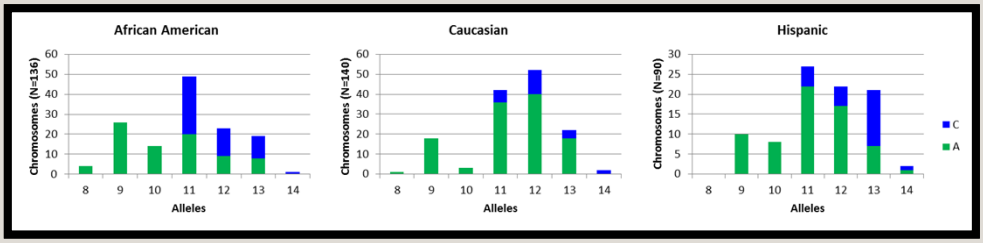


38

Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
Single "Old" SNP (not population or allele specific)	D16S539	rs1728369	0.404	0.171	0.278	4	4	4	5

D16S539: rs1728369

Well distributed across populations *and alleles*



39

Category	Locus	rs Number	Minor Allele Frequency			Number of Associated STR Alleles			Alleles Gained with Flanking SNPs
			African American	Caucasian	Hispanic	African American	Caucasian	Hispanic	
Multiple SNPs in Haplotype	D5S818	G-T 4bp 3'	0.316	0.257	0.189	7	5	6	17
		rs25768	0.228	0.257	0.156	5	5	5	
		rs146841551	0.029	-	-	1	-	-	
		rs541272009	0.007	-	-	1	-	-	

D5S818:
rs73801920
rs25768

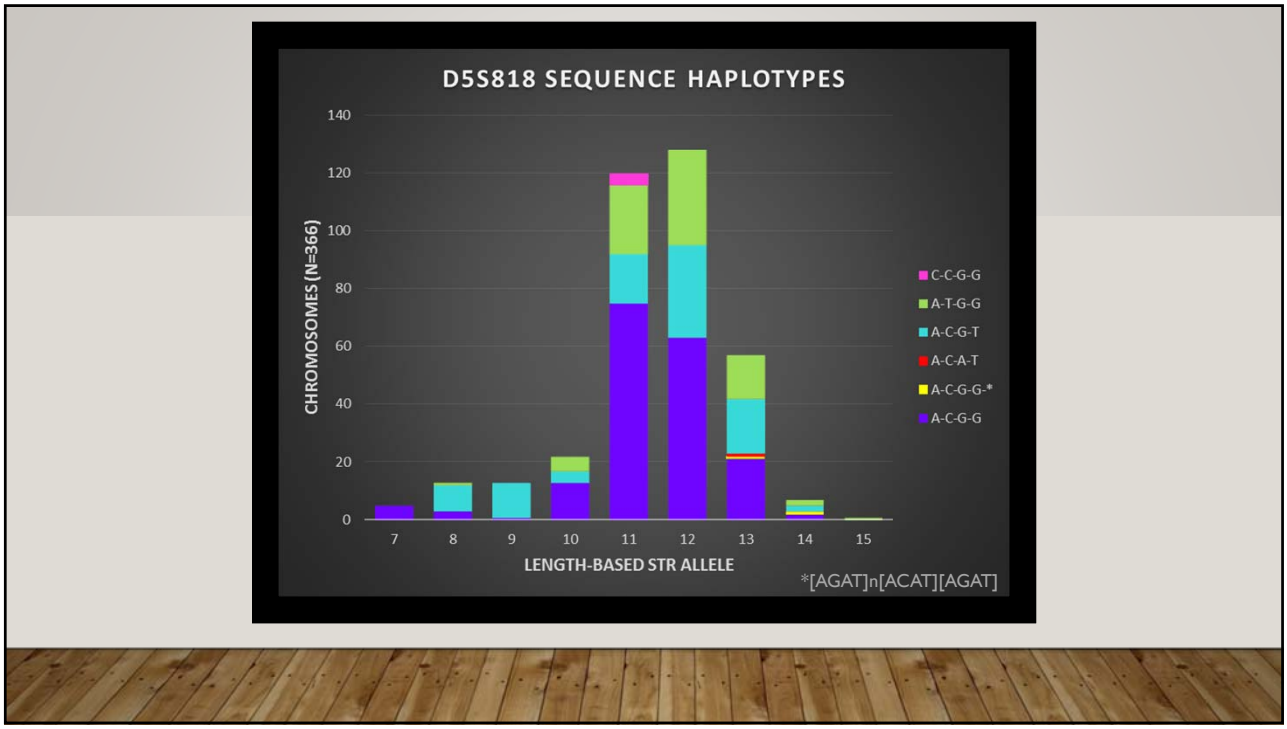
} Well distributed across populations and alleles

two additional rare SNPs

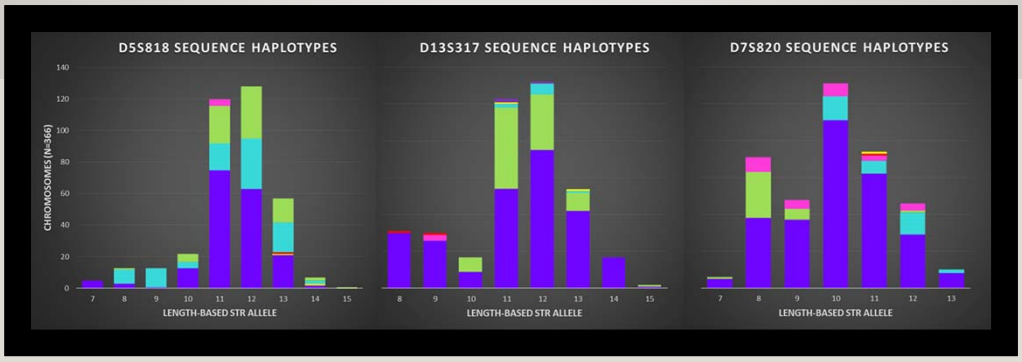
40


D5S818 STR ALLELES
CHROMOSOMES (N=366)
140
120
100
80
60
40
20
0
7
8
9
10
11
12
13
14
15
LENGTH-BASED STR ALLELE

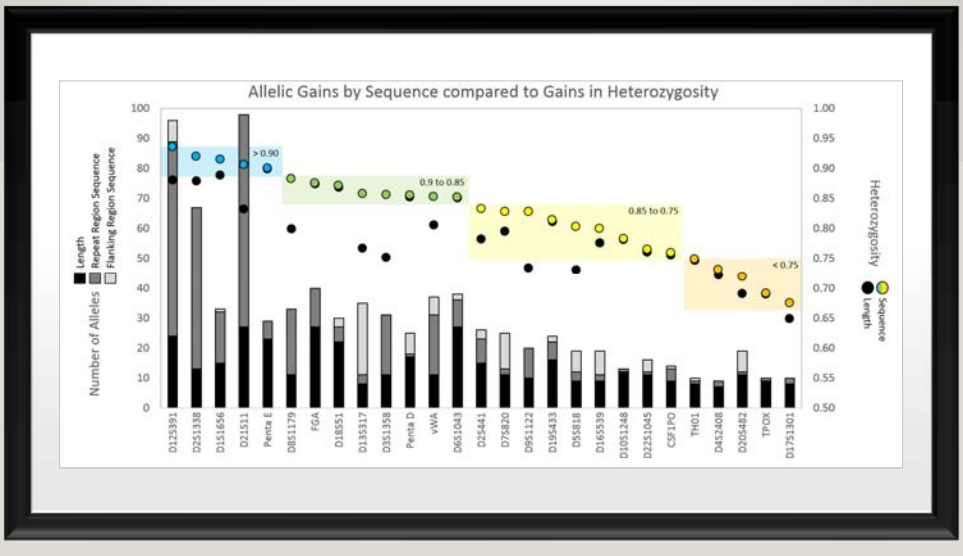
41


D5S818 SEQUENCE HAPLOTYPES
CHROMOSOMES (N=366)
140
120
100
80
60
40
20
0
7
8
9
10
11
12
13
14
15
LENGTH-BASED STR ALLELE
C-C-G-G
A-T-G-G
A-C-G-T
A-C-A-T
A-C-G-G-*
A-C-G-G
*[AGAT]n[ACAT][AGAT]

42

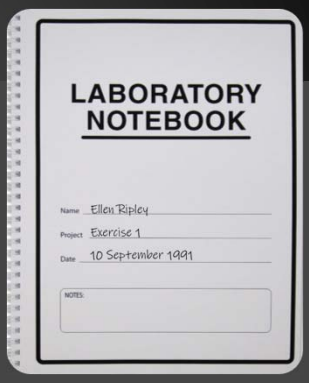

D5S818 SEQUENCE HAPLOTYPES
D13S317 SEQUENCE HAPLOTYPES
D7S820 SEQUENCE HAPLOTYPES
CHROMOSOMES (N=366)
LENGTH-BASED STR ALLELE
LENGTH-BASED STR ALLELE
LENGTH-BASED STR ALLELE

43


Allelic Gains by Sequence compared to Gains in Heterozygosity
Number of Alleles
Length
Repeat Region Sequence
Flanking Region Sequence
Heterozygosity
Sequence
Length
> 0.90
0.9 to 0.85
0.85 to 0.75
< 0.75
D12S391
D2S1338
D16S539
D21S11
Penta E
D8S1179
FGA
D18S51
D13S317
D3S1358
Penta D
vWA
D6S1043
D2S441
D7S820
D9S1122
D19S433
D5S818
D16S539
D10S1248
D22S1045
CSF1PO
TH01
D4S2408
D20S482
TPOX
D17S1301

44

EXERCISE 1



45

HISTORICAL NOMENCLATURE

46

1991
Early paper with nomenclature info

Locus Name: based on GenBank locus designations

- e.g. HUMHPRTB

Motif: lowest alphabetical representation of the STR

- e.g. GATA > AGAT
- TH01!!!

Number of repeats based on sequencing data from at least two alleles

DNA Typing and Genetic Mapping with Trimeric and Tetrameric Tandem Repeats

Al Edwards,* Andrew Civitello,* Holly A. Hammond,* and C. Thomas Caskey*†

Summary

Introduction

47

1997
DNA recommendations of the ISFH

A common nomenclature for STRs is a prerequisite for interlaboratory reproducibility and for the exchange and comparison of data.

For many loci there has been an enormous increase of information about sequence and substructure which has repeatedly raised questions relating to nomenclature.

EDITORIAL

W. Bär · B. Brinkmann · B. Budowle · A. Carracedo P. Gill · P. Lincoln · W. Mayr · B. Olaisen

DNA recommendations

Further report of the DNA Commission* of the ISFH regarding the use of short tandem repeat systems

Introduction

1. Sequence and repeat designation of STRs

48

1997
DNA recommendations of the ISFH

Strand

- protein coding genes or pseudogenes: use coding strand
- no connection to coding genes: sequence originally described in literature or 1st public DB entry

Motif (unless already established):

- First 5' nts that can define a motif

If multiple nomenclatures exist, prioritize the one which most closely adheres. If not possible, then use the first documented.

W. Bär · B. Brinkmann · B. Budowle · A. Carracedo
P. Gill · P. Lincoln · W. Mayr · B. Olaisen

DNA recommendations

Further report of the DNA Commission* of the ISFH regarding the use of short tandem repeat systems

49

1997
DNA recommendations of the ISFH

Allele Numbering for non-straightforward loci

- If sequence variation is present, use full number of repeats including variation
- Complex loci: repeat nomenclature should have a mathematical relationship to length of a consensus allele
- Example D21S11

repeats	length	relation
27	213bp	27x4+105*=213
31	229bp	31x4+105*=229
33.2	239bp	33x4+2+105*=239

*105bp = sum of 5'+3' flank and 43bp interspersed element

W. Bär · B. Brinkmann · B. Budowle · A. Carracedo
P. Gill · P. Lincoln · W. Mayr · B. Olaisen

DNA recommendations

Further report of the DNA Commission* of the ISFH regarding the use of short tandem repeat systems

"Authors should follow these recommendations and state they did so."

50

2006
Recommendations of the ISFG

Allele number derived from total number of contiguous variant and non-variant repeats

Single repeat units adjacent to main array and of the same sequence as main variable repeat should be included

- e.g. [GATA]n $[GACA]_2$ [GATA] = n+2+1

If repetitive motifs are not adjacent and have ≤3 units and show no variation, they should not be included

- e.g. [GATA]n $[GACA]_2$ $\mathbf{N_8}$ $[GATA]_3$ = n+2
  - If N ≤ 4 nt, include
  - If N >4 nt, exclude

Short communication

DNA Commission of the International Society of Forensic Genetics (ISFG): An update of the recommendations on the use of Y-STRs in forensic analysis

L. Gusmão, J.M. Butler, A. Carracedo, P. Gill, M. Kayser, W.R. Mayr, N. Morling, M. Prinz, L. Roewer, C. Tyler-Smith, P.M. Schneider

51

2006
Recommendations of the ISFG

Indels within repeat are counted

e.g. $[CTTTT]_8$ C $[CTTTT]_3$ = 11.1

Indels in flank are designated separately

e.g. 11 (U40Tins)

(this only works when sequencing)

Ambiguous indels assigned to highest numbered end of homopolymer

e.g. C-AAAAAAA$[GATA]_6$… = U9Adel

e.g. …$[TATC]_6$TTTTTTT-GC = D9Tdel

Short communication

DNA Commission of the International Society of Forensic Genetics (ISFG): An update of the recommendations on the use of Y-STRs in forensic analysis

L. Gusmão, J.M. Butler, A. Carracedo, P. Gill, M. Kayser, W.R. Mayr, N. Morling, M. Prinz, L. Roewer, C. Tyler-Smith, P.M. Schneider

52

2006
Recommendations of the ISFG

Indels within repeat are counted

- e.g. $[CTTTT]_8$ C $[CTTTT]_3$ = 11.1

Indels in flank are designated separately

- e.g. 11 (U40Tins)

(this only works when sequencing)

Ambiguous indels assigned to highest numbered end of homopolymer

- e.g. C-AAAAAAA$[GATA]_6$… = U9Adel
- e.g. …$[TATC]_6$TTTTTTT-GC = D9Tdel

DNA Commission of the International Society of Forensic Genetics (ISFG): An update of the recommendations on the use of Y-STRs in forensic analysis

53

2016
Considerations of the ISFG

Consideration 1: MPS analysis should be performed with software that allows STR sequences to be exported and stored in databases as sequence (text) strings to capture the maximum consensus sequence information.

Forensic Science International: Genetics

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

54

2016
Considerations of the ISFG

Consideration 2. The forward strand direction assigned in the human genome has been constant for all assemblies published since the first draft in 2001 and can be used to align STR sequences.

55

2016
Considerations of the ISFG

Consideration 3. The choice of reference sequence is crucial for standardizing STR nomenclature systems. At the time of writing, GRCh38 is the most up-to-date sequence assembly and is recommended as the framework with which to define repeat region structure for sequence alignment and for the mapping of sequence features such as SNPs.

56

2016
Considerations of the ISFG

Consideration 4. Further work is needed to translate the nomenclature of STR loci thus far coded relative to the reverse strand and repeat region start and end points. There is a need to strictly define these and other anchor points to specify the repeat regions.

57

2016
Considerations of the ISFG

Consideration 5. Although simple STR nomenclature systems may be required at some point in the future … comprehensive STR nomenclature systems are preferred for early adopters... Backward compatibility to the repeat-based nomenclature derived from CE needs to be maintained to preserve the universal applicability of established national STR databases.

58

2016
Considerations of the ISFG

Consideration 6. To account for relevant genetic variation outside common repeat regions, STR sequences stored as sequence strings should include flanking sequences as well as the genome coordinates of the sequence read start and end points.

59

2016
Considerations of the ISFG

Consideration 7. Updated allele frequency databases will be necessary to take full advantage of the increased power of discrimination offered by MPS generated STR data. A unified nomenclature system is needed to ensure compatibility of worldwide population databases.

60

2016
Considerations of the ISFG

Consideration 8. Future forensic MPS multiplexes would benefit from retention of past markers for backward compatibility and a marker selection process based on population data, molecular biology, sequencing chemistry, and a continued dialogue between the forensic community and commercial suppliers.

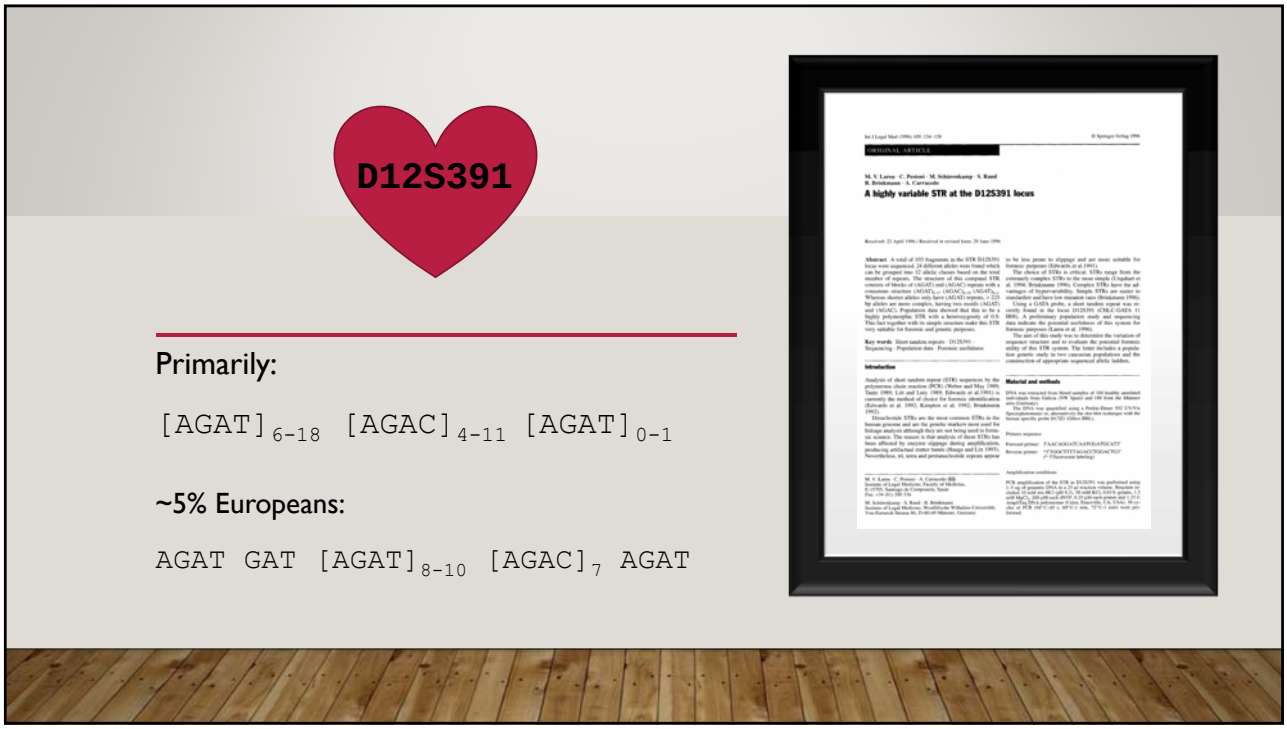

Forensic Science International: Genetics

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

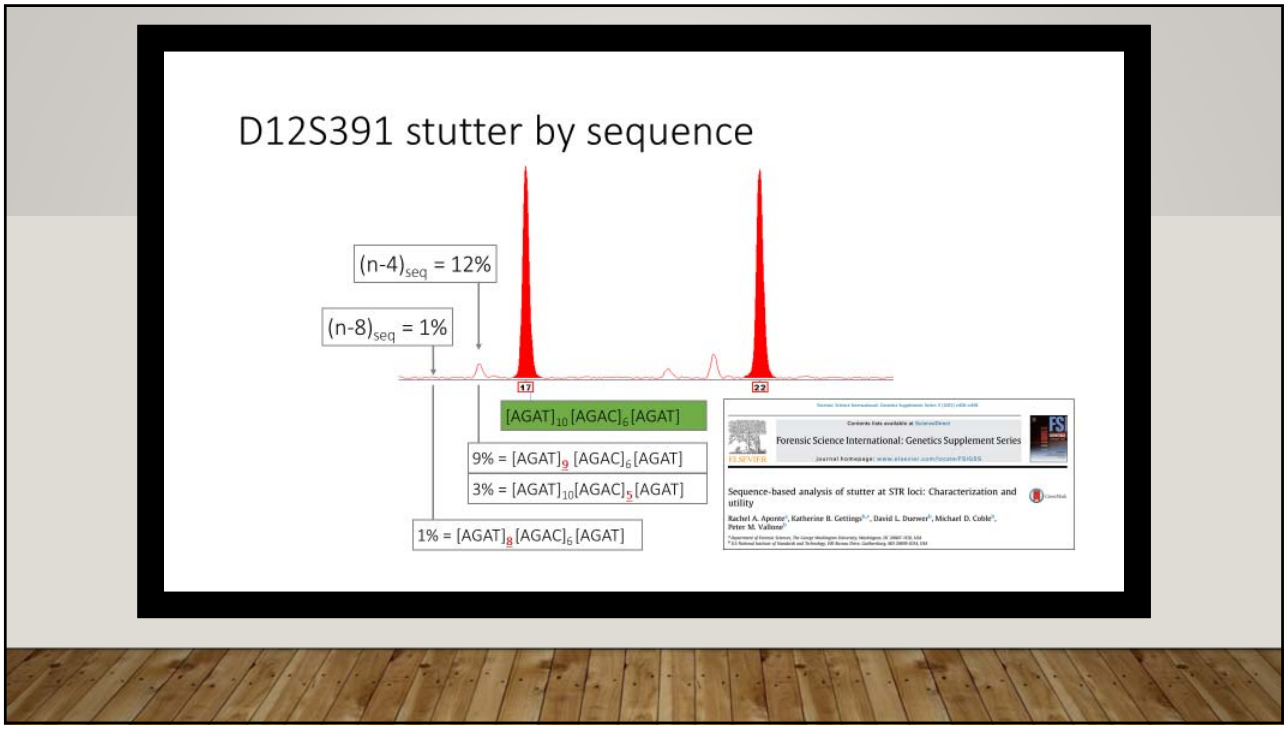
61

Fun with D12S391

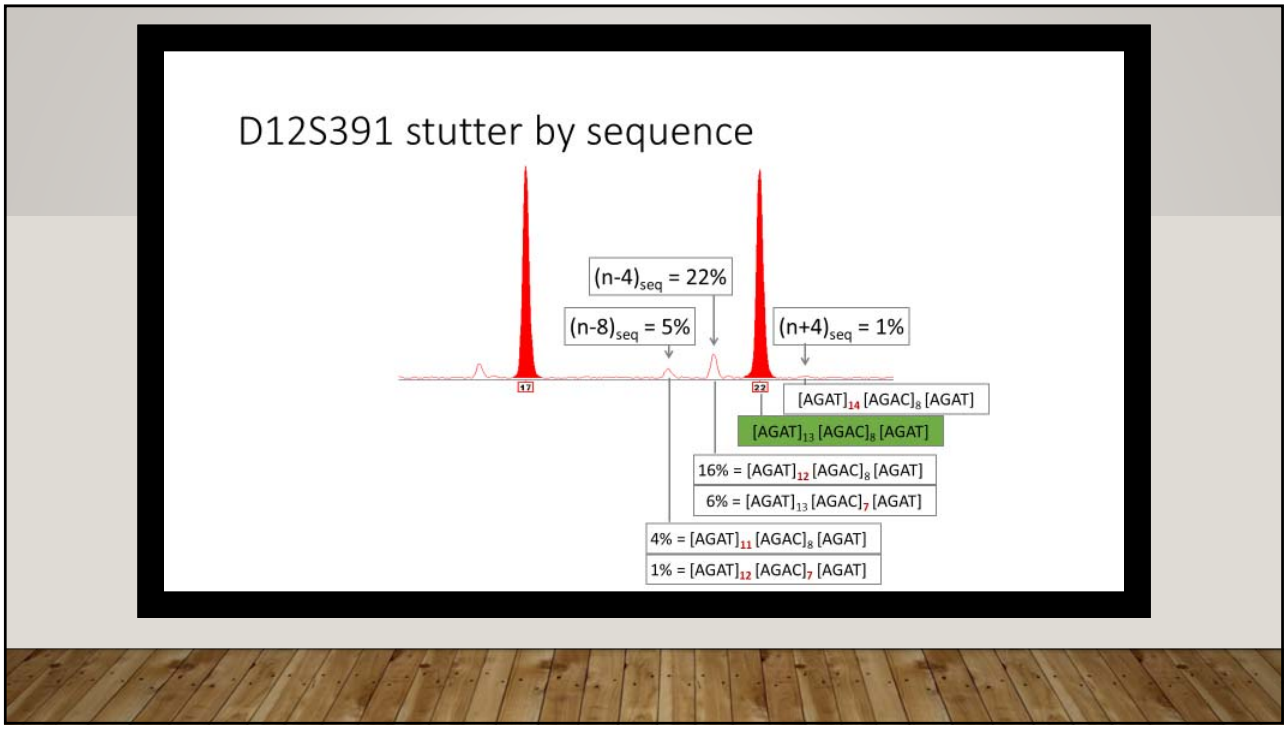
62


D12S391
Primarily:
[AGAT]6-18 [AGAC]4-11 [AGAT]0-1
~5% Europeans:
AGAT GAT [AGAT]8-10 [AGAC]7 AGAT
A highly variable STR at the D12S391 locus

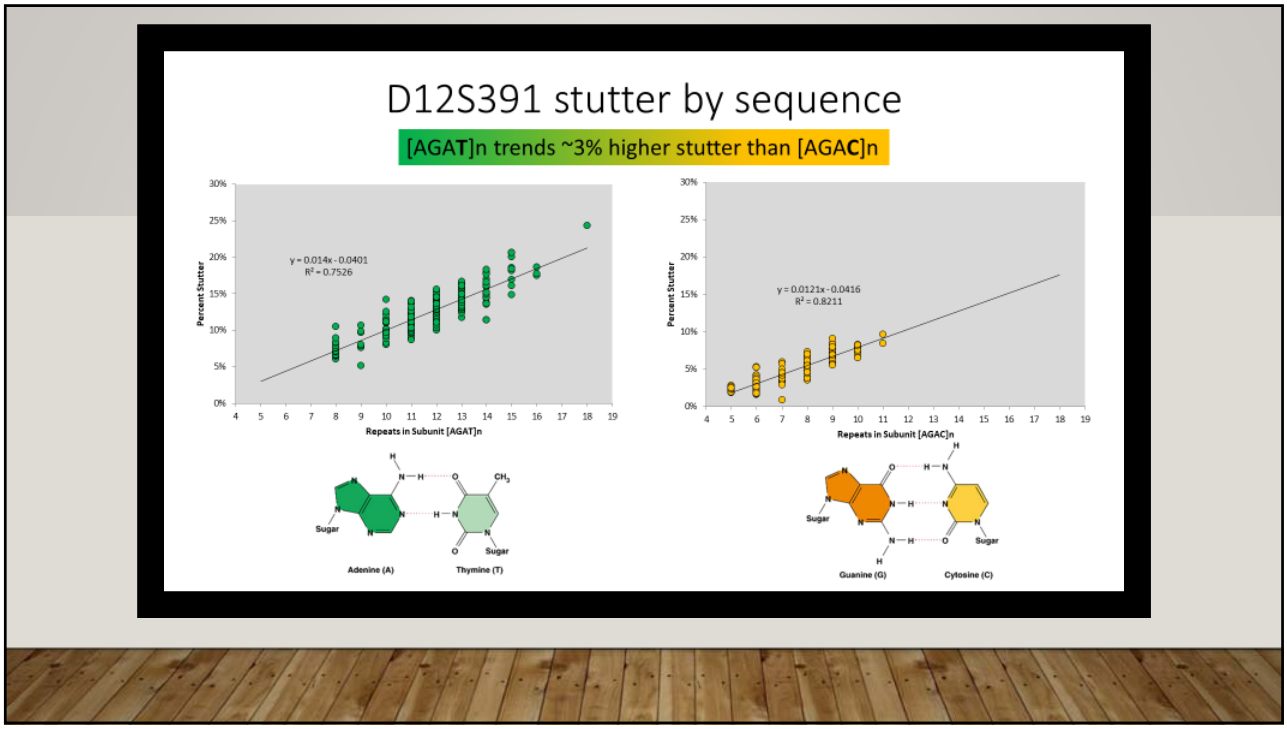
63


D12S391 stutter by sequence
(n-4)seq = 12%
(n-8)seq = 1%
17
22
[AGAT]10[AGAC]6[AGAT]
9% = [AGAT]9 [AGAC]6[AGAT]
3% = [AGAT]10[AGAC]5[AGAT]
1% = [AGAT]8[AGAC]6[AGAT]
Forensic Science International: Genetics Supplement Series
Sequence-based analysis of stutter at STR loci: Characterization and utility

64


D12S391 stutter by sequence
(n-4)seq = 22%
(n-8)seq = 5%
(n+4)seq = 1%
17
22
[AGAT]14 [AGAC]8 [AGAT]
[AGAT]13 [AGAC]8 [AGAT]
16% = [AGAT]12 [AGAC]8 [AGAT]
6% = [AGAT]13 [AGAC]7 [AGAT]
4% = [AGAT]11 [AGAC]8 [AGAT]
1% = [AGAT]12 [AGAC]7 [AGAT]

65


D12S391 stutter by sequence
[AGAT]n trends ~3% higher stutter than [AGAC]n
y = 0.014x - 0.0401
R² = 0.7526
Percent Stutter
Repeats in Subunit [AGAT]n
y = 0.0121x - 0.0416
R² = 0.8211
Percent Stutter
Repeats in Subunit [AGAC]n
Sugar
Adenine (A)
Thymine (T)
Guanine (G)
Cytosine (C)

66

Forensic Science International: Genetics

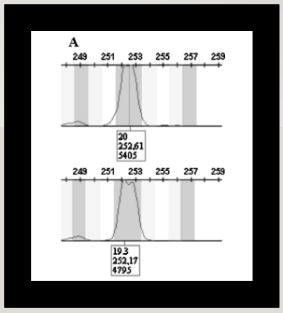
journal homepage: www.elsevier.com/locate/fsig

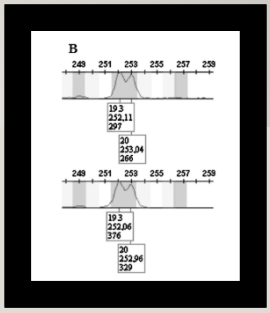
Case Report

Non-uniform phenotyping of D12S391 resolved by second generation sequencing

S. Dahgaard, E. Rockenbauer, A. Buchard, H.S. Mogensen, R. Frank-Hansen, C. Børsting *, N. Morling

Single Source samples show varying levels of 1bp resolution:





This has implications for validations and mixture analysis:

	Allele	Repeat Region Sequence	A	C	G	T	MW	delta	if F is labeled
Sample 1	19.3	AGAT GAT [AGAT]10 [AGAC]7 AGAT	39	7	20	13	10709.93		
Sample 2	20	[AGAT]13 [AGAC]6 AGAT	40	6	20	14	10860.08	150.15	easier
Sample 3	20	[AGAT]12 [AGAC]8	40	8	20	12	10830.04	120.11	harder

67

STRBase & STRSeq

68

STRBase

Short Tandem Repeat DNA Internet DataBase

69

STRBase: a short tandem repeat DNA database for the human identity testing community

Christian M. Ruitberg, Dennis J. Reeder and John M. Butler*

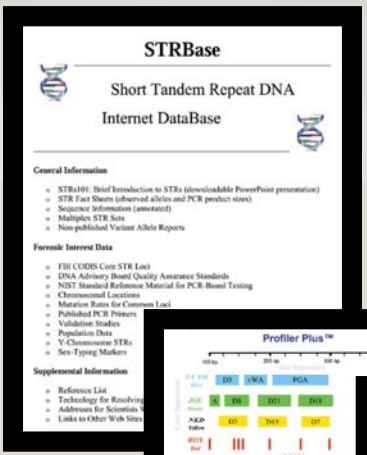
Biotechnology Division, National Institute of Standards and Technology, 100 Bureau Drive, Mail Stop 8311, Gaithersburg, MD 20899-8311, USA

ABSTRACT

The National Institute of Standards and Technology (NIST) has compiled and maintained a Short Tandem Repeat DNA Internet Database (http://www.cstl.nist.gov/biotech/strbase/) since 1997 commonly referred to as STRBase. This database is an information resource for the forensic DNA typing community with details on commonly used short tandem repeat (STR) DNA markers. STRBase consolidates and organizes the abundant literature on this subject to facilitate on-going efforts in DNA typing. Observed alleles and annotated sequence for each STR locus are described along with a review of STR analysis technologies. Additionally, commercially available STR multiplex kits are described, published polymerase chain reaction (PCR) primer sequences are reported, and validation studies conducted by a number of forensic laboratories are listed. To supplement the technical information, addresses for scientists and hyperlinks to organizations working in this area are available, along with the comprehensive reference list of over 1300 publications on STRs used for DNA typing purposes.

INTRODUCTION

CONTENT OF STRBase



STRBase

Short Tandem Repeat DNA Internet DataBase

General Information

- STRs101: Brief Introduction to STRs (downloadable PowerPoint presentation)
- STR Fact Sheets (observed alleles and PCR product sizes)
- Sequence Information (annotated)
- Multiplex STR Sets
- Non-published Variant Allele Reports

Forensic Interest Data

- FBI CODIS Core STR Loci
- DNA Advisory Board Quality Assurance Standards
- NIST Standard Reference Material for PCR-Based Testing
- Chromosomal Locations
- Mutation Rates for Common Loci
- Published PCR Primers
- Validation Studies
- Population Data
- Y-Chromosome STRs
- Sex-Typing Markers

Supplemental Information

- Reference List
- Technology for Resolving
- Addresses for Scientists
- Links to Other Web Sites

D8S1179

General Information

Other Names D6S502, D8

Chromosomal Location 8q

GenBank Accession GO8710, has 12 repeats

Repeat: [TCTA] = GenBank top strand

PCR Primer Information

Reported Primers

Ref.

PCR Primer Sequences

Set 1 369 F - TTTTTGTATTCACATGTACATTCG - 3' R - CGTAGCTATAATTATGTTCATTTTCA - 3'

Set 2 PCARE AmpFlSTR® Profiler Plus™

Set 3 Promega GenePrint® PowerPlex™ 2.1, GenePrint® PowerPlex™ 16

PCR Product Sizes of Observed Alleles

Allele (Repeat #)	Set 1	Set 2	Set 3	Repeat Structure	Ref.
7	157 bp	123 bp	200 bp	$[TCTA]_3$	T16
8	161 bp	127 bp	205 bp	$[TCTA]_4$	360
9	165 bp	131 bp	211 bp	$[TCTA]_5$	360
10	169 bp	135 bp	215 bp	$[TCTA]_6$	360
11	173 bp	139 bp	219 bp	$[TCTA]_7$	360
12	177 bp	143 bp	223 bp	$[TCTA]_8$	360
13	181 bp	147 bp	227 bp	$[TCTA]_9[TCTG]_1[TCTA]_3$	360
14	185 bp	151 bp	231 bp	$[TCTA]_{10}[TCTG]_1[TCTA]_3$	360
15	189 bp	155 bp	235 bp	$[TCTA]_{11}[TCTG]_1[TCTA]_3$	360
16	193 bp	159 bp	239 bp	$[TCTA]_{12}[TCTG]_1[TCTA]_3$	360
17	197 bp	163 bp	243 bp	$[TCTA]_{13}[TCTG]_1[TCTA]_3$	360
18	201 bp	167 bp	247 bp	$[TCTA]_{14}[TCTG]_1[TCTA]_3$	360
19	205 bp	171 bp	251 bp	$[TCTA]_{15}[TCTG]_1[TCTA]_3$	T16

Profiler Plus™

D3 vWA FGA

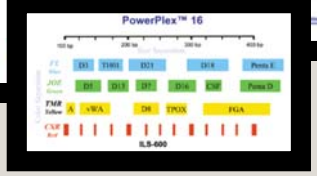
D8 D21 D18

D5 D13 D7

AGCT Yellow

ROX Red

G5500



PowerPlex™ 16

D3 TH01 D21 D18 Penta E

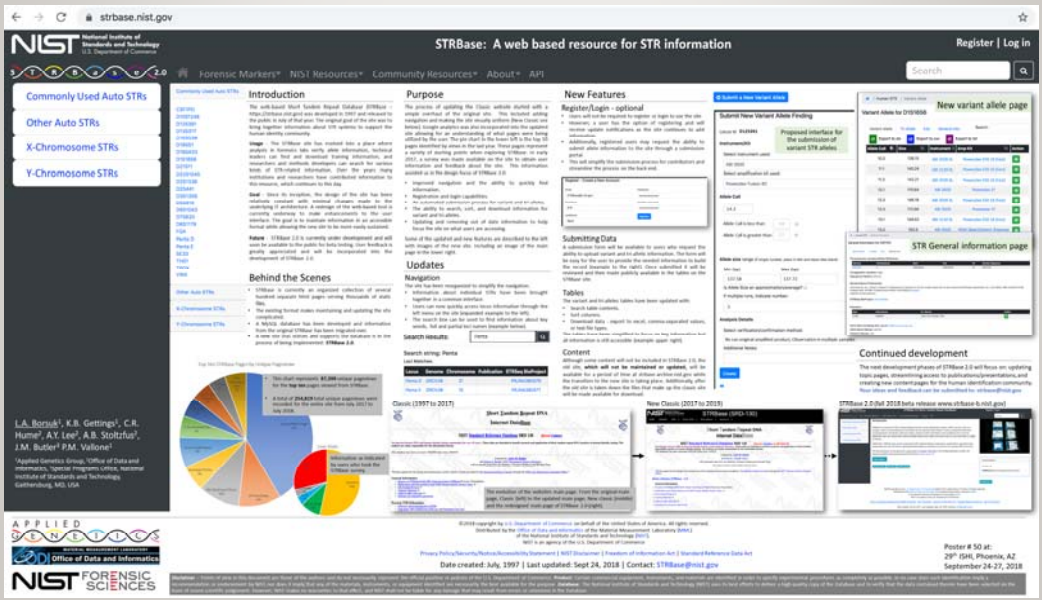
D5 D13 D7 D16 CSF Penta D

TMR Yellow A vWA D8 TPOX FGA

CXR Red

ILS-600

70



71

STRSeq

The STR Sequencing Project

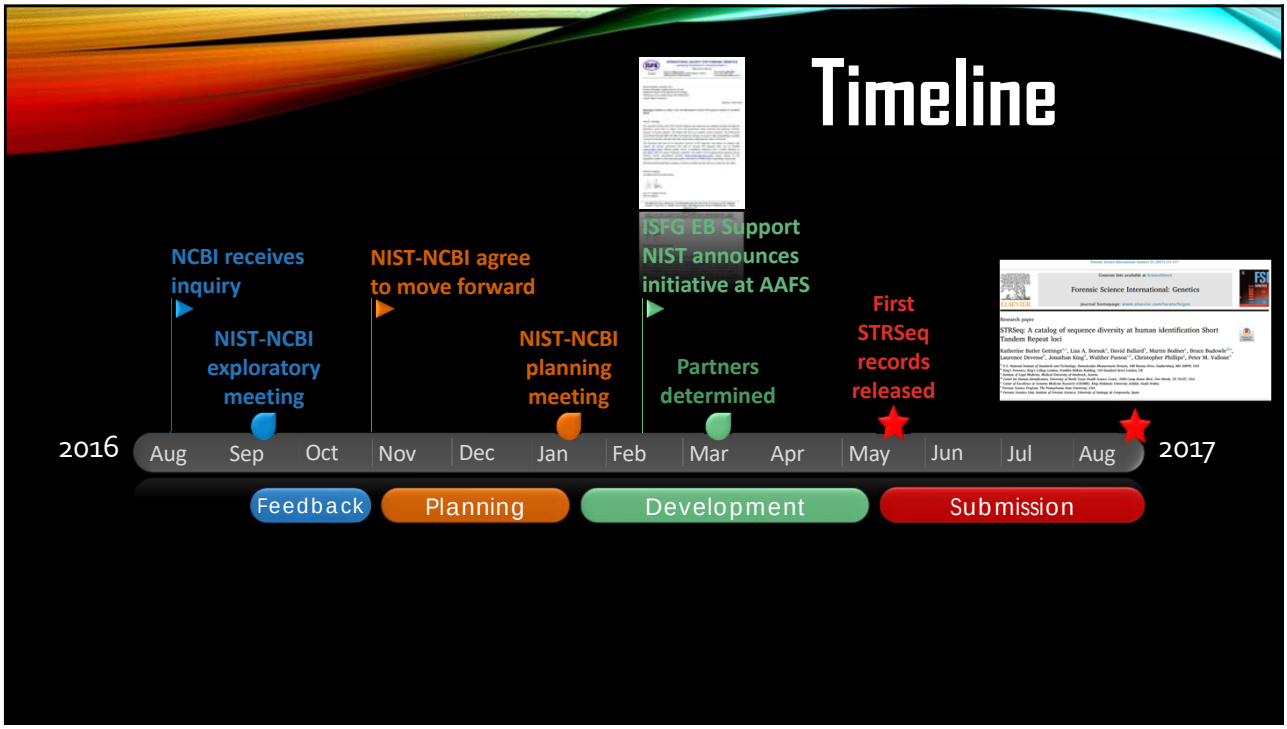
72


STRSeq:
GUARDIANS OF THE GALAXY

73


Inception
RefSeq: NCBI Reference Sequence Database
A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.
Locus·Reference·Genomic
LRG sequences provide a stable genomic DNA framework for reporting mutations with a permanent ID and core content that never changes.

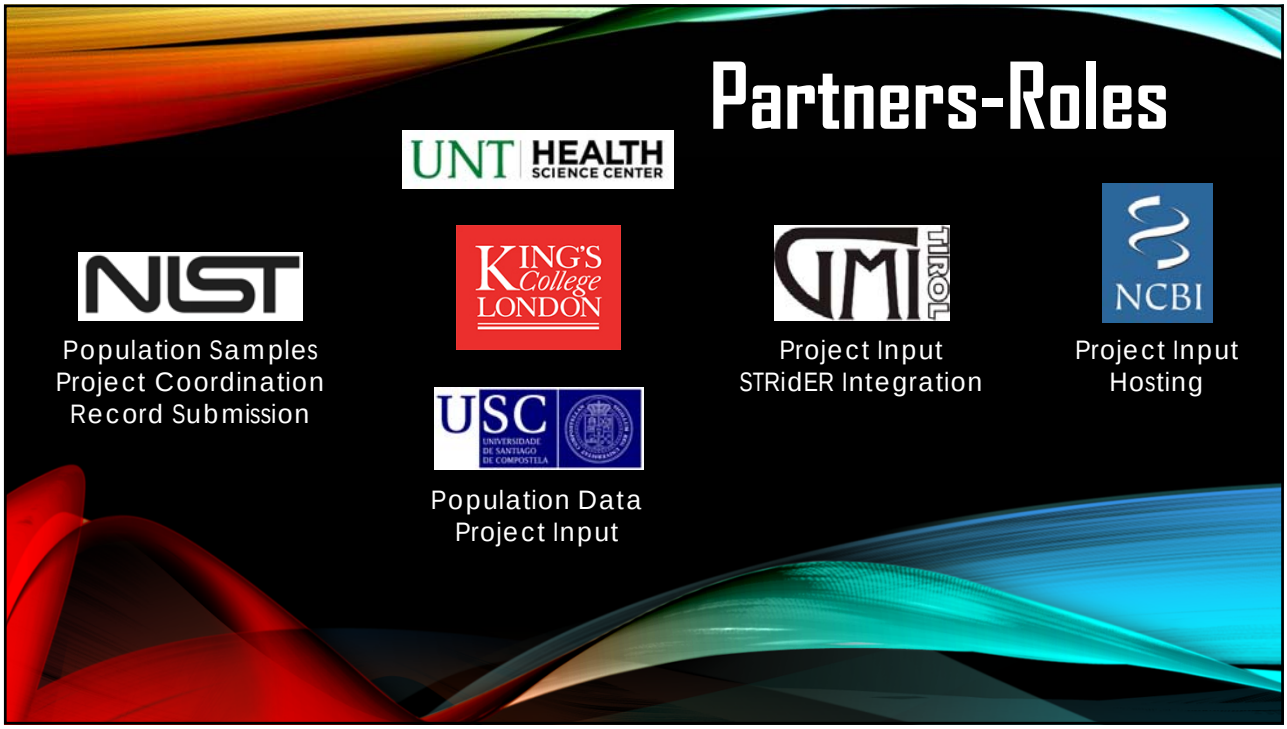
74


Timeline
NCBI receives inquiry
NIST-NCBI exploratory meeting
NIST-NCBI agree to move forward
NIST-NCBI planning meeting
ISFG EB Support NIST announces initiative at AAFS
Partners determined
First STRSeq records released
Forensic Science International: Genetics
STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci
2016
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
2017
Feedback
Planning
Development
Submission

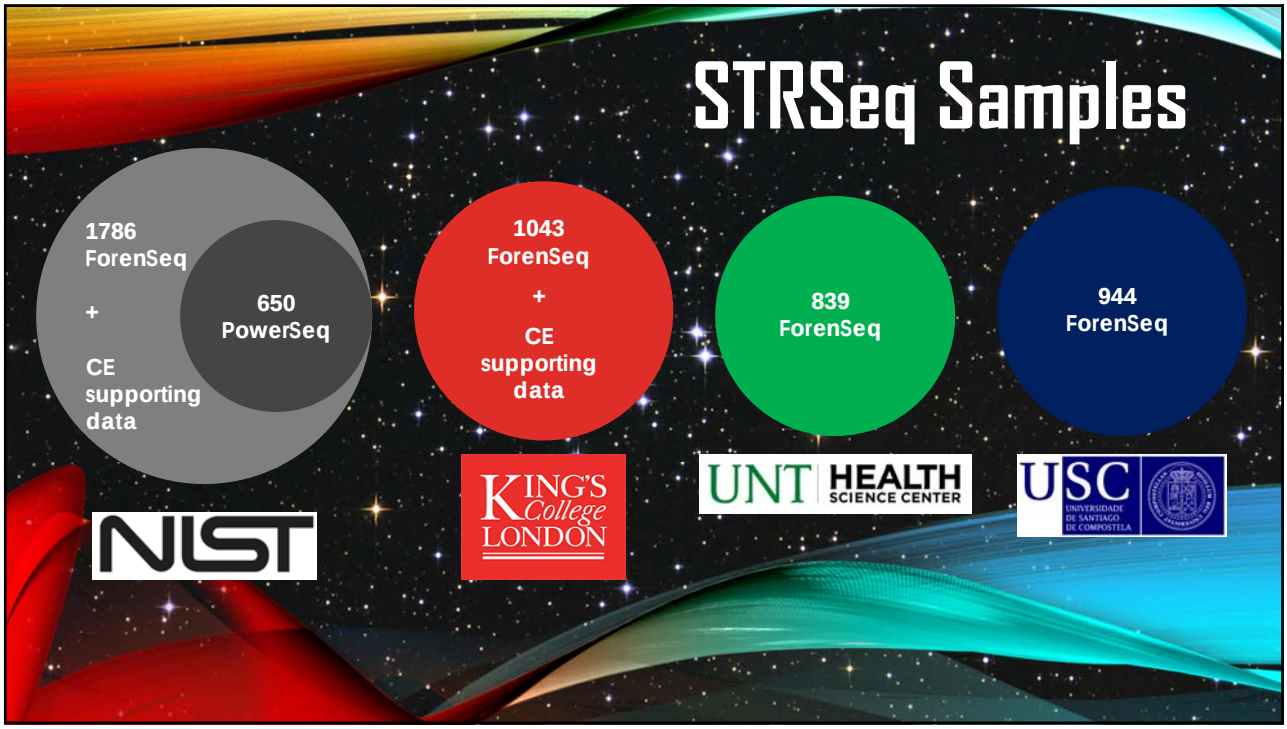
75


Partnership

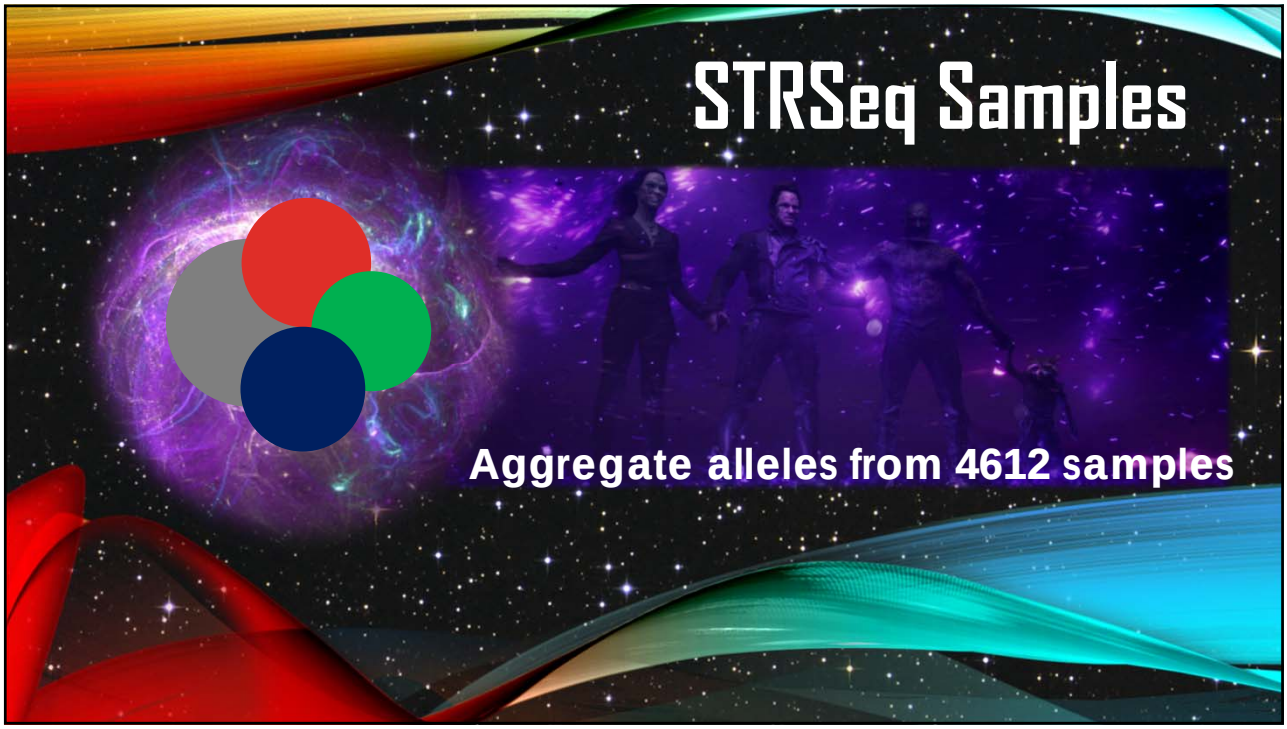
76


Partners-Roles
UNT HEALTH SCIENCE CENTER
NIST
Population Samples
Project Coordination
Record Submission
KING'S College LONDON
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
Population Data
Project Input
GMI TIROL
Project Input
STRidER Integration
NCBI
Project Input
Hosting

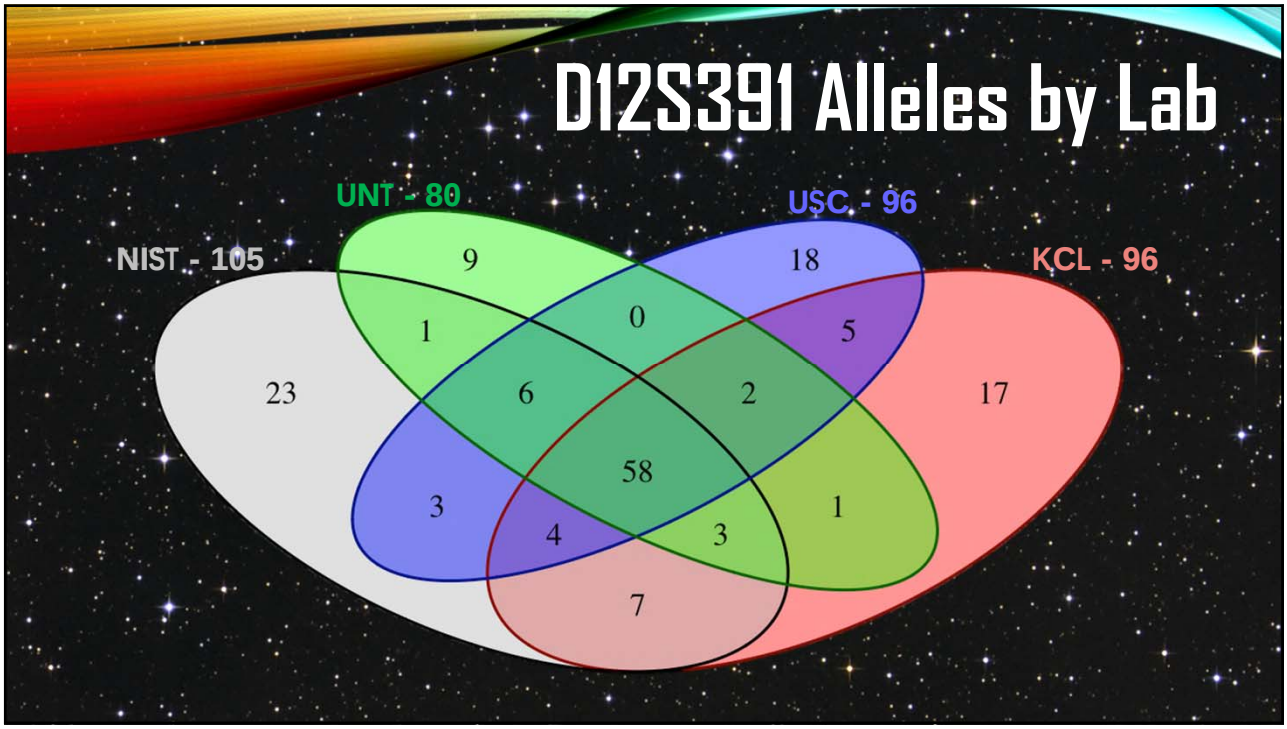
77


STRSeq Samples
1786 ForenSeq + CE supporting data
650 PowerSeq
1043 ForenSeq + CE supporting data
839 ForenSeq
944 ForenSeq
NIST
KING'S College LONDON
UNT HEALTH SCIENCE CENTER
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

78

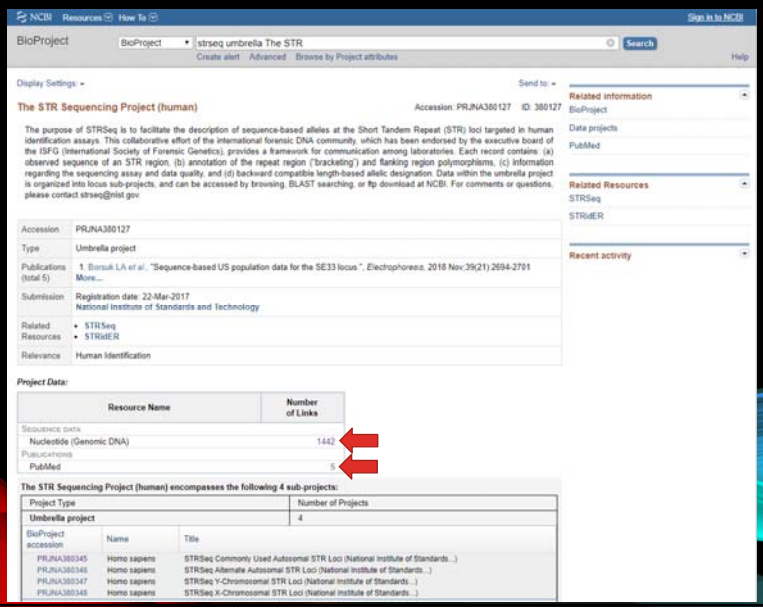

STRSeq Samples
Aggregate alleles from 4612 samples

79

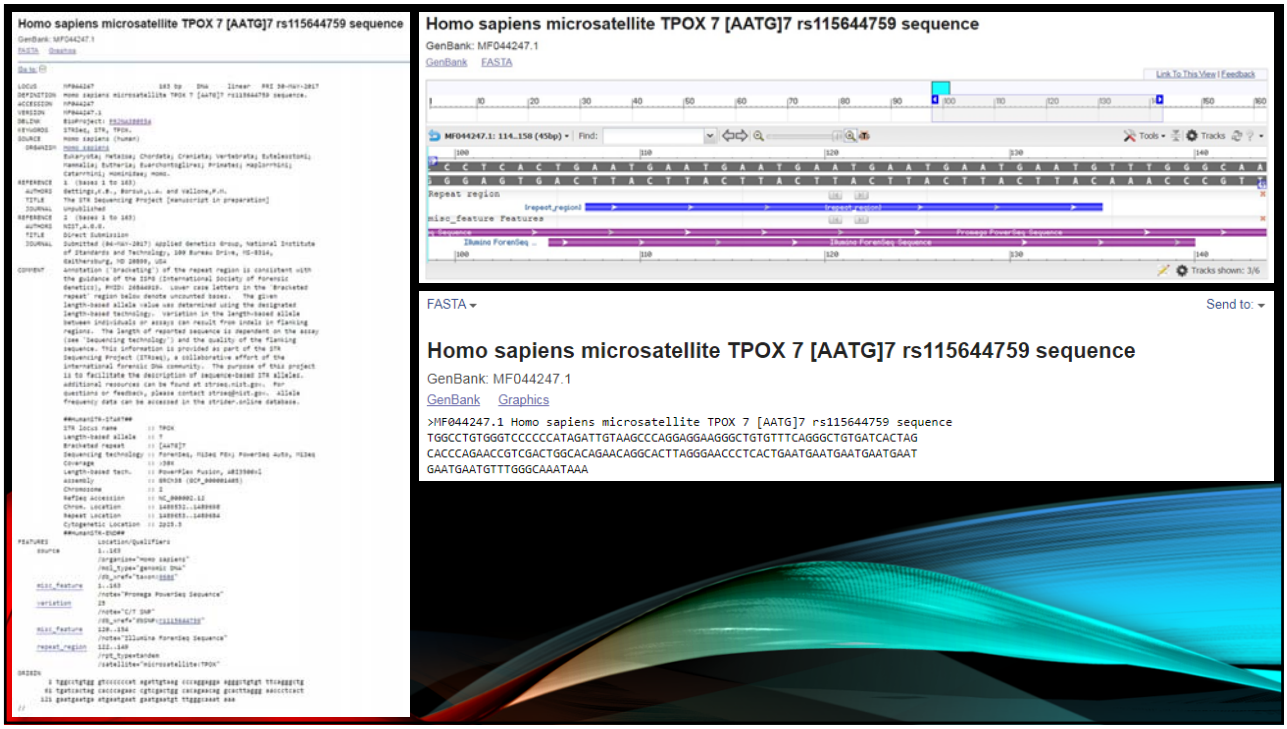

D12S391 Alleles by Lab
UNT - 80
USC - 96
NIST - 105
KCL - 96
9
18
1
0
5
23
6
2
17
58
3
1
4
3
7

80

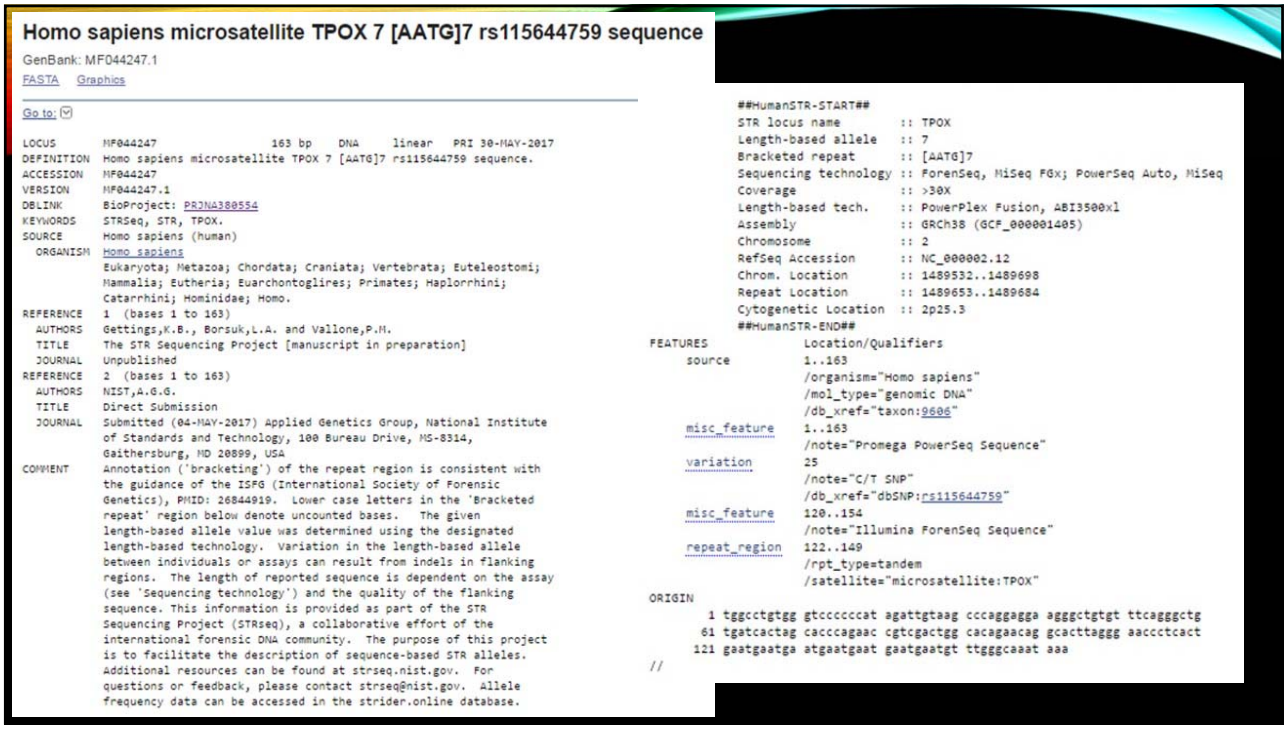
https://www.ncbi.nlm.nih.gov/bioproject/380127



81



82


Homo sapiens microsatellite TPOX 7 [AATG]7 rs115644759 sequence
GenBank: MF044247.1
FASTA Graphics
Go to:
LOCUS MF044247 163 bp DNA linear PRI 30-MAY-2017
DEFINITION Homo sapiens microsatellite TPOX 7 [AATG]7 rs115644759 sequence.
ACCESSION MF044247
VERSION MF044247.1
DBLINK BioProject: PRJNA380554
KEYWORDS STRSeq, STR, TPOX.
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 163)
AUTHORS Gettings,K.B., Borsuk,L.A. and Vallone,P.M.
TITLE The STR Sequencing Project [manuscript in preparation]
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 163)
AUTHORS NIST,A.G.G.
TITLE Direct Submission
JOURNAL Submitted (04-MAY-2017) Applied Genetics Group, National Institute of Standards and Technology, 100 Bureau Drive, MS-8314, Gaithersburg, MD 20899, USA
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 (see 'Sequencing technology') 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 :: TPOX
Length-based allele :: 7
Bracketed repeat :: [AATG]7
Sequencing technology :: ForenSeq, MiSeq FGx; PowerSeq Auto, MiSeq
Coverage :: >30X
Length-based tech. :: PowerPlex Fusion, ABI3500xl
Assembly :: GRCh38 (GCF_000001405)
Chromosome :: 2
RefSeq Accession :: NC_000002.12
Chrom. Location :: 1489532..1489698
Repeat Location :: 1489653..1489684
Cytogenetic Location :: 2p25.3
##HumanSTR-END##
FEATURES Location/Qualifiers
source 1..163
/organism="Homo sapiens"
/mol_type="genomic DNA"
/db_xref="taxon:9606"
misc_feature 1..163
/note="Promega PowerSeq Sequence"
variation 25
/note="C/T SNP"
/db_xref="dbSNP:rs115644759"
misc_feature 120..154
/note="Illumina ForenSeq Sequence"
repeat_region 122..149
/rpt_type=tandem
/satellite="microsatellite:TPOX"
ORIGIN
1 tggcctgtgg gtccccccat agattgtaag cccaggagga agggctgtgt ttcagggctg
61 tgatcactag cacccagaac cgtcgactgg cacagaacag gcacttaggg aaccctcact
121 gaatgaatga atgaatgaat gaatgaatgt ttgggcaaat aaa
//

83


Homo sapiens microsatellite TPOX 7 [AATG]7 rs115644759 sequence
GenBank: MF044247.1
GenBank FASTA
Link To This View | Feedback
MF044247.1: 114..158 (45bp)
Find:
Tools
Tracks
Repeat region
[repeat_region]
misc_feature Features
Promega PowerSeq Sequence
Illumina ForenSeq ...
Illumina ForenSeq Sequence
Tracks shown: 3/6

84

STRSeq in Bioinformatics

Standalone and API BLAST

Download BLAST
Get BLAST databases and executables

Use BLAST API
Call BLAST from your application

Use BLAST in the cloud
Start an instance at a cloud provider

Embedding the NCBI Sequence View in Web Content

Introduction

The NCBI Graphical Sequence Viewer (SV) is a general purpose tool for viewing biological sequence data. The Sequence Viewer has a very rich set of options and can display virtually any sequence. It can be embedded in a wide variety of web pages serving many different needs. This page has examples showing best practice for embedding Sequence Viewer with several different sets of options.

Homo sapiens microsatellite TPOX 7 [AATG]7 rs115644759 sequence

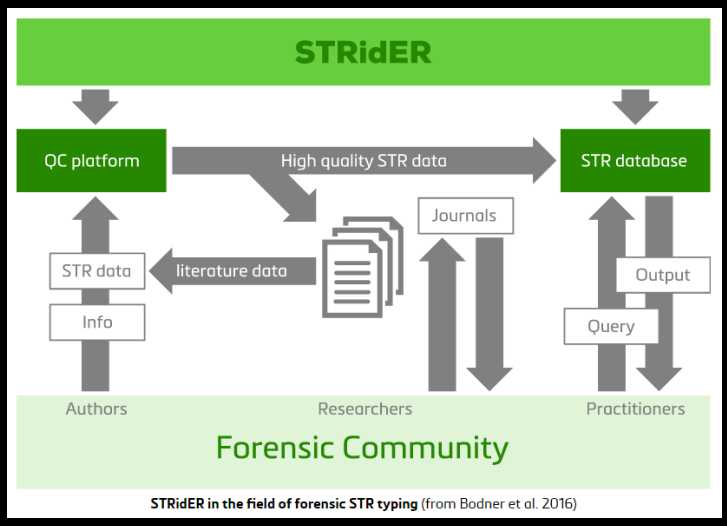
GenBank: MF044247.1

GenBank FASTA

85

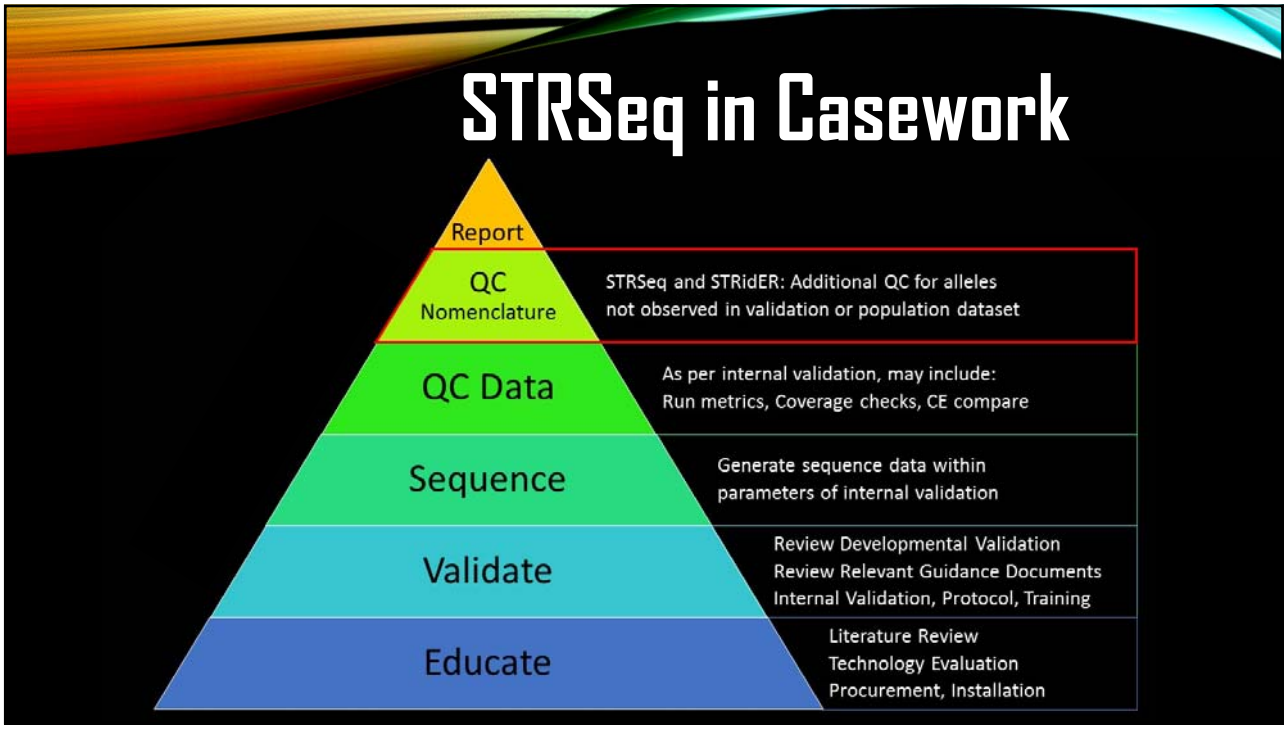
STRSeq in Population Data

STRSeq



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

86



87

FAQs:

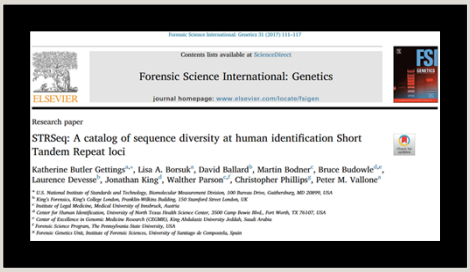
How do you decide which loci to include?

Why do some loci not have records yet?

Where does the data come from?

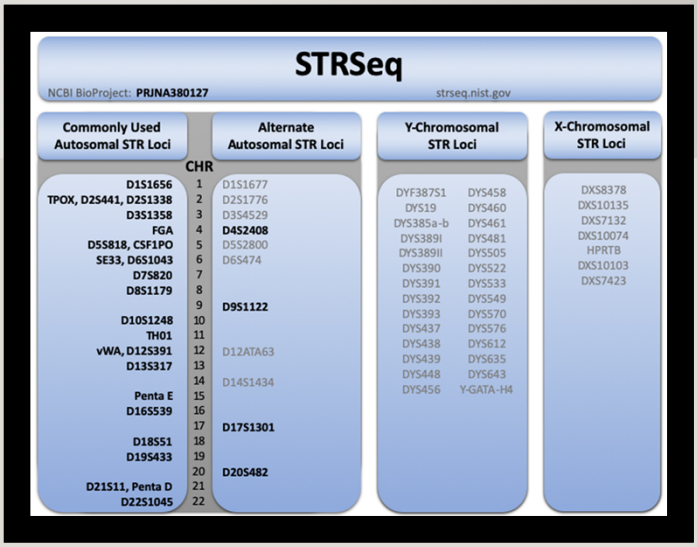
Where are the allele frequencies?

Can I send you a sequence?

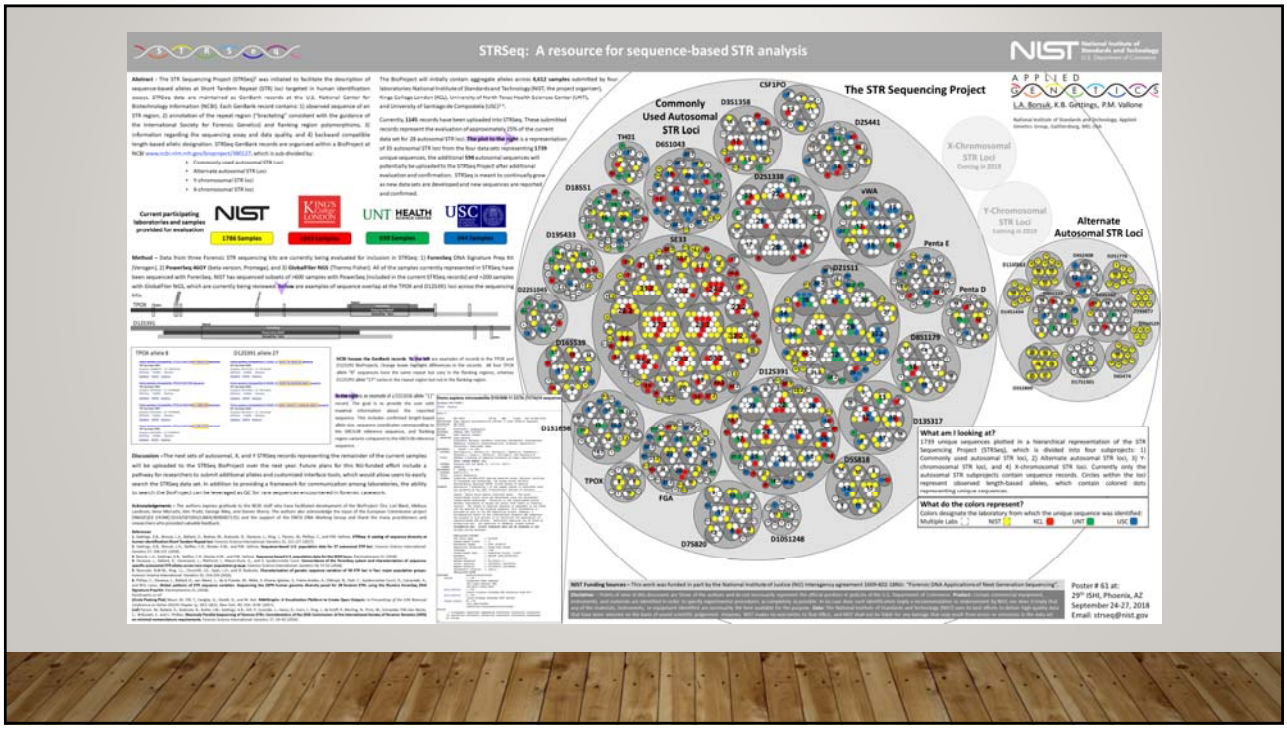

Forensic Science International: Genetics

Research paper

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

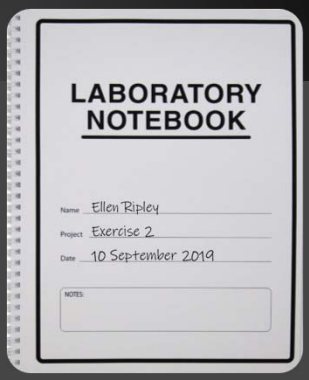


88



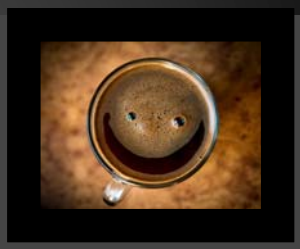
89

EXERCISE 2



90

COFFEE BREAK
meet back in 30 minutes



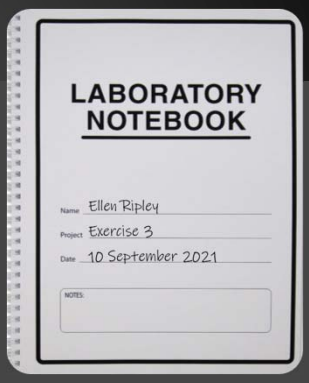
91

NEW IDEAS IN NOMENCLATURE

SID & STRNaming

92

EXERCISE 3



93

STRAND WG Efforts & WRAP UP

94

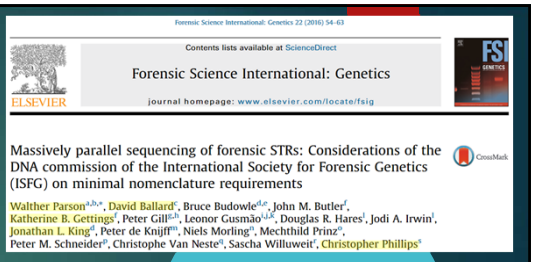
STRAND working group

align | name | define

95

STRAND working group

align | name | define



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]

Our mission is to harmonize related efforts across member laboratories:

- STRidER STR sequence quality control
- STRSeq catalog of sequences
- STRaitRazor bioinformatic freeware
- Forensic STR Sequence Guide

and to characterize additional STR loci present in the genome which may be useful for forensic purposes in the future.

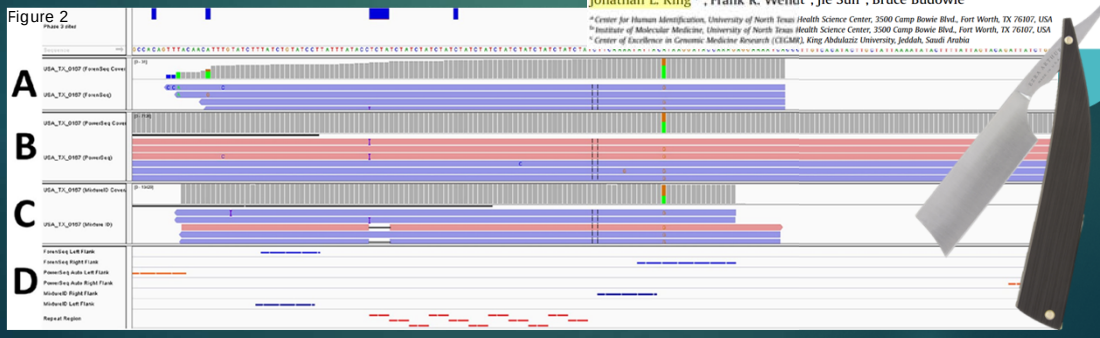
96

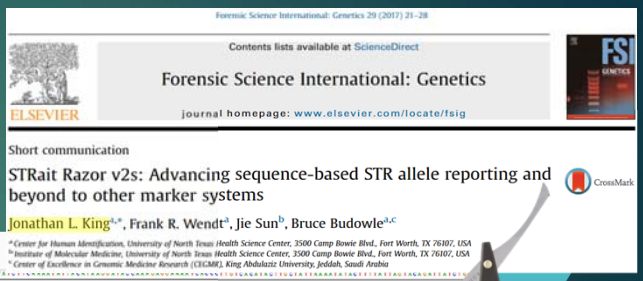
STRAND *working group*

align | name | define

- STRait Razor Agnostic Freeware

UNT HEALTH SCIENCE CENTER





Forensic Science International: Genetics 29 (2017) 21–28

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Short communication

STRait Razor v2s: Advancing sequence-based STR allele reporting and beyond to other marker systems

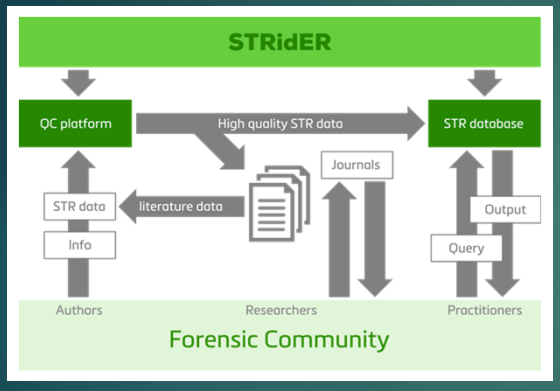
Jonathan L. King[a,*], Frank R. Wendt[a], Jie Sun[b], Bruce Budowle[a,c]

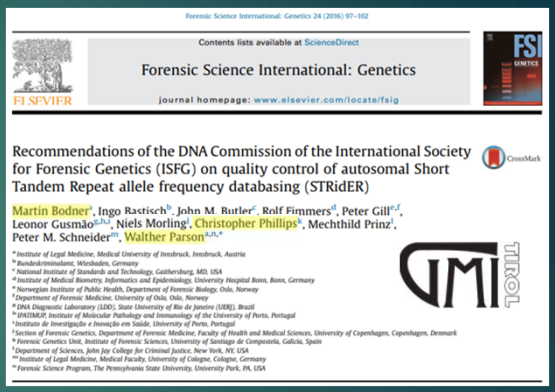
[a] Center for Human Identification, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA
[b] Institute of Molecular Medicine, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA
[c] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

97

STRAND *working group*

align | name | define





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,*]

*Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria
[b] Bundeskriminalamt, Wiesbaden, Germany
[c] National Institute of Standards and Technology, Gaithersburg, MD, USA
[d] Institute of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Bonn, Germany
[e] Norwegian Institute of Public Health, Department of Forensic Biology, Oslo, Norway
[f] Department of Forensic Medicine, University of Oslo, Oslo, Norway
[g] DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Brazil
[h] IPATIMUP, Institute of Molecular Pathology and Immunology of the University of Porto, Portugal
[i] Instituto de Investigação e Inovação em Saúde, University of Porto, Portugal
[j] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[k] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Galicia, Spain
[l] Department of Sciences, John Jay College for Criminal Justice, New York, NY, USA
[m] Institute of Legal Medicine, Medical Faculty, University of Cologne, Cologne, Germany
[n] Forensic Science Program, The Pennsylvania State University, University Park, PA, USA

The existing architecture of STRidER allows for the implementation of nucleotide sequence strings and thus is fully compatible with the QC of population data generated by MPS.

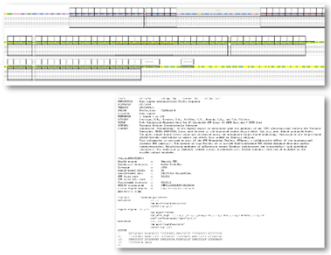
strider.online

98

STRAND *working group*
align | name | define

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



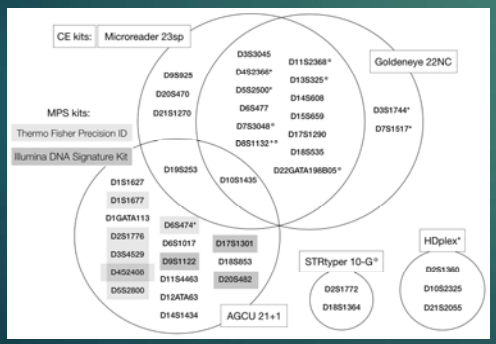
- translation into CE allele
- inspection of flanking region <u>where available</u>

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

99

STRAND *working group*
align | name | define

- Forensic STR Sequence Guide

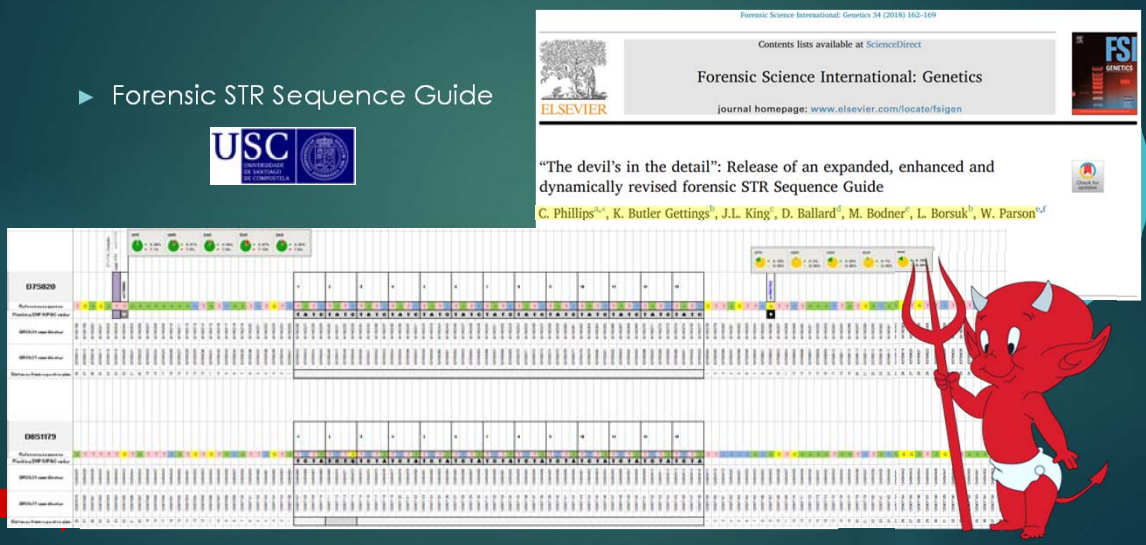


100

STRAND *working group*

align | name | define

- Forensic STR Sequence Guide



101

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:GML,Austria | L.Borsuk&K.Gettings:NST,US | J.King:UNT,US | C.Phillips:USC,Spain

102

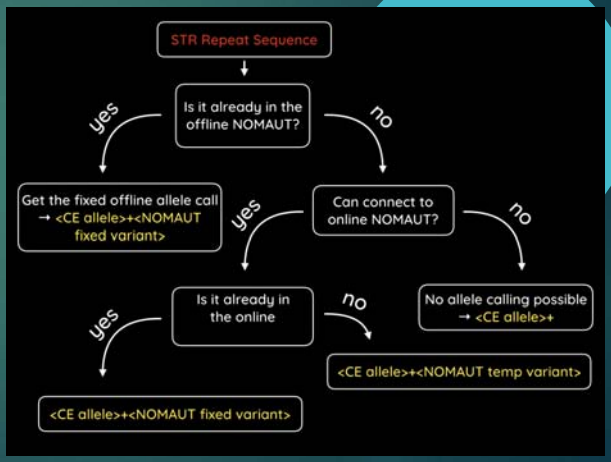
STR Nomenclature Meeting

Formats for STR Sequences

- Short Designator

+ NOM
✓ AUT

Sascha Willuweit
Charité Medical University

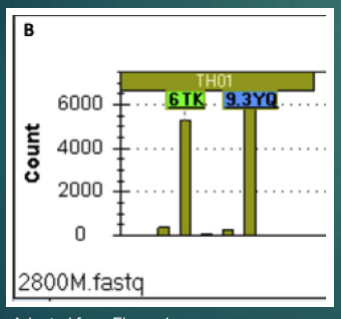


103

STR Nomenclature Meeting

Formats for STR Sequences

Short Designator



Adapted from Figure 1

Forensic Science International: Genetics 42 (2019) 14-20

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

A nomenclature for sequence-based forensic DNA analysis

Brian Young*, Tom Faris, Luigi Armogida

NicheVision Forensics, LLC. 526 South Main St. Akron, OH, 44311, USA

Adapted from Table 2

Sequence Type	Read Count	Sequence	SID Code
Allele	5,272	TCCATGGTGAATGAATGAATGAATGAATGAATGAGGGAAATAAGG	6 TK
Sequence Artifact	35	TCCATGGTGAATGAAcGAATGAATGAATGAATGAGGGAAATAAGG	6 VS
N-1 Stutter	368	TCCATGGTGAATGAATGAATGAATGAGGGAAATAAGG	5 LI
Allele	6,653	TCCATGGTGAATGAATGAATGAATGAATGATGAATGAATGAATGAGGGAAATAAGG	9.3 YQ
Sequence Artifact	11	TaCATGGTGAATGAATGAATGAATGAATGATGAATGAATGAATGAGGGAAATAAGG	9.3 MS
Sequence Artifact	11	TCCATGGTGAATGAATGAATaAATGAATGAATGATGAATGAATGAATGAGGGAAATAAGG	9.3 ZK
N-1 Stutter	232	TCCATGGTGAATGAATGAATGAATGAATGATGAATGAATGAATGAGGGAAATAAGG	8.3 WC

104

STR Nomenclature Meeting

Formats for STR Sequences

Short Designator

Longest Uninterrupted Stretch (LUS) Concept for Sequence Allele Designation

Was not proposed as sequence allele *nomenclature*

- Aim was representation of sequence alleles for probabilistic genotyping
- Practical, near-term solution for labs already employing prob gen for casework
  - Lack of program options for sequence alleles could be a barrier to NGS adoption
  - Could be used in mixture *interpretation* even prior to storage of sequence alleles in databases


Forensic Science International: Genetics

Use of the LUS in sequence allele designations to facilitate probabilistic genotyping of NGS-based STR typing results

Rebecca S. Just*, Jodi A. Irwin

Example D8S1179 alleles

Length (RU)	Bracketed format	LUS length	LUS Allele Designation (RU_LUS)
12	[TCTA]**12**	12	**12_12**
12	[TCTA]2 TCTG [TCTA]**9**	9	**12_9**
12	TCTA TCTG [TCTA]**10**	10	**12_10**

UNCLASSIFIED//FOUO 3

105

STR Nomenclature Meeting

Formats for STR Sequences

Bracketed Repeat

STRNaming from NFI

- Jerry Hoogenboom & Kris van der Gaag

```
CE11_TATC[8]TGTC[1]TATC[3]AATC[1]ATCT[3]
CE11_TATC[10]AATC[3]ATCT[3]
CE11_TATC[11]AATC[2]ATCT[3]
CE11_TATC[12]AATC[1]ATCT[3]
CE11_TATC[12]AATC[1]ATCT[3]_-24G>A
CE11_TATC[12]AATC[1]ATCT[3]_-25C>T
CE11_TATC[13]ATCT[3]
CE12_TATC[7]TATT[1]TATC[5]AATC[1]ATCT[3]
CE12_TATC[12]AATC[2]ATCT[3]
CE12_TATC[13]AATC[1]ATCT[3]
CE12_TATC[13]AATC[1]ATCT[3]_-24G>A
CE12_TATC[13]AATC[1]ATCT[3]_-25C>T
CE12_TATC[13]AATC[2]ATCT[2]
CE12_TATC[14]ATCT[3]
CE13_TATC[13]AATC[2]ATCT[3]
CE13_TATC[14]AATC[1]ATCT[3]
CE13_TATC[14]AATC[1]ATCT[3]_-24G>A
CE13_TATC[14]AATC[1]ATCT[3]_-25C>T
CE13_TATC[15]AATC[1]ATCT[3]_+9GTCT>-
CE13_TATC[15]ATCT[3]
```

106

Flanking Region Polymorphism Nomenclature

Option	Description	Example	Requires Range	Requires Reference Genome	Requires InDel Alignment Parameters
1	Report differences relative to a reference genome	Chr5:123775552:C>A	Yes	Yes	Yes
2	Report differences relative to repeat region	+4C>A	Yes	Sort of	Yes
3	Report rs number and change	rs73801920 C>A	Yes	No, but	No

107

STR Nomenclature Meeting

Formats for STR Sequences

Full String = Unequivocal Record

Storage method/location is lab-determined

"At this time, forensic DNA databasing software (e.g. CODIS) is generally not equipped to store or search STR sequence strings.

Such databases primarily contain convicted offender samples; therefore, enabling STR sequence storage or search capabilities may be of limited use until laboratories begin routinely sequencing this sample type.

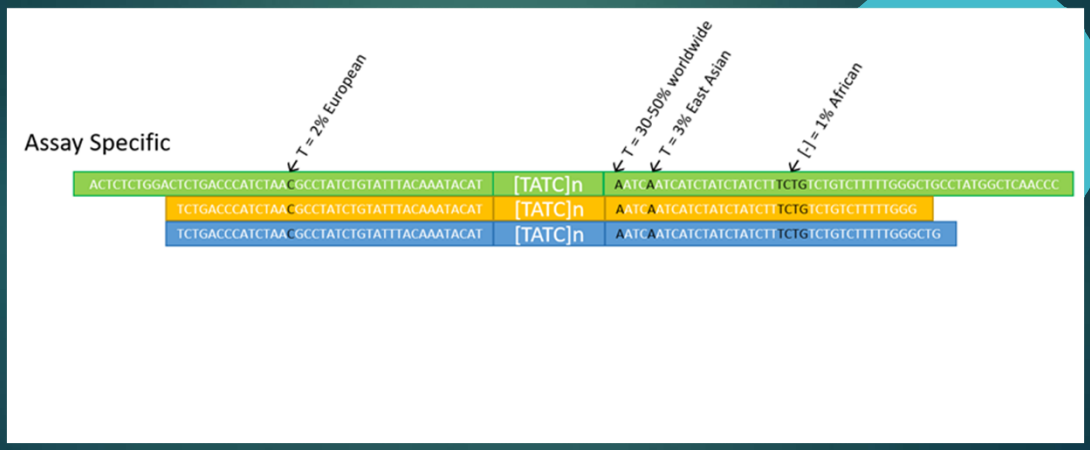
In the interim, length based (numerical allele) profiles can be developed via STR sequencing assays.

Analysts confirming interlaboratory matches could compare sequence data, when applicable."

108

STR Nomenclature Meeting

Defined Coordinates



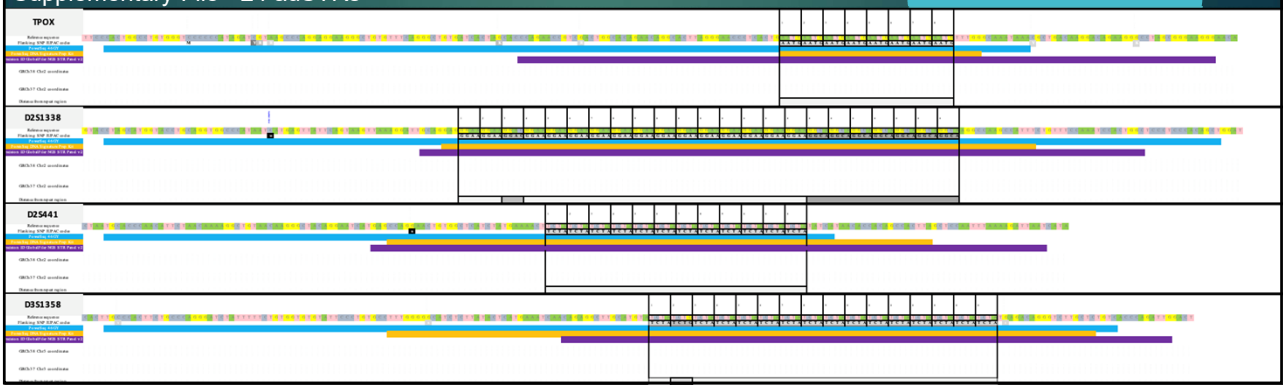
109

STR Nomenclature Meeting

Defined Coordinates

PowerSeq 46GY	GeneMarker NGS Range
ForenSeq DNA Signature Prep Kit	UAS Flanking Region Report Range
Precision ID GlobalFiler NGS v2	Converge .bed file range

Supplementary File - 24 auSTRs

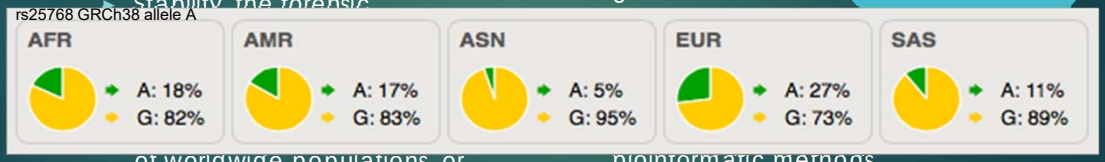


110

STR Nomenclature Meeting

Forensic Specific Reference

- Advantages
  - Elimination of rare SNP alleles in STR flanking regions, incorporation of known insertions
  - Stability, the forensic
- Disadvantages
  - Significant effort would be required for curation, maintenance, version control, and enforcement of general use within the



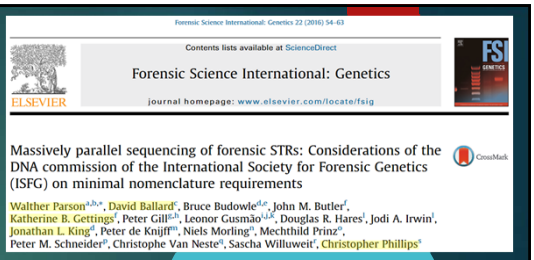
of worldwide populations, or representative of maximal complexity

bioinformatic methods

111

STRAND *working group*

align | name | define


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]

Our mission is to harmonize related efforts across member laboratories:

- STRidER STR sequence quality control
- STRSeq catalog of sequences
- STRaitRazor bioinformatic freeware
- Forensic STR Sequence Guide

and to characterize additional STR loci present in the genome which may be useful for forensic purposes in the future.

112

STRAND *working group*

align | name | define

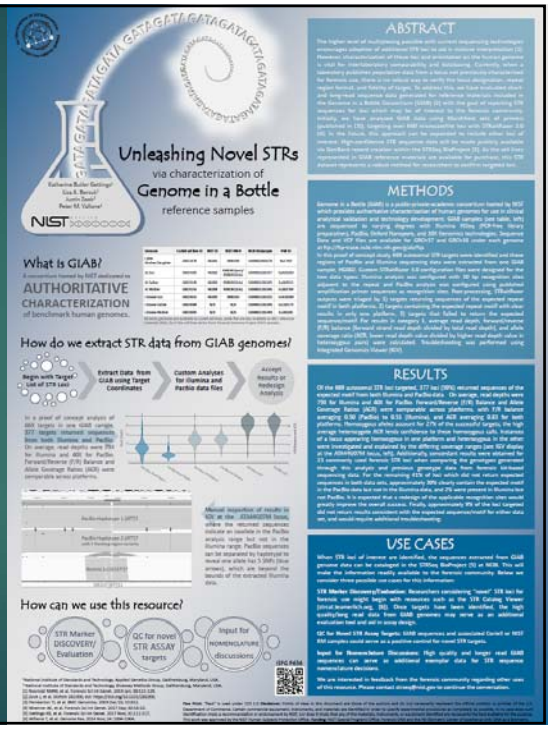
Genome in a Bottle (GIAB)

7 Coriell cell lines

- One individual and two trios
- PCR-free prep, HiSeq, PacBio, ONT, 10X

Analyzing STR regions in GIAB samples

- Any "novel" marker can be characterized
- Proof of concept targeting >600 STRs



QUESTIONS?

enjoy your week at the
28th ISFG in Prague

Questions later? katherine.gettings@nist.gov strseq@nist.gov strandwg@gmail.com