WORKSHOP #10
International Society for Forensic Genetics
Tuesday, August 29, 2017

Autosomal STR Markers and Interpretation

Organized by John M. Butler and Lisa Borsuk
U.S. National Institute of Standards and Technology

Time	Topic	Presenter
Understanding STR Markers and Measurements		
09:00 – 09:30	Introductions & Expectations Reviewed, STR Kits & Measurement Techniques	John Butler
09:30 – 10:00	STR Markers Commonly Used	Lisa Borsuk
10:00 – 10:30	Interpretation Issues	John Butler
10:30 – 11:00	Length vs Sequence Information: Lessons Learned from TPOX and SE33	Lisa Borsuk
11:00 – 11:30	BREAK	
Communicating and Sharing STR Information		
11:30 – 12:00	STR Nomenclature, STRSEQ, STRidER	Katherine Gettings
12:00 – 12:30	STRBase and Revisions Planned	John Butler & Lisa Borsuk
12:30 – 13:00	Other Uses with Forensic STR Markers and Potential Privacy Concerns	John Butler

Core STR Sets

United States:

CODIS 13 *(1997-2017)*: TH01, TPOX, CSF1PO, FGA, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11

CODIS 20 *(2017-present)*: TH01, TPOX, CSF1PO, FGA, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, D1S1656, D2S441, D2S1338, D10S1248, D12S391, D19S433, D22S1045

European Standard Set:

ESS 7 *(1998-2009)*: TH01, FGA, vWA, D3S1358, D8S1179, D18S51, D21S11

ESS 12 *(2009-present)*: TH01, FGA, vWA, D3S1358, D8S1179, D18S51, D21S11, D1S1656, D2S441, D10S1248, D12S391, D22S1045

United Kingdom:

SGM *(1995-1999)*: TH01, FGA, vWA, D8S1179, D18S51, D21S11, amelogenin

SGM Plus *(1999-2014):* TH01, FGA, vWA, D8S1179, D18S51, D21S11, D2S1338, D3S1338, D16S539, D19S433, amelogenin

DNA-17 *(2014-present)*: TH01, FGA, vWA, D3S1358, D8S1179, D18S51, D21S11, D1S1656, D2S441, D10S1248, D12S391, D22S1045, D2S1338, D16S539, D19S433, SE33, amelogenin

Australia:

Profiler Plus: FGA, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D18S51, D21S11, amelogenin

PowerPlex 21: FGA, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D18S51, D21S11, D1S1656, D2S1338, D6S1043, D12S391, D16S539, D19S433, CSF1PO, Penta D, Penta E, TH01, TPOX, amelogenin

29 August 2017

Workshop #10

Autosomal STR Markers and Interpretation

John M. Butler, Ph.D.
Special Programs Office
U.S. National Institute of Standards and Technology

Lisa Borsuk, M.S.
Applied Genetics Group
U.S. National Institute of Standards and Technology

Katherine Gettings, Ph.D.
Applied Genetics Group
U.S. National Institute of Standards and Technology

Current Forensic DNA Testing

From presentation given by John Butler to forensic science managers at Interpol in October 2016

- **Short tandem repeat (STR) markers** are used
  - **Typically 15 to 22 STRs examined** with commercial kits (e.g., Identifiler, PowerPlex 16, NGM, GlobalFiler, Fusion)
- STR length (and sequence) varies among individuals
  - DNA molecules are **labeled with fluorescent dyes and separated by size using CE** (capillary electrophoresis)
  - **Only the STR length is measured** against an internal size standard and calibrated with an allelic ladder (which is a combination of the most common possibilities of alleles)
- **National DNA databases** using STR markers now exist in >50 countries (>75 million STR profiles total)
  - **Having core STR markers in common is critical** to enable comparisons across laboratories and between countries

Purpose and Value of this Workshop

- **Aid understanding of autosomal STR markers widely used in forensic genetics and issues involved with data interpretation**
- Autosomal STR markers will likely be used for years to come
  - National **DNA databases continue to expand** (~75-100 million STR profiles worldwide)
  - Recent rapid growth in the number of available STR typing kits due to (1) **expansion of core loci** in Europe (2011; 7 → 12) and United States (2017; 13 → 20) and (2) **patent coverage expiring**
  - Possible STR typing methodologies are expanding due to (1) **new CE instruments**, (2) **rapid DNA systems**, and (3) **massively parallel sequencing** (next-generation sequencing) technologies
- **Receive input on revisions to the NIST STRBase website**

Workshop#10 Outline

Autosomal STR Markers and Interpretation

Time *(approximate)*	Topic	Presenter
	Understanding STR Markers and Measurements	
09:00 – 09:30	Introductions & Expectations Reviewed, STR Kits & Measurement Techniques	John Butler
09:30 – 10:00	STR Markers Commonly Used	Lisa Borsuk
10:00 – 10:30	Interpretation Issues	John Butler
10:30 – 11:00	Length vs Sequence Information: Lessons Learned from TPOX and SE33	Lisa Borsuk
11:00 – 11:30	BREAK	
	Communicating and Sharing STR Information	
11:30 – 12:00	STR Nomenclature, STRSEQ, STRidER	Katherine Gettings
12:00 – 12:30	STRBase and Revisions Planned	John Butler & Lisa Borsuk
12:30 – 13:00	Other Uses with Forensic STR Markers and Potential Privacy Concerns	John Butler

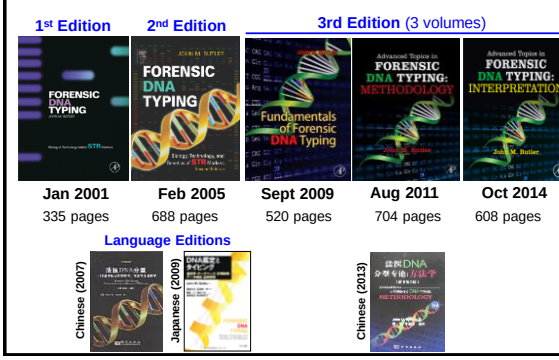
29 August 2017

Workshop #10

Introductions

John M. Butler, Ph.D.
Special Programs Office
U.S. National Institute of Standards and Technology

Forensic DNA Typing Textbooks Have Set the Standard for the Field



Review of STR Allele Sequence Variation

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

STR allele sequence variation: Current knowledge and future issues

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

[a] US National Institute of Standards and Technology, Biomolecular Measurement Division, 100 Bureau Drive, Gaithersburg, MD 20899, USA
[b] The George Washington University, Department of Forensic Sciences, 2100 Foxhall Road NW, Washington, DC 20007, USA
[c] US. National Institute of Standards and Technology, Special Programs Office, 100 Bureau Drive, Gaithersburg, MD 20899, USA

ARTICLE INFO

Article history:
Received 13 February 2015
Received in revised form 20 May 2015
Accepted 10 June 2015

Keywords:
Forensic DNA
Short tandem repeat
STR loci
DNA database
Sequence variation
STR nomenclature

ABSTRACT

This article reviews what is currently known about short tandem repeat (STR) allelic sequence variation in and around the twenty-four loci most commonly used throughout the world to perform forensic DNA investigations. These STR loci include D1S1656, TPOX, D2S441, D2S1338, D3S1358, FGA, CSF1PO, D5S818, SE33, D6S1043, D7S820, D8S1179, D10S1248, TH01, vWA, D12S391, D13S317, Penta E, D16S539, D18S51, D19S433, D21S11, Penta D, and D22S1045. All known reported variant alleles are compiled along with genomic information available from GenBank, dbSNP, and the 1000 Genomes Project. Supplementary files are included which provide annotated reference sequences for each STR locus, characterize genomic variation around the STR repeat region, and compare alleles present in currently available STR kit allelic ladders. Looking to the future, STR allele nomenclature options are discussed as they relate to next generation sequencing efforts underway.

Published by Elsevier Ireland Ltd.

Gettings, K.B., Aponte, R.A., Vallone, P.M., Butler, J.M. (2015). STR allele sequence variation: current knowledge and future issues. *Forensic Science International: Genetics, 18,* 118-130

Introductions & Expectations

1. **Your Name?**
2. **Your Laboratory/Employer?**
3. **Your Experience with STRs?**
   a) 0-2 years
   b) 2-10 years
   c) >10 years
4. **What you hope to learn in this workshop?**

Greg Matheson on Forensic Science Philosophy

The CAC News – 2nd Quarter 2012 – p. 6
"Generalist vs. Specialist: a Philosophical Approach"
http://www.cacnews.org/news/2ndq12.pdf

- If you want to be a technician, performing tests on requests, then just focus on the policies and procedures of your laboratory. If you want to be a scientist and a professional, learn the policies and procedures, but go much further and learn the philosophy of your profession. **Understand the importance of why things are done** the way they are done, the scientific method, the viewpoint of the critiques, the issues of bias and the importance of ethics.

Advantages for STR Markers

- Small product sizes are generally **compatible with degraded DNA** and PCR enables recovery of information from **small amounts of material**
- Numerous alleles per locus aid **mixture interpretation**
- Multiplex amplification with fluorescence detection enables **high power of discrimination** in a single test
- Commercially available in an **easy to use kit format**
- Uniform set of **core STR loci** provide capability for national (and international) sharing of criminal DNA profiles

Standard Approaches Enable Reliable DNA Data Comparison

- **Core loci**
  - In 1997, U.S. selected 13 core STR markers
  - Europe moved from 7 to 12 core STR loci in November 2011
  - U.S. moved to 20 core STRs in January 2017
- **Common data formats**
  - ISFG DNA Commission allele nomenclature designation recommendations
  - ANSI/NIST-ITL standard for data storage and transmission
- **Commercial STR kits**
  - Consistent allelic ladders
- **Certified reference materials**
  - NIST SRM 2391c (certified values for STR allele measurements)

National Institute of Standards and Technology (NIST)

- Started in 1901 with roots back to the Constitution
- Name changed to **National Institute of Standards and Technology (NIST)** from National Bureau of Standards in 1988
- Primary campus in Gaithersburg, Maryland (just outside of Washington, D.C.)
- Part of the U.S. Department of Commerce
- >3,000 employees and >2,000 associates
- Supply >1300 reference materials
- Defines official time for the U.S.

DNA reference materials can help calibrate laboratory results and enable traceability to a common standard

DNA SRM 2391c Certificate Updated 3 April 2015

Will soon be moving to SRM 2391d

National Institute of Standards & Technology
Certificate of Analysis
Standard Reference Material® 2391c
PCR-Based DNA Profiling Standard

What's New? *Addition of Sanger sequencing analysis; additional STR genotyping test kits used towards certification; extension of certification date; editorial changes*

Certified Genotypes/Haplotypes
25 autosomal STR loci and amelogenin
29 Y-STR loci

6 Components
A (single-source female genomic DNA)
B (single-source male genomic DNA)
C (single-source male genomic DNA)
D (3:1 mixture of A and C)
E (female cells on 903 paper)
F (male cells on FTA paper)

Reference Genotypes
26 autosomal STRs

Information Genotypes/Haplotypes
1 autosomal STR: Penta C
12 X-STR loci
30 InDels (DIPlex)

STR Kit Coverage
Thermo Fisher Applied Biosystems (Foster City, CA): AmpFlSTR Identifiler, Identifiler Plus, NGM, NGM SElect, Cofiler, Profiler, Profiler Plus, Profiler Plus ID, SGM Plus, SEfiler, MiniFiler, GlobalFiler, YFiler, YFiler Plus
Promega Corporation (Madison, WI): PowerPlex 16, 16 HS, ESX 17, ESI 17, ES, SS, ESI 17 Pro, ESI 17 Fast, ESX 17 Fast, 18D, 21, CS7, Fusion, Y, Y23
Qiagen (Hilden, Germany): Investigator ESSplex, IDplex, ESSplex SE, ESSplex SE Plus, ESSplex SE GO!, IDplex Plus, IDplex GO!, 24plex, 24plex GO!, Argus X-12, DIPlex

Allele Sequences Provided in New SRM 2391c Certificate to Aid Use with Next-Generation Sequencing

Table 14. Autosomal STR Sequencing for Component E

SRM 2391c – Component E				
Marker	Length-based Types	Sanger Result	Repeat Structure –Allele 1	Repeat Structure –Allele 2
D1S1656	11, 16.3	11, 16.3	[TAGA]$_{11}$ [TG]$_5$	[TAGA]$_4$ TGA [TAGA]$_{11}$ TAGG [TG]$_5$
D2S1338	19, 20	19, 20	[TGCC]$_2$ [TTCC]$_{12}$	[TGCC]$_2$ [TTCC]$_{13}$
D2S441	10, 10	10, 10	[TCTA]$_{10}$	[TCTA]$_2$ TCTG [TCTA]$_{10}$
D3S1358	14, 15	14, 15	TCTA [TCTG]$_2$ [TCTA]$_{11}$	TCTA [TCTG]$_2$ [TCTA]$_{12}$
D5S818	11, 13	11, 13	[AGAT]$_{11}$	[AGAT]$_{13}$
D6S1043	11, 11	11, 11	[AGAT]$_{11}$	[AGAT]$_{11}$
D7S820	8, 10	8, 10	[GATA]$_8$	[GATA]$_{10}$
D8S1179	11, 13	11, 13	[TCTA]$_{11}$	[TCTA] TCTG [TCTA]$_{11}$
D8S1115	9, 16	9, 16	[ATT]$_9$	[ATT]$_{16}$
D10S1248	14, 14	14, 14	[GGAA]$_{14}$	[GGAA]$_{14}$
D12S391	17, 22	17, 22	[AGAT]$_{10}$ [AGAC]$_6$ AGAT	[AGAT]$_{13}$ [AGAC]$_8$ AGAT
D13S317	8, 12	8, 12	[TATC]$_8$	[TATC]$_{12}$, A→T SNP 1 bp ds from repeat
D16S539	11, 12	11, 12	[GATA]$_{11}$	[GATA]$_{12}$
D18S51	14, 17	14, 17	[AGAA]$_{14}$	[AGAA]$_{17}$
D19S433	14, 14	14, 14	[AAGG] AAAG [AAGG] TAGG [AAGG]$_{12}$	[AAGG] AAAG [AAGG] TAGG [AAGG]$_{12}$
D21S11	29, 30	29, 30	[TCTA]$_4$ [TCTG]$_6$ {[TCTA]$_3$ TA [TCTA]$_3$ TCA [TCTA]$_2$ TCCATA} [TCTA]$_{11}$	[TCTA]$_2$ TATA [TCTA]$_3$ [TCTG]$_5$ {[TCTA]$_3$ TA [TCTA]$_3$ TCA [TCTA]$_2$ TCCATA} [TCTA]$_{11}$

SRM 2391c Certificate of Analysis (issued 3 April 2015)

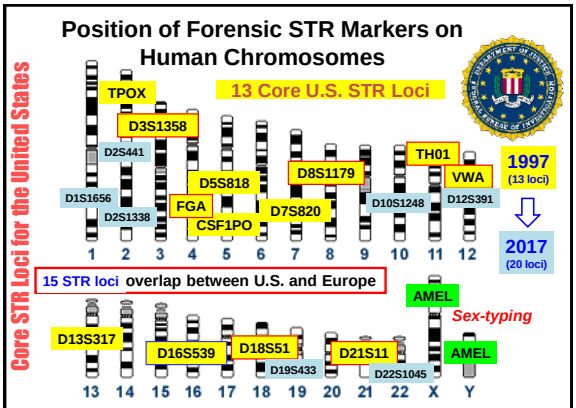
STR Kits

NIST Disclaimer

Points of view are those of the presenters and do not necessarily represent the official position or policies of the National Institute of Standards and Technology.

Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

Position of Forensic STR Markers on Human Chromosomes



U.S. has moved to 20 core loci

Locus
CSF1PO
D3S1358
D5S818
D7S820
D8S1179
D13S317
D16S539
D18S51
D21S11
FGA
TH01
TPOX
vWA
D1S1656
D2S441
D2S1338
D10S1248
D12S391
D19S433
D22S1045

Red is for original CODIS Core 13 Loci.
Blue is for new additional CODIS Core Loci.

Forensic Science International: Genetics 17 (2015) 33–34
Contents lists available at ScienceDirect
Forensic Science International: Genetics
journal homepage: www.elsevier.com/locate/fsig

Letter to the Editor

Required in U.S. starting January 1, 2017

Selection and implementation of expanded CODIS core loci in the United States

"The CODIS Core Loci Working Group selected a consortium of 11 CODIS laboratories…these laboratories performed validation experiments…

With the assistance of the National Institute of Standards and Technology (NIST), the data generated through these validation studies were compiled, reviewed and analyzed."

Hares, D.R. (2015) Selection and implementation of expanded CODIS core loci in the United States. *Forensic Sci. Int. Genet.* 17:33-34

U.S. Core Loci Expansion Efforts

More loci added as databases grew…

U.S. Core Loci Goals Announced
1990: 4 VNTRs (RFLP)
1997: 13 STRs (PCR)
2011: 20+ STRs

1990 U.S. began with 4 RFLP VNTRs
Baechtel et al. (1991) **Combined DNA Index System (CODIS) proposed**
1994
1996 Early STR kits
1997 DNA Identification Act (federal law)
1998 **Launch of U.S. National DNA Database (NDIS)** *Initial QAS released*
2000 Budowle et al. (1998) Initial CODIS core loci (13 STRs)
2002 NDIS exceeds 1 million profiles
2003 16plex STR kits; President's DNA Initiative: Debbie Smith Act increases funding for DNA databases
2009 Europe expands from 7 to 12 loci
2011 NDIS exceeds 10 million profiles
2012 PowerPlex Fusion and GlobalFiler **24plex kits available**
2015 Implementation to be required 2 years after announcement

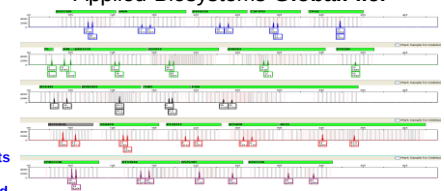
CODIS Core Loci WG recommend new loci

Hares (2012a, 2012b) Letters to editor (1) & (2) announcing proposed new loci

Butler et al. (2012) NIST population data published covering all 29 kit STR loci

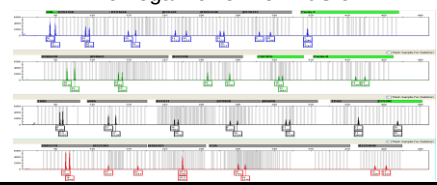
QAS: Quality Assurance Standards

Applied Biosystems **GlobalFiler**



Both kits were released in Sept 2012

Promega PowerPlex **Fusion**



Autosomal STR Typing Kits (for CE use)

Promega	**Thermo Fisher Scientific** (formerly Applied Biosystems)	**QIAGEN**
PowerPlex	*AmpFlSTR*	*Investigator*
PowerPlex 16 (HS)	Profiler Plus	ESSplex Plus
PowerPlex 18D	COfiler	ESSplex SE Plus (QS, GO!)
PowerPlex ESI 16 (Fast)	Profiler	Hexaplex ESS
PowerPlex ESX 16 (Fast)	SGM Plus	Nonaplex ESS
PowerPlex ESI 17 Pro (Fast)	SEfiler Plus	Decaplex SE
PowerPlex ESX 17 (Fast)	SinoFiler	IDplex Plus (GO!)
PowerPlex 21	MiniFiler	HDplex
PowerPlex CS7	**Identifiler (Direct, Plus)**	
	VeriFiler Express	**24plex** (QS, GO!)
PowerPlex Fusion	NGM	
PowerPlex Fusion 6C	NGM SElect (Express)	
PowerPlex 35GY 8C	NGM Detect	
	GlobalFiler (Express)	

Other Published STR Kits & Assays

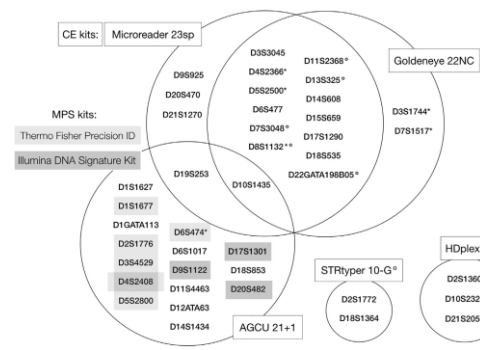
- **Spain** • **I-DNASE21 System** (21 autosomal STRs, amelogenin)
  - Aznar, J.M., et al. (2014). I-DNASE21 system: development and SWGDAM validation of a new STR 21-plex reaction. *Forensic Science International: Genetics, 8(1),* 10-19.
- **China** (Guangzhou) • **HomyGene19+14Y System** (18 autosomal STRs, 14 Y-STRs, amelogenin)
  - Du, W., et al. (2017). Developmental validation of the HomyGene19+14Y System. *International Journal of Legal Medicine, 131(3),* 605-620.
- **China** (Xinxiang) • **GoldenEye 20A Kit** (19 autosomal STRs, amelogenin)
  - Huang, Y.M., et al. (2013). Assessment of application value of 19 autosomal short tandem repeat loci of GoldenEye 20A kit in forensic paternity testing. *International Journal of Legal Medicine, 127(3),* 587-590.
- **China** (Beijing) • **Rapid 21-plex System** (20 autosomal STRs, amelogenin)
  - Yang, M., et al. (2016). Development of a rapid 21-plex autosomal STR typing system for forensic applications. *Electrophoresis, 37,* 2789-2799.
- **China** (Shanghai) • **Expressmarker 16+10Y & 16+18Y Kits** (15 autosomal STRs, 10/18 Y-STRs, amelogenin)
  - Zhou, H., et al. (2016). Developmental validation of forensic DNA-STR kits: Expressmarker 16+10Y and Expressmarker 16+18Y. *Forensic Science International: Genetics, 24,* 1-17.
- **China** (Xi'an) • **AGCU 21+1 STR Kit** (21 non-core autosomal STRs, amelogenin)
  - Zhu, B.F., et al. (2015). Developmental validation of the AGCU 21+1 STR kit: a novel multiplex assay for forensic application. *Electrophoresis, 36,* 271-276.
- **China** (Shantou) • **SureID PanGlobal System** (24 autosomal STRs, 2 Y markers, amelogenin)
  - Liu, Y., et al. (2017). Developmental validation of a 6-dye typing system with 27 loci and application in Han population of China. *Scientific Reports, 7,* 4706.

Recent Articles on STR Markers

- Gettings, K.B., et al. (2015). STR allele sequence variation: current knowledge and future issues. *Forensic Science International: Genetics, 18,* 118-130.
- Parson, W., et al. (2016). Massively parallel sequencing of forensic STRs: considerations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. *Forensic Science International: Genetics, 22,* 54-63.
- C. Phillips (2017) A genomic audit of newly-adopted autosomal STRs for forensic identification. *Forensic Sci. Int. Genet.* 29: 193-204.

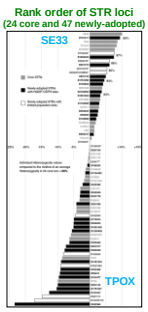
See other references in workshop reference list

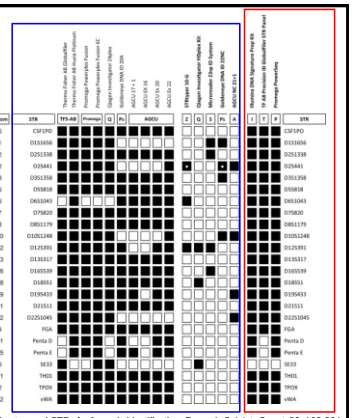
Multiplex Combinations of 41 Newly-Adopted STRs



C. Phillips (2017) A genomic audit of newly-adopted autosomal STRs for forensic identification. *Forensic Sci. Int. Genet.* 29: 193-204.

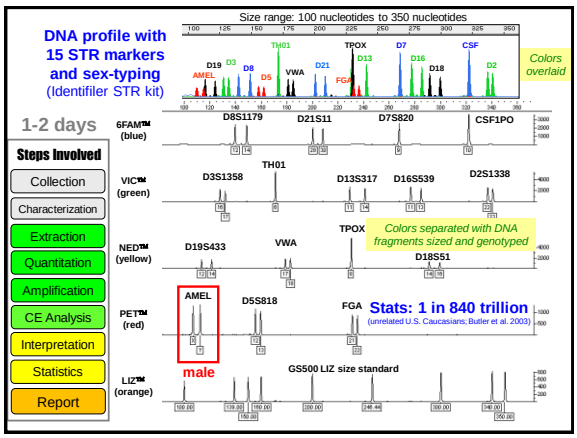
STR markers present in kits for **CE** and **MPS**

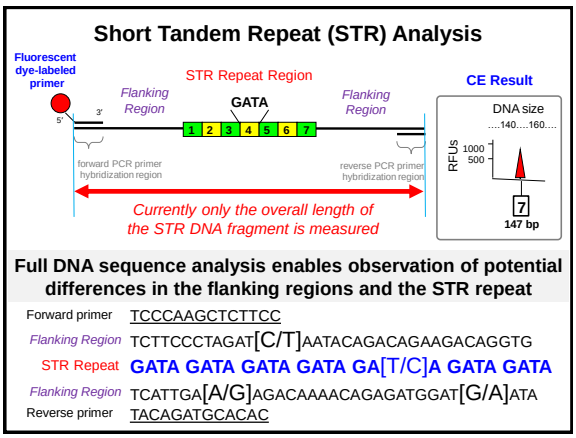




C. Phillips (2017) A genomic audit of newly-adopted autosomal STRs for forensic identification. *Forensic Sci. Int. Genet.* 29: 193-204.

STR Measurement Systems





STR Measurement Techniques

- **Length-based** measurements
  - Gels
  - Capillary electrophoresis
  - Mass spectrometry
- **Sequence-based** measurements
  - Sanger sequencing
  - Massively parallel sequencing (MPS or NGS)

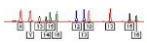
J.M. Butler (2011) *Advanced Topics in Forensic DNA Typing: Methodology*. Table 6.1

Genetic Analyzers from Applied Biosystems

ABI Genetic Analyzer	Years Released for Human ID	Number of Capillaries	Laser	Polymer delivery	Other features
373 (gel system)	1992-2003	--	40 mW Ar+ (488/514 nm)	--	PMTs and color filter wheel for detection
377 (gel system)	1995-2006	--	40 mW Ar+ (488/514 nm)	--	CCD camera
310	1995-	1	10 mW Ar+ (488/514 nm)	syringe	Mac operating system & Windows NT (later)
3100	2000-2005	16	25 mW Ar+ (488/514 nm)	syringe	
3100-Avant	2002-2007	4	25 mW Ar+ (488/514 nm)	syringe	
3130	2003-2011	4	25 mW Ar+ (488/514 nm)	pump	
3130xl	2003-2011	16	25 mW Ar+ (488/514 nm)	pump	
3500	2010-	8	10-25 mW diode (505 nm)	new pump	110V power; **RFID-tagged reagents**; **.hid files**; normalization & 6-dye detection possible
3500xl	2010-	24			
3700	2002-2003	96	25 mW Ar+ (488/514 nm)	cuvette-based	Split beam technology
3730	2005-	48	25 mW Ar+ (488/514 nm)	pump	
3730xl	2005-	96	25 mW Ar+ (488/514 nm)	pump	

Information courtesy of Michelle S. Shepherd, Applied Biosystems, LIFE Technologies.

New Instruments and Approaches Becoming **Available** for the Forensic Genetics Community



- The ABI 310, 3100, 3130, 3500 Genetic Analyzers have been the only capillary electrophoresis (CE) instrument available for >20 years
- It remains to be seen how the community will be impacted by the introduction of new CE systems, rapid DNA, and MPS/NGS sequencing

RapidHIT ID System

Ion S5 MPS Sequencer

Length vs Sequence Measurements

- **STR typing**
  - Established use for 20 years
  - Currently only length-based
- **Sequence analysis**
  - Under development
  - Provides **additional information** regarding STR alleles
  - Provides **additional capabilities** (such as determining eye color, ancestry, separating identical twins, etc.)

14

STR length = 14 repeats

sequence analysis

$[TCTA]_1[TCTG]_1[TCTA]_{12}$

$[TCTA]_1[TCTG]_2[TCTA]_{11}$

allele sequence diversity exists
(but is not measured if we only examine the overall length)

Challenges with sequence information
- How to describe and name the sequenced STR alleles
- How to store and analyze large amounts of data
- How to address privacy concerns with gaining more genomic information

Guidance on STR Analysis

Types of Standards

physical (measurement) standards

Certified reference material to aid with calibration of measurements

http://www.nist.gov/srm/

documentary (technical) standards

Specific requirements for the operation of a laboratory related to management system and competence

Current Hierarchy of Standards
for Accrediting Bodies to Use
in Auditing U.S. Forensic DNA Laboratories

International Laboratory Accreditation Cooperation (ILAC)
G19:08/2014 Modules in a Forensic Science Process

2017

ISO/IEC 17025:2005 General requirements for the competence of testing and calibration laboratories

2017

The FBI Quality Assurance Standards (2011) serve as supplemental materials to ISO/IEC 17025 for DNA audits

2017 2016

SWGDAM guidelines (interpretation, validation, etc.) provide further information but are not audited against

FBI **Quality Assurance Standards** (QAS)

- **DNA Identification Act of 1994**
  - Requires FBI Laboratory, those labs receiving federal funds, and those labs using the National DNA Index System (NDIS) to comply
- FBI Laboratory's DNA Advisory Board (**DAB**)
  - Met from 1995 to 2000 to discuss and draft QAS
  - FBI Director issued initial QAS in October 1998 (caseworking) and April 1999 (databasing)
- Scientific Working Group on DNA Analysis Methods (**SWGDAM**)
  - assumed responsibility for QAS revisions when DAB was dissolved
  - QAS revisions released in July 2009 and September 2011
- **QAS audit documents are used by accrediting bodies** such as ASCLD/LAB in audits of DNA laboratories as **supplemental material to the ISO/IEC 17025 standard**

The FBI Quality Assurance Standards

Begun in 1998/99 with updates via SWGDAM in 2009, 2011, 2014, 2017, …


QUALITY ASSURANCE STANDARDS FOR DNA DATABASING LABORATORIES

QUALITY ASSURANCE STANDARDS FOR FORENSIC DNA TESTING LABORATORIES

This document consists of definitions and standards. The standards are quality assurance measures that place specific requirements on the laboratory. Equivalent measures not outlined in this document may also meet the standard if determined sufficient through an accreditation process.

EFFECTIVE DATE:

These standards shall take effect September 1, 2011.

REFERENCES: Federal Bureau of Investigation, "Quality Assurance Standards for Forensic DNA Testing Laboratories" and "Quality Assurance Standards for Convicted Offender DNA Databasing Laboratories," Forensic Science Communications, July 2000, Volume 2, Number 3.

THE FBI QUALITY ASSURANCE STANDARDS

THE FBI QUALITY ASSURANCE STANDARDS

AUDIT FOR

FORENSIC DNA TESTING LABORATORIES

IN ACCORDANCE WITH

THE QUALITY ASSURANCE STANDARDS

FOR

FORENSIC DNA TESTING LABORATORIES

EFFECTIVE SEPTEMBER 1, 2011

1. SCOPE
2. DEFINITIONS
3. QUALITY ASSURANCE PROGRAM
4. ORGANIZATION AND MANAGEMENT
5. PERSONNEL
6. FACILITIES
7. EVIDENCE *(SAMPLE)* CONTROL
8. VALIDATION
9. ANALYTICAL PROCEDURES
10. EQUIPMENT CALIBRATION AND MAINTENANCE
11. REPORTS
12. REVIEW
13. PROFICIENCY TESTING
14. CORRECTIVE ACTION
15. AUDITS
16. SAFETY
17. OUTSOURCING

http://www.fbi.gov/about-us/lab/biometric-analysis/codis/qas-standards-for-forensic-dna-testing-laboratories-effective-9-1-2011
http://www.fbi.gov/about-us/lab/biometric-analysis/codis/qas-standards-for-dna-databasing-laboratories-effective-9-1-2011

Scientific Working Group on DNA Analysis Methods (SWGDAM)

http://www.swgdam.org/

- **Established in November 1988 by FBI Laboratory**
  - Named Technical Working Group on DNA Analysis Methods (TWGDAM) for the first decade
- **Comprised of ~50 scientists from U.S. and Canada**
  - Typically 20-25 voting members and the rest as invited guests
- European Network of Forensic Science Institutes (ENFSI) DNA Working Group representative often attends
- **Three day meetings held semiannually every January and July**
- **Current committees (6) and working groups (4):**
  - Autosomal STR Interpretation, Combined DNA Index System, Quality Assurance, Rapid DNA, Lineage Marker, Laboratory Operations, Next Generation Sequencing, Verbal Equivalent, Forensic Serology (coming), and Contextual Bias (coming)
- Previous committees:
  - RFLP, PCR, mitochondrial DNA, Y-STR, mass spectrometry, training, validation, expert systems, missing persons/mass disasters, mixture interpretation, enhanced methods and interpretation, probabilistic genotyping

Current SWGDAM Guidelines

Hyperlinks to documents available on SWGDAM.org

Release Date	Guidelines	Previous Versions (TWGDAM)
2017	**STR Interpretation** Guidelines (90 pages)	2000, 2010
2017	Contamination Prevention and Detection Guidelines (29 pages)	--
2016	**Validation** Guidelines for Forensic DNA Analysis Methods (15 pages)	1991, 1995, 2004, 2012
2015	Validation of Probabilistic Genotyping Systems (12 pages)	--
2015	Collection and Serological Examination of Biological Evidence (19 pages)	--
2014	Guidelines for **Missing Persons Casework** (28 pages)	--
2014	Interpretation Guidelines for **Y-Chromosome STRs** (20 pages)	2009
2014	STR Enhanced Detection Methods (22 pages)	--
2013	**Training** Guidelines (30 pages)	2001
2013	**Mitochondrial DNA Analysis Interpretation** Guidelines (23 pages) & Mitochondrial DNA Nomenclature Examples (5 pages)	1993, 2003

Public Comments Can Now Be Made on Draft SWGDAM Documents


PUBLIC COMMENTS PAGE

In accordance with the SWGDAM Bylaws (Section V.C.5), SWGDAM will make any new or revised guidance or standard document(s) available for public comment for a minimum of 30 days. Generally, SWGDAM attempts to review its guidance documents within 5 years of their issuance and is usually actively revising at least one of its guidance documents at any given time. SWGDAM strongly encourages the forensic DNA community or other interested group to comment on the SWGDAM documents currently in this stage of development. Upon receipt, these comments will be forwarded to the appropriate SWGDAM Committee for consideration and may be incorporated into the final document considered for approval by the SWGDAM Membership.

Alternately, SWGDAM may publish a response to a specific suggestion or recommendation on its FAQ Page for general information purposes. SWGDAM will make all reasonable efforts to advise the forensic DNA community of these documents currently available for public comment. SWGDAM strongly encourages all interested parties to regularly monitor SWGDAM.org for the guidance document(s) or standard document(s) currently available for public comment. Please use the contact portal below for providing comments on the SWGDAM document(s) available for public comment.

SWGDAM Documents Available for Public Comment

The following guidance or standards document(s) is/are currently available for public comment until September 22, 2017:

QUALITY ASSURANCE STANDARDS FOR FORENSIC DNA TESTING LABORATORIES

Details: Following completion of a multi-year review, SWGDAM is recommending changes to the Forensic QAS.

http://swgdam.org/public_review.html

Thank you for your attention

Acknowledgments:
Peter Vallone

Contact Information

John M. Butler
NIST Fellow & Special Assistant
to the Director for Forensic Science
Special Programs Office
john.butler@nist.gov
+1-301-975-4049

Lisa Borsuk
Research Scientist, Bioinformatics
Applied Genetics Group
lisa.borsuk@nist.gov
+1-301-975-5405

Katherine Gettings
Research Biologist
Applied Genetics Group
katherine.gettings@nist.gov
+1-301-975-6401

A copy of this presentation is available at:
http://strbase.nist.gov/NISTpub.htm

Questions?

29 August 2017

Workshop #10

STR Markers

Lisa Borsuk, M.S.
Applied Genetics Group
U.S. National Institute of Standards and Technology

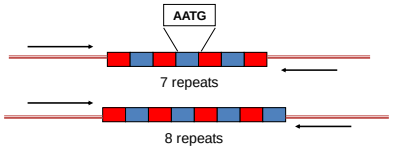
Product Disclaimer

- **We will mention commercial STR kit names and information, but we are in no way attempting to endorse any specific products.**
- **NIST Disclaimer**: Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose. All work presented has been reviewed and approved by the NIST Human Subjects Protections Office.
- **Points of view are the speakers** and do not necessarily represent the official position of the National Institute of Standards and Technology or the U.S. Department of Justice. **The NIST Applied Genetics Group receives or has received funding from the FBI Laboratory and the National Institute of Justice.**

Outline

- Overview of STRs
  - General information
  - Core Autosomal Sets of STRs
- CE and NGS kits
  - Examples of 6 and 8 dye kits
  - Examples of NGS kits
- Example – D1S1656

Short Tandem Repeats (STRs)



the repeat region is variable between samples while the flanking regions where PCR primers bind are constant

Homozygote = both alleles are the same length

Heterozygote = alleles differ and can be resolved from one another

STRs 101 intro.pptx from STRBase.NIST.gov

General Information

- Desirable Features for STRs
  - High heterozygosity
  - Regular repeat unit
    - 3-6 bases in length
    - 4 base repeats are most common
  - Distinguishable alleles
  - Robust amplification
- Types of STRs
  - Simple – one repeat sequence
  - Compound – two or more repeat sequences
  - Complex repeats
  - Hypervariable repeats

STRs 101 STRBase.NIST.gov

Core Autosomal STR Loci

Listed are the primary Autosomal STRs available in kits D6S1043, Penta E, and Penta D are additional STRs present in many kits

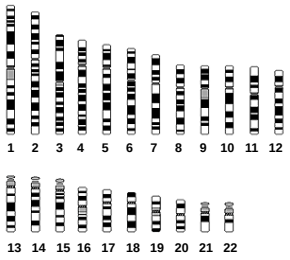
Chr	STR Locus	Repeat		US Core Loci 20	European Standard Set 12	ESS Additional Loci	Interpol Standard Set
1q42	D1S1656	[CCTA]a [TCTA]b	Compound	*	*		
2p25.3	TPOX	[AATG]a	Simple	*			
2p14	D2S441	[TCTA]a		*	*		
2q35	D2S1338	[GGAA]a [GGCA]b		*		*	
3p21.31	D3S1358	TCTA [TCTG]a [TCTA]b		*	*		*
4q31.3	FGA	[GGAA]a GGAG [AAAG]b AGAA AAAA [GAAA]c	Complex	*	*		*
5q23.2	D5S818	[ATCT]a		*			
5q33.1	CSF1PO	[ATCT]a		*			
6q14	SE33	[CTTT]a [TT]0-1 [CT]b [CTTT]c	Hypervariable			*	*
7q21.11	D7S820	[TATC]a		*			
8q24.13	D8S1179	[TCTA]a [TCTG]0-2 [TCTA]b		*	*		*
10q26.3	D10S1248	[GGAA]a		*	*		
11p15.5	TH01	[AATG]a ATG 0-1 [AATG]b		*	*		*
12p13.31	vWA	TAGA TGGA [TAGA]a [CAGA]b [TAGA]c		*	*		*
12q13.2	D12S391	[AGAT]a [AGAC]b AGAT 0-1		*	*		
13q31.1	D13S317	[TATC]a		*			
16q24.1	D16S539	[GATA]a		*		*	
18q21.33	D18S51	[AGAA]a		*	*		*
19q12	D19S433	[CCTT]a CCTA [CCTT]b CTTT [CCTT]c		*		*	
21q21.1	D21S11	[TCTA]a [TCTG]b [TCTA]c TA [TCTA]d TCA [TCTA]e TCCATA [TCTA]f		*	*		*
22q12.3	D22S1045	[ATT]a ACT [ATT]2		*	*		
Xp, Yp	Amelogenin	--		*		*	*

STRs

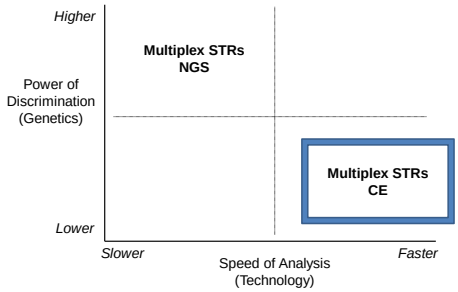
Additional Autosomal STRs
D4S2408 D17S1301
D5S2800 **~55 +**

Y STRs
DYS391
DYS385 a/b
Y-GATA-H4
~110 +

X STRs
DXS10101
DXS7423
HPRTB
~20 +

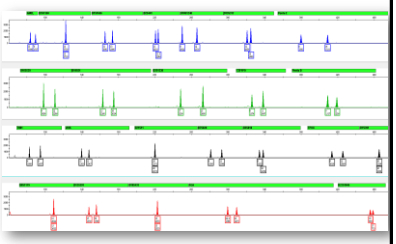


DNA Typing Technologies

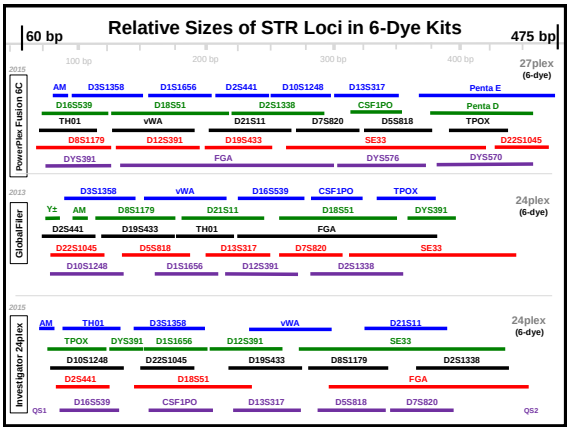


Capillary Electrophoresis

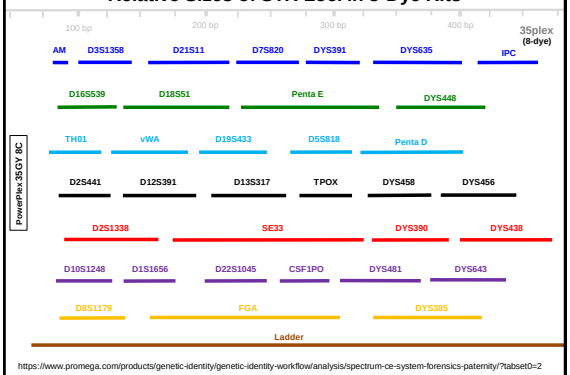
- **Length based** STR results
- Fast turn around time from collection to results
- Accurate reproducible results
- Multiplexed Kits
  - ~20 Loci plus



Relative Sizes of STR Loci in 6-Dye Kits

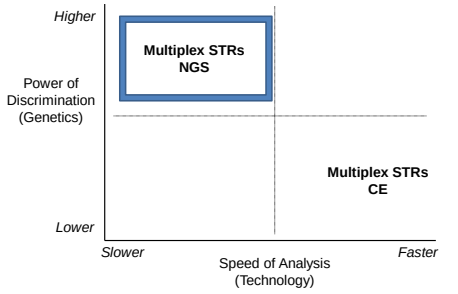


Relative Sizes of STR Loci in 8-Dye Kits



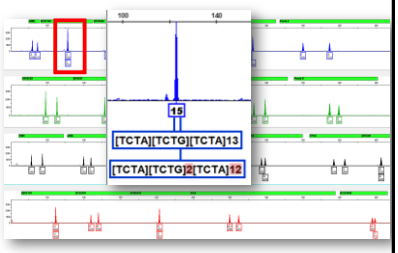
https://www.promega.com/products/genetic-identity/genetic-identity-workflow/analysis/spectrum-ce-system-forensics-paternity/?tabset0=2

DNA Typing Technologies

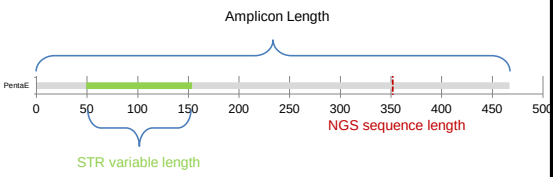


Capillary Electrophoresis

- **Length based** STR results
- Fast turn around time from collection to results
- Accurate reproducible results
- Multiplexed Kits
  - ~20 Loci plus
- No Sequence Information

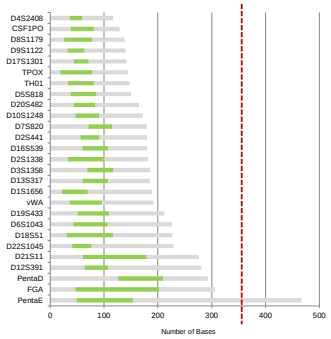


Example NGS Range



Illumina ForenSeq (Set B) – STR Amplicon Sizes

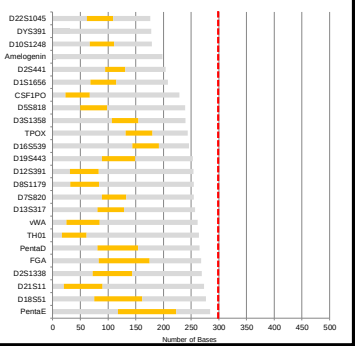
- 58 STRs plus Amelogenin
  - 27 – Autosomal STRs
  - 7 – X STRs
  - 24 – Y STRs
- 172 SNPs
  - 94 – Identity
  - 56 – Ancestry
  - 22 – Phenotype



Based on ForenSeq Manual Sept2015

Promega PowerSeq – STR Amplicon Sizes

- 45 STRs plus Amelogenin
  - 22 Autosomal STRs
  - 23 Y STRs
  - Additional mitochondrial targeted sequencing



ThermoFisher Precision ID

- GlobalFiler NGS STR Panel
  - 30 STRs plus Amelogenin
    - 29 Autosomal STRs
    - 1 Y STR

Additional Panels

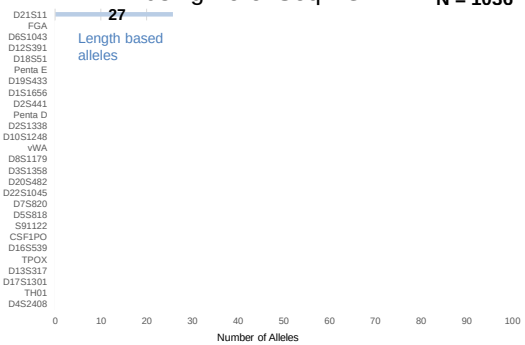
- Ancestry Panel
  - 165 SNPs
- Identity Panel
  - 124 SNPs
- mtDNA Whole Genome Panel
- mtDNA Control Region Panel

Locus	Source	Repeat Structure	Type	Chromosome
AMEL-X	GF	Indel	Sex determination	X
AMEL-Y	GF	Indel	Sex determination	Y
TPOX	GF	AATG	CODIS	2
D8S1358	GF	TCTA/TCTG	CODIS	3
FGA	GF	CTTT/TTCC	CODIS	4
CSF1PO	GF	AGAT	CODIS	5
D5S818	GF	AGAT	CODIS	5
D7S820	GF	GATA	CODIS	7
D8S1179	GF	TCTA/TCTG	CODIS	8
TH01	GF	TCAT	CODIS	11
vWA	GF	TCTA/TCTG	CODIS	12
D13S317	GF	TATC	CODIS	13
D16S539	GF	GATA	CODIS	16
D18S51	GF	AGAA	CODIS	18
D21S11	GF	TCTA/TCTG	CODIS	21
D1S1656	GF	TAGA	Expanded CODIS	1
D2S1338	GF	TGCC/TTCC	Expanded CODIS	2
D2S441	GF	TCTA/TCAA	Expanded CODIS	2
D10S1248	GF	GGAA	Expanded CODIS	10
D12S391	GF	AGAT/AGAC	Expanded CODIS	12
D19S433	GF	AAGG/TAGG	Expanded CODIS	19
D22S1045	GF	ATT	Expanded CODIS	22
DYS391	GF	TCTA	Expanded CODIS	Y
D6S1043	Ohira	AGAT/AGAC	MPS	6
D6S1800	MPS	GATA/GATC	MPS	6
D6S474	MPS	GATA/GACA	MPS	6
D12ATA63	MPS	TAA/CAA	MPS	12
D14S1434	MPS	CTGT/CTAT	MPS	14
D15S1677	MPS	TTCC	MPS	15
D15S1776	MPS	AGAT	MPS	2
D3S4529	MPS	ATCT	MPS	3
D4S2408	MPS	ATCT	MPS	4

https://www.thermofisher.com/order/catalog/product/A30939

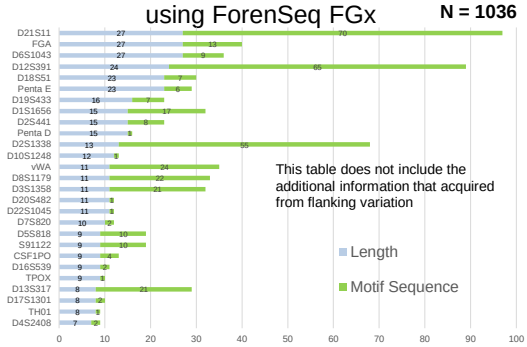
Autosomal STRs – NIST Population Data using ForenSeq FGx

N = 1036

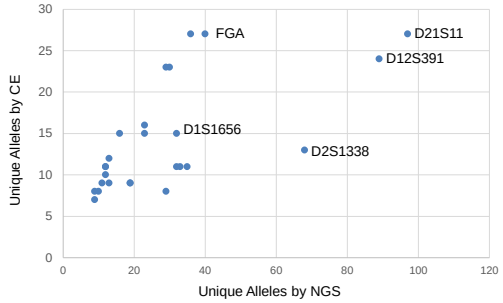


Autosomal STRs – NIST Population Data using ForenSeq FGx

N = 1036



Numbers of Unique Alleles for CE vs. NGS for each autosomal locus in ForenSeq



NIST 1036 Population Samples

Resources

- STRBase
  - STRBase.NIST.gov
- Journal article searches
  - Pubmed - www.ncbi.nlm.nih.gov
- Books
  - Advanced Topics in Forensic DNA Typing: Methodology – J.M. Butler 2012

D1S1656 – General Information

- Chromosome location
  - Chr1 : 23,076,906 - 23,076,983 (GRCh38)
- Repeat – note that the [ca]5 is not counted as part of the repeat
  - [ca]5 [TCTA]n
  - Other observed repeat patterns
    - [ca]5 CCTA [TCTA]n
    - [ca]5 CTTA [TCTA]n
    - [ca]5 CCTA [TCTA]n TCA [TCTA]n
- Standard Sets
  - USA (CODIS core-loci)
  - ENFSI/European Union (ESS)
  - Interpol

D1S1656 – CE Kits

- Thermo Fisher
  - NGM
  - NGM SElect/Detect (Express)
  - Globalfiler (Express)
  - Verifiler
- Promega
  - PowerPlex ESX 16 (Fast)
  - PowerPlex ESX 17 (Fast)
  - PowerPlex ESI 16 (Fast)
  - PowerPlex ESI 17 (Fast)
  - PowerPlex 21
  - PowerPlex Fusion
  - PowerPlex Fusion 6C
  - PowerPlex 35GY 8C
- Qiagen
  - Investigator ESSplex Plus Kit
  - Investigator ESSplex SE Plus/GO!/QS
  - Investigator Nonaplex ESS Kit
  - Investigator Hexaplex ESS Kit
  - Investigator 24plex
- Gordiz
  - COrDIS Plus
- Peoplespot
  - Goldeneye DNA ID 22NC

D1S1656 – CE Kit Ladders

Promega	ESI 17 Pro	9	10	11	12	13	14	14.3	15	15.3	16	16.3	17	17.3	18	18.3	19	19.3	20.3
Promega	ESX 16/17	9	10	11	12	13	14	14.3	15	15.3	16	16.3	17	17.3	18	18.3	19	19.3	20.3
Promega	Fusion	9	10	11	12	13	14	14.3	15	15.3	16	16.3	17	17.3	18	18.3	19	19.3	20.3
ThermoFisher	NGMSElect	9	10	11	12	13	14	14.3	15	15.3	16	16.3	17	17.3		18.3		19.3	20.3
ThermoFisher	GlobalFiler	9	10	11	12	13	14	14.3	15	15.3	16	16.3	17	17.3		18.3		19.3	20.3
Qiagen	ESSplex		10	11	12	13	14		15		16		17	17.3		18.3		19.3	
Qiagen	24plex		10	11	12	13	14	14.3	15	15.3	16	16.3	17	17.3	18	18.3		19.3	20.3

Alleles present in the ladders of some example kits
Colors represent Dyes used by the kits

D1S1656 – NGS Kits

- Illumina
  - ForenSeq Sets A and B
- Promega
  - PowerSeq Auto
- Thermo Fisher
  - GlobalFiler NGS STR Panel

D1S1656 – Observed Lengths and Sequences

Allele	Bracket	Reference	Platform
7	[ca]5[TCTA]7[]	publication	FGx/MiSeq
8	[ca]5[TCTA]8	Phillips et al. (2011)	Sanger
9	[ca]5[TCTA]9	Phillips et al. (2011)	Sanger
10	[ca]5[TCTA]10	Lareu et al. (1998)	Sanger
10	[ca]5[CCTA][TCTA]9	Phillips et al. (2011)	Sanger
10.3		Variant Alleles/STRBase	CE
11	[ca]5[TCTA]11	Lareu et al. (1998)	Sanger
11	[ca]5[CCTA][TCTA]10	Lareu et al. (1998)	Sanger
11.1		Variant Alleles/STRBase	CE
11.3		Variant Alleles/STRBase	CE
...	...[82 rows] removed[]...	...	...
17	[ca]5[CTTA][TCTA]16	Gettings et al. (2017)	FGx/MiSeq
17	[ca]5[TCTA]17	Novroski et al. (2016)	FGx/MiSeq
17	[ca]5[CCTA][TCTA]16	Lareu et al. (1998)	Sanger
17.1	[ca]6[TCTA]12[TCA][TCTA]4	Novroski et al. (2016)	FGx/MiSeq
17.2		Variant Alleles/STRBase	CE
17.3	[ca]5[CCTA][TCTA]12[TCA][TCTG][TCTA]3	Gettings et al. (2017)	FGx/MiSeq
17.3	[ca]5[TCTA]13[TCA][TCTA]4	Gettings et al. (2017)	FGx/MiSeq
17.3	[ca]5[CCTA][TCTA]12[TCA][TCTA]4	Lareu et al. (1998)	Sanger
18	[ca]5[CCTA][TCTA]17	Phillips et al. (2011)	Sanger
18.1		Variant Alleles/STRBase	CE
18.2		Variant Alleles/STRBase	CE
18.3	[ca]5[CCTA][TCTA]13[TCA][TCTA]4	Lareu et al. (1998)	Sanger
19	[ca]5[CCTA][TCTA]18	Novroski et al. (2016)	FGx/MiSeq
19.3	[ca]5[CCTA][TCTA]14[TCA][TCTA]4	Lareu et al. (1998)	Sanger
20		STRBase/bot/KGRP	CE
20.3	[ca]5[CCTA][TCTA]15[TCA][TCTA]4[]	KCL	FGx/MiSeq
21		STRBase/bot/KGRP	CE
21.3		Variant Alleles/STRBase	CE

Conclusions

- More information is being accumulated for STRs important to the forensic community
- More STRs are being explored and incorporated into routine analysis
- There is a lot of data out there and more is coming

29 August 2017

Workshop #10

Interpretation Issues

John M. Butler, Ph.D.
Special Programs Office
U.S. National Institute of Standards and Technology

Steps in Forensic DNA Analysis

Gathering the Data

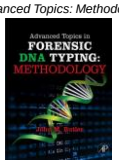
Understanding Results Obtained & Sharing Them

Collection/Storage/ Characterization → Extraction/ Quantitation → Amplification/ Marker Sets → Separation/ Detection → Data → Stats → Report

Interpretation

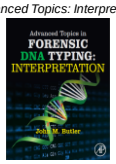
>1300 pages of information with >5000 references cited in these two books

Advanced Topics: Methodology



August 2011

Advanced Topics: Interpretation



October 2014

Ian Evett on Interpretation

"The crucial element that the scientist brings to any case is the *interpretation* of those observations. This is the heart of forensic science: it is where the scientist adds value to the process."

Evett, I.W., et al. (2000). The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice, 40,* 233-239.

Critical Challenges Faced Today

- **Success of DNA testing** → significant growth in sample submissions → sample backlogs
  - Laboratory automation and expert system data review
  - Restrictive case acceptance policies to avoid law enforcement investigator 'swab-athons' at crime scenes
- **Greater detection sensitivity** → more complex DNA mixtures and low-template DNA with 'touch' evidence
  - Probabilistic genotyping to cope with increase in data interpretation uncertainty
  - Use of a complexity threshold to avoid "skating on thin ice"

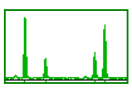
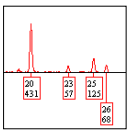
Butler, J.M. (2015) The future of forensic DNA analysis. *Phil. Trans. R. Soc.* B 370: 20140252

5 Reasons that DNA Results Are Becoming More Challenging to Interpret

1. **More sensitive DNA test results**
2. **More touch evidence samples** that are poor-quality, low-template, complex mixtures
3. **More options exist** for statistical approaches involving probabilistic genotyping software
4. **Many laboratories are not prepared** to cope with complex mixtures
5. **More loci being added** because of the large number of samples in DNA databases

http://www.cstl.nist.gov/strbase/pub_pres/Butler-DNA-interpretation-AAFS2015.pdf

Math Analogy to DNA Evidence

$2 + 2 = 4$	$2x^2 + x = 10$	$\int_{x=0}^{\infty} f(x)dx$
Basic Arithmetic	**Algebra**	**Calculus**
		
Single-Source DNA Profile (DNA databasing)	**Sexual Assault Evidence** (2-person mixture with high-levels of DNA)	**Touch Evidence** (>2-person, low-level, complex mixtures perhaps involving relatives)

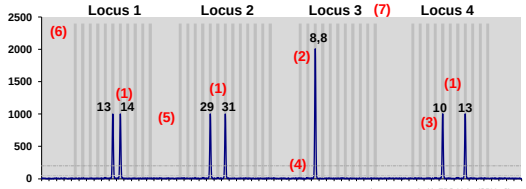
http://www.cstl.nist.gov/strbase/pub_pres/Butler-DNA-interpretation-AAFS2015.pdf

Many laboratories are not prepared
to cope with complex mixtures

- Have **appropriate validation studies** been performed to inform proper interpretation protocols? (curriculum & classroom instruction)
- Are **appropriately challenging proficiency tests** being given? (graded homework assignments)
- **Would we want to go into a calculus exam only having studied algebra and having completed homework assignments involving basic arithmetic?**

http://www.cstl.nist.gov/strbase/pub_pres/Butler-DNA-interpretation-AAFS2015.pdf

Using **Ideal Data** to Discuss Principles

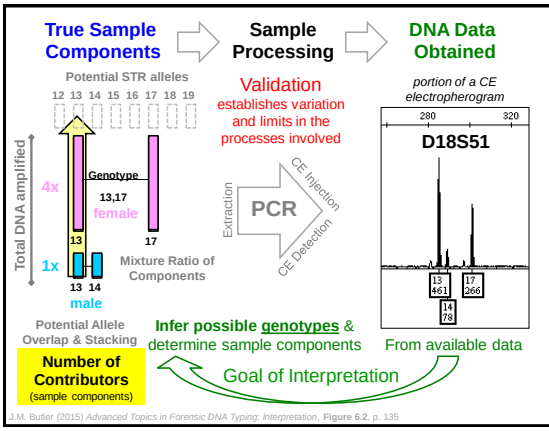


(1) 100% PHR (Hb) between heterozygous alleles
(2) Homozygotes are exactly twice heterozygotes due to allele sharing
(3) No peak height differences exist due to size spread in alleles (any combination of resolvable alleles produces 100% PHR)
(4) No stutter artifacts enabling mixture detection at low contributor amounts
(5) Perfect inter-locus balance
(6) Completely repeatable peak heights from injection to injection on the same or other CE instruments in the lab or other labs
(7) *Genetic markers that are so polymorphic all profiles are fully heterozygous with distinguishable alleles enabling better mixture detection and interpretation*

J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Figure 1.5**, p. 11

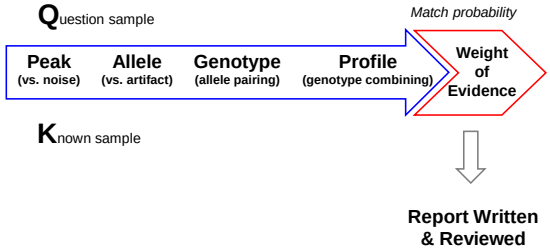
Challenges in Real-World Data

- **Stochastic (random) variation** in sampling each allele during the PCR amplification process
  - This is highly affected by DNA quantity and quality
  - Imbalance in allele sampling gets worse with low amounts of DNA template and higher numbers of contributors
- **Degraded DNA** template may make some allele targets unavailable
- **PCR inhibitors** present in the sample may reduce PCR amplification efficiency for some alleles and/or loci
- **Overlap of alleles** from contributors **in DNA mixtures**
  - Stutter products can mask true alleles from a minor contributor
  - Allele stacking may not be fully proportional to contributor contribution

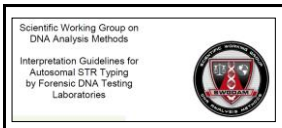


J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Figure 6.2**, p. 135

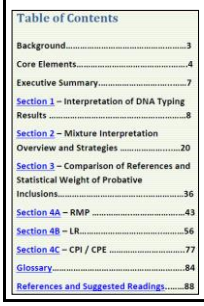
Steps in DNA Interpretation



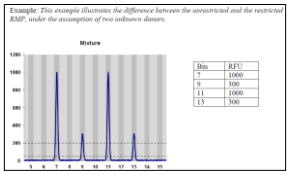
J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Figure 1.3**, p. 7



2017 SWGDAM Guidelines



- 90 pages in length
- Provides worked examples with different statistical approaches



Available at https://www.swgdam.org/publications

Know the Limits of What You Can Do

Butler, J.M. (2015) *Advanced Topics in Forensic DNA Typing: Interpretation* (Elsevier Academic Press: San Diego), pp. 159-182

CHAPTER
7

Low-Level DNA and Complex Mixtures

"The limits of each DNA typing procedure should be understood, especially when the DNA sample is small, is a mixture of DNA from multiple sources, or is contaminated with interfering substances."
NRC I, 1992, p. 8

"For the complex DNA profile, there is no predominant or overarching standard interpretation method."
Peter Gill (Gill et al. 2012, report to the UK Forensic Science Regulator, p. 18)

"The limits of each DNA typing procedure should be understood, especially when the DNA sample is small, is a mixture of DNA from multiple sources…" (NRC I, 1992, p. 8)

ABI Genetic Analyzer Data Collection

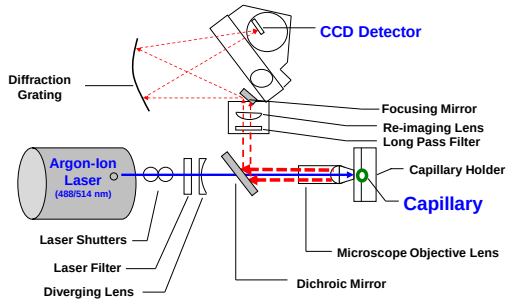
Advanced Topics in Forensic DNA Typing: Interpretation, Chapter 2

Advanced Topics in Forensic DNA Typing: Methodology, Chapter 6

Key Points on Data Collection

- **On-scale data** of STR allele peaks are important to interpretation (both lower and upper limits exist for reliable data)
- Data signals from ABI Genetic Analyzers are processed by **proprietary algorithms** that include variable binning (adjustment for less sensitive fluorescent dyes), baselining, smoothing, and multi-componenting for separating color channels
- **Instrument sensitivities vary** due to different lasers, detectors, and optical alignment (remember that signal strength is in "relative fluorescence units", RFUs)

Optics for ABI 310 Genetic Analyzer

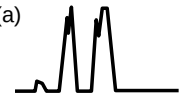


J.M. Butler (2012) *Advanced Topics in Forensic DNA Typing: Methodology*, Figure 6.6

STR Typing Works Best in a Narrow Window of DNA Template Amounts

Off-scale data with flat-topped peaks

(a)



Too much DNA amplified

Or injected onto CE

Allele dropout due to stochastic effects

(b)

Too little DNA amplified

"Just right"

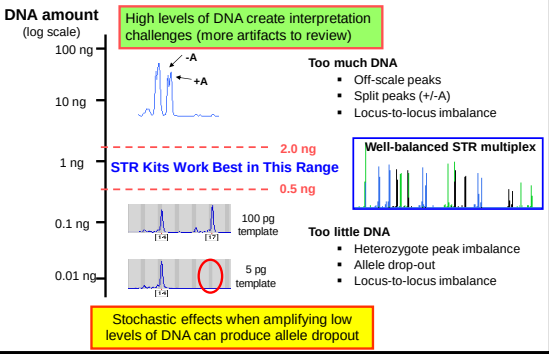
(c)

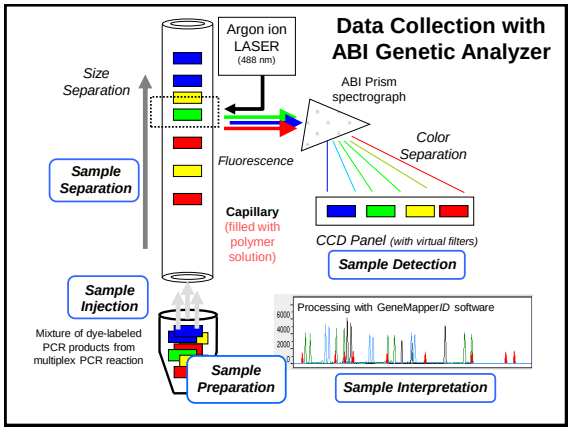
Within optimal range

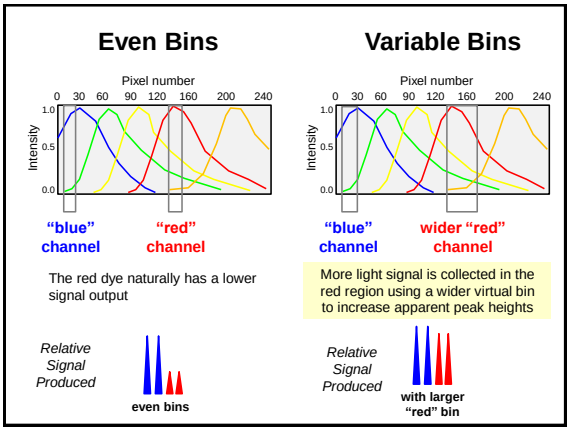
Typically best results are seen in the 0.5 ng to 1.5 ng range for most STR kits

Impact of DNA Amount into Multiplex PCR Reaction

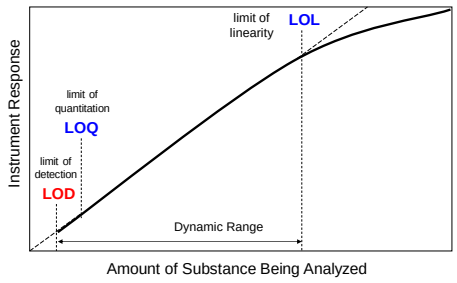
We generally aim for 0.5-2 ng







Useful Range of an Analytical Method



J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Figure 2.3**, p. 31

Impact of Setting Thresholds Too High or Too Low

If	Then
Threshold is set too high…	Analysis may miss low-level legitimate peaks (**false negative conclusions produced**)
Threshold is set too low…	Analysis will take longer as artifacts and baseline noise must be removed from consideration as true peaks during data review (**false positive conclusions produced**)

J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Table 2.3**, p. 44

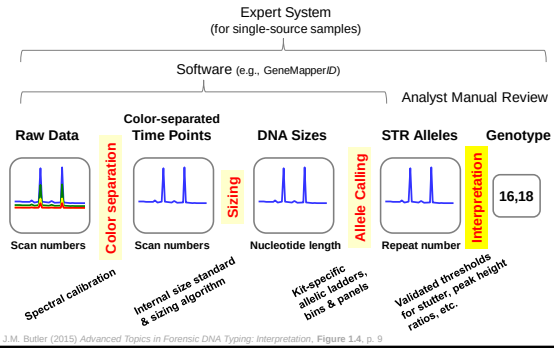
STR Alleles and PCR Amplification Artifacts

Advanced Topics in Forensic DNA Typing: Interpretation, Chapter 3

Key Points with STR Alleles

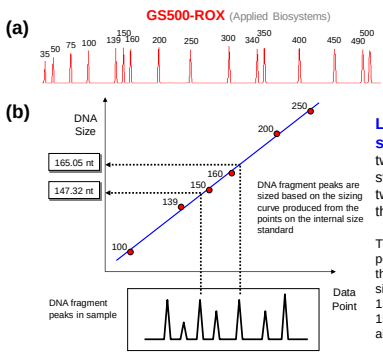
- STR allele designations are made by comparing the relative size of sample peaks to allelic ladder allele sizes
- A common, calibrated STR allele nomenclature is essential in order to compare data among laboratories
- STR allele sizes are based on a measure of the relative electrophoretic mobility of amplified PCR products (defined by primer positions) compared to an internal size standard using a specific sizing algorithm
- STR alleles can vary in their overall length (number of repeat units), with their internal sequence of repeats, and in the flanking region

Transformation of Information at a Single STR Locus during Data Processing



J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Figure 1.4**, p. 9

DNA Size Standard and Sizing Algorithm



Local Southern sizing algorithm uses two peaks (from the size standard) above and two peaks below below the unknown peak

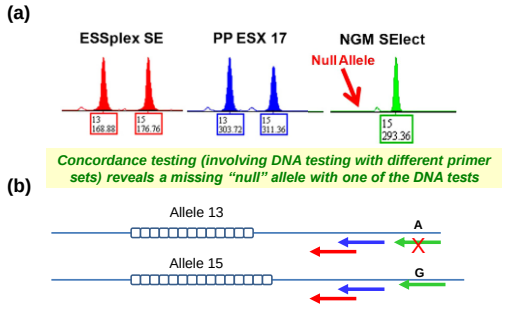
The 147.32 nt example peak would be defined by the relative positions of the size standard 100 nt and 139 nt peaks below and the 150 nt and 160 nt peaks above

J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Box 3.1**, p. 54

Null Alleles

- Allele is present in the DNA sample but fails to be amplified due to a **nucleotide change in a primer binding site**
- Allele dropout is a problem because a heterozygous sample appears falsely as a homozygote
- Two PCR primer sets can yield different results on samples originating from the same source
- This phenomenon impacts DNA databases
- Large concordance studies are typically performed prior to use of new STR kits

Primer Binding Site Mutations can lead to "null alleles" that are not detected



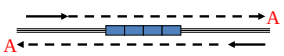
Concordance testing (involving DNA testing with different primer sets) reveals a missing "null" allele with one of the DNA tests

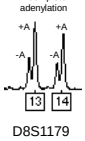
J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Figure 4.8**, p. 100

Non-Template Addition

- Taq polymerase will often add an extra nucleotide to the end of a PCR product; most often an "A" (termed "adenylation")
- **Dependent on 5'-end of the reverse primer; a "G" can be put at the end of a primer to promote non-template addition**
- Can be enhanced with extension soak at the end of the PCR cycle (e.g., 15-45 min @ 60 or 72 ºC) – to give polymerase more time
- Excess amounts of DNA template in the PCR reaction can result in incomplete adenylation (not enough polymerase to go around)

Best if there is NOT a mixture of "+/- A" peaks
(desirable to have full adenylation to avoid split peaks)

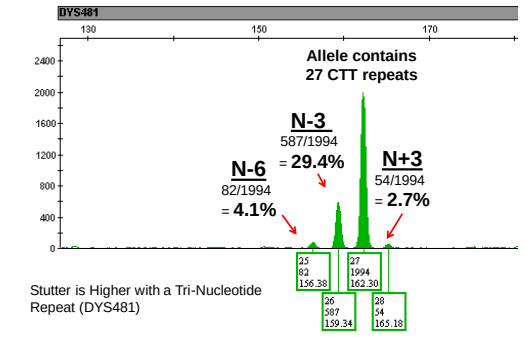




Stutter Products

- Peaks that show up primarily one repeat less than the true allele as a result of strand slippage during DNA synthesis
- Stutter is less pronounced with larger repeat unit sizes
  (dinucleotides > tri- > tetra- > penta-)
- Longer repeat regions generate more stutter
- Each successive stutter product is less intense
  (allele > repeat-1 > repeat-2)
- Stutter peaks make mixture analysis more difficult

Stutter Artifacts Complicate Interpretation of STR Results



Stutter is Higher with a Tri-Nucleotide Repeat (DYS481)

J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Figure 3.10**, p. 71

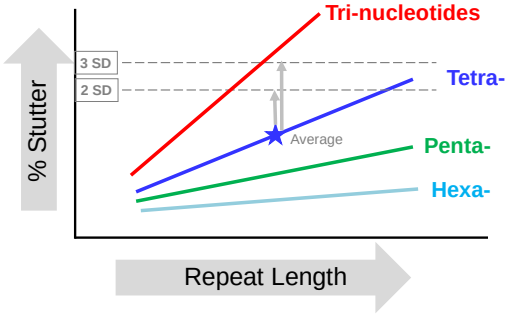
Types of STR Repeat Units

Requires size based DNA separation to resolve different alleles from one another

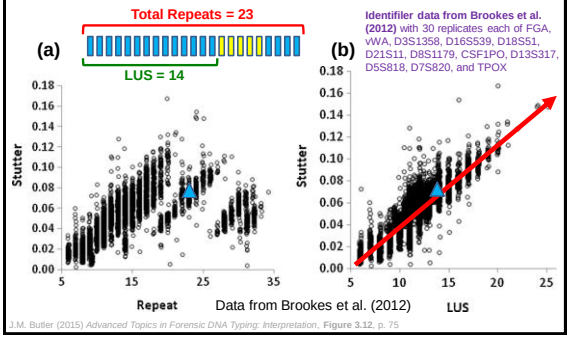
- **Di**nucleotide (CA)(CA)(CA)(CA)
- **Tri**nucleotide (GCC)(GCC)(GCC)
- **Tetra**nucleotide (AATG)(AATG)(AATG)
- **Penta**nucleotide (AGAAA)(AGAAA)
- **Hexa**nucleotide (AGTACA)(AGTACA)

Short tandem repeat (STR) = microsatellite = simple sequence repeat (SSR)

Simplified Illustration of Stutter Trends



Stutter Ratios Model Better with Longest Uninterrupted Stretch (LUS) Compared to Total Repeat Length



J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Figure 3.12**, p. 75

STR Genotypes

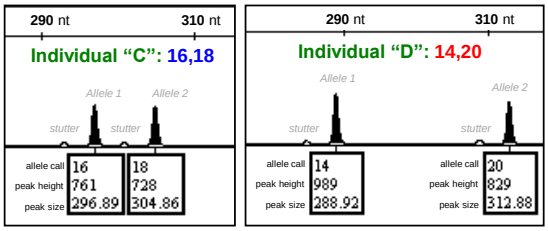
Heterozygote Balance, Stochastic Effects, etc.

Advanced Topics in Forensic DNA Typing: Interpretation, Chapter 4

Key Points with STR Genotypes

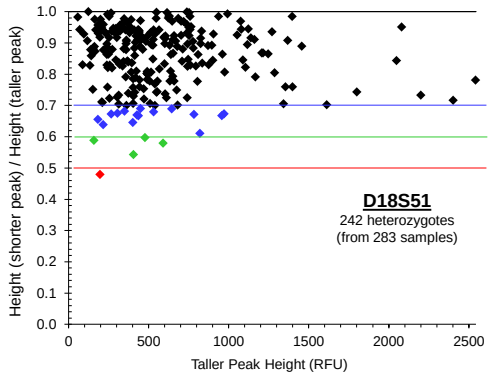
- In heterozygous loci, the two alleles should be equal in amount; however, stochastic effects during PCR amplification (especially when the amount of DNA being amplified is limited) create an imbalance in the two detected alleles
- Heterozygote balance (Hb) or peak height ratios (PHRs) measure this level of imbalance
- Under conditions of extreme imbalance, one allele may "drop-out" and not be detected
- Stochastic thresholds are sometimes used to help assess the probability of allele drop-out in a DNA profile

D18S51 Results from Two Samples



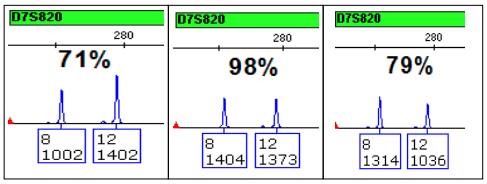
Peak Height Ratios (PHRs) or Heterozygote balance (*Hb*)

728/761 = 0.957= **95.7%** 829/989 = 0.838 = **83.8%**



J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Figure 4.2**, p. 90

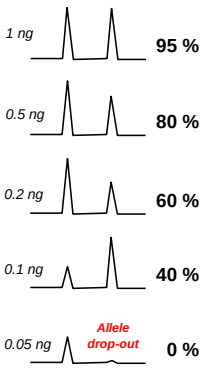
Natural Variation in Peak Height Ratio During Replicate PCR Amplifications



The heights of the peaks will vary from sample-to-sample, even for the same DNA sample amplified in parallel

Slide from Charlotte Word (ISHI 2010 mixture workshop)

Hypothetical Heterozygote Alleles



Heterozygote balance typically decreases with DNA template level

In the extreme, one of the alleles fails to be amplified (this is known as allele drop-out)

J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Figure 4.3**, p. 92

STR Profiles

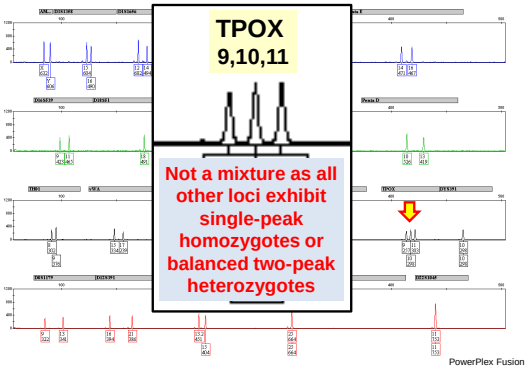
Multiplex PCR, Tri-Alleles, Amelogenin, and Partial Profiles

Advanced Topics in Forensic DNA Typing: Interpretation, Chapter 5

Key Points with STR Profiles

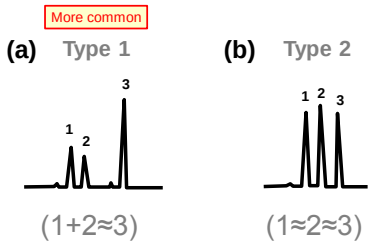
- **Tri-allelic patterns occasionally occur** at STR loci (~1 in every 1000 profiles) and are due to copy number variation (CNVs) in the genome
- **Due to potential deletions of the amelogenin Y region, additional male confirmation markers are used in newer 24plex STR kits**
  - The amelogenin gene is found on both the X and Y chromosomes and portions of it can be targeted to produce assays that enable gender identification as part of STR analysis using commercial kits
- **Partial profiles can result** from low amounts of DNA template or DNA samples that are damaged or broken into small pieces or contain PCR inhibitors

Single-Source DNA Sample Exhibiting a TPOX Tri-Allelic Pattern



J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Figure 5.2**, p. 113

Types of Tri-Allelic Patterns



This classification scheme was developed by Tim Clayton and colleagues at the UK Forensic Science Service (Clayton et al. 2004, *J. Forensic Sci.* 49: 1207-1214)

J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Figure 5.3**, p. 114

Tri-Allelic Patterns Occur about 1 in 1000 Profiles

but the frequency varies across STR loci

D.N.A. BOX 5.2

ESTIMATION OF AVERAGE FREQUENCY FOR TRI-ALLELIC PATTERNS IN STR PROFILES

Several years ago, DNA analyst Malena Jimenez from the Missouri State Highway Patrol supplied the NIST STRBase website with information on tri-alleles observed during analysis of 69,600 convicted offender samples. Below is a summary of the number of reported tri-alleles for each locus examined with their PowerPlex 16 single-source profiles.

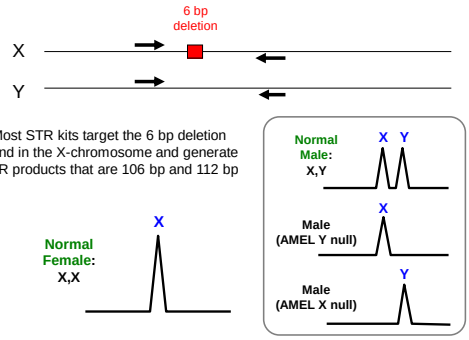
Noting a total of 68 tri-alleles observed across these 69,600 samples, a tri-allelic pattern is expected to occur about once every one thousand samples on average. However, as can be seen with this data, the distribution of tri-allelic patterns is not even across loci.

Source: Steven Myers, California Department of Justice presentation provided to the author based on data collected from the NIST STRBase website at http://www.cstl.nist.gov/strbase/tri_tab.htm.

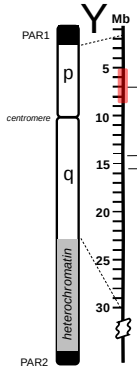
STR Locus	# Tri-Alleles	STR Locus	# Tri-Alleles	STR Locus	# Tri-Alleles
CSF1PO	1	D3S1358	2	D16S539	3
FGA	11	D5S818	1	D18S51	3
TH01	0	D7S820	0	D21S11	9
TPOX	9	D8S1179	2	Penta D	3
vWA	10	D13S317	4	Penta E	10

J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Box 5.2**, p. 114

Amelogenin Sex-Typing Assay



J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Figure 5.4**, p. 119



Relative Positions Along the Y-Chromosome of Amelogenin (AMEL Y) and Male Confirmation Markers Used in Newer STR Kits

SRY

AMEL Y

DYS391

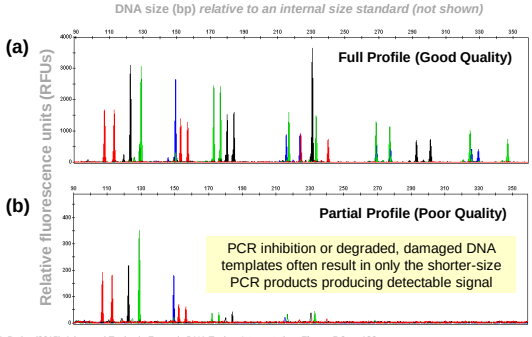
Y-InDel (M175, rs203678)

Deletions of the Y-chromosome can encompass >1 Mb around the AMEL Y region
(DYS458 from Y-STR kits is often lost in these situations)

STR Kit	Male Confirmation Marker(s)
PowerPlex Fusion	DYS391
GlobalFiler	DYS391, Y-InDel
Investigator 24plex	DYS391, Y-InDel

J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Figure 5.5**, p. 122

Partial Profiles Can Occur from Poor Quality DNA or Low Amounts of DNA Template

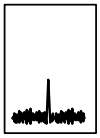
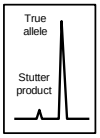
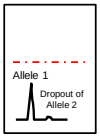
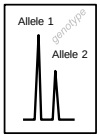


J.M. Butler (2015) *Advanced Topics in Forensic DNA Typing: Interpretation*, **Figure 5.6**, p. 122

Data Interpretation Overview

The Steps of Data Interpretation

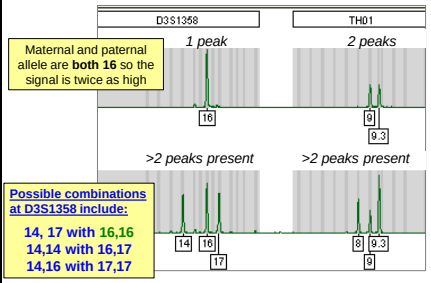
Peak (vs. noise) → **Allele** (vs. artifact) → **Genotype** (allele pairing) → **Profile** (genotype combining)

Analytical Threshold	Expected Stutter %	Stochastic Threshold	Peak Height Ratio (PHR)
			

Next step: Examine feasible genotypes to deduce possible contributor profiles

Moving from individual locus genotypes to profiles of potential contributors to the mixture is dependent on mixture ratios and numbers of contributors

Single-Source Sample vs Mixture Results

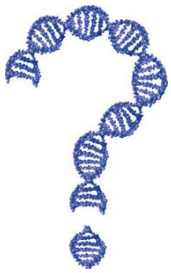


Single-Source

Mixture

Multiple possible combinations could have given rise to the mixture observed here

Thank you for your attention!



Contact info:
john.butler@nist.gov
+1-301-975-4049

A copy of this presentation will be available at:
http://strbase.nist.gov/training.htm

29 August 2017

Workshop #10

Length vs Sequence Information:

Lessons Learned from TPOX and SE33

Lisa Borsuk, M.S.
Applied Genetics Group
U.S. National Institute of Standards and Technology

Official Disclaimer

- The opinions and assertions contained herein are solely those of the author and are not to be construed as official or as views of the U.S. Department of Commerce.
- Commercial equipment, instruments, software, or materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the U.S. Department of Commerce, nor does it imply that any of the materials, instruments, software or equipment identified are necessarily the best available for the purpose.
- All work presented has been reviewed and approved by the NIST Human Subjects Protections Office.

Outline

- What is the NIST Population Set and how is it used?
  - Overview of STRs
  - The sequencing of the NIST Population Set
  - Analysis of the NIST Population Set
- Analysis Pipeline
  - Modifications
  - Analysis of results
- TPOX – a quick look at how simple STRs can be affected by sequencing
- SE33 – a deeper look into a highly polymorphic STR marker

Population Samples

- **1036 Good Quality Single Source Samples**
  - Self Identified
    - 342 African American
    - 97 Asian
    - 361 Caucasian
    - 236 Hispanic
- Unrelated Anonymous Individuals – mostly male to include the Y Chromosome
  - Male – 1032
  - Female – 4
- Collected over a number of years from a variety of sources

>1000 DNA extracts

NIST Population Samples

- Highly Characterized Set
- Reliably report a number of DNA markers for human identity testing
  - Short Tandem Repeats (STR)
  - Single Nucleotide Polymorphism (SNP)
  - Insertions and or Deletions (INDEL)
- Generated data for population frequencies
- Tested using a variety of methods
  - Capillary Electrophoresis (CE) kits for STR genotyping and confirmation

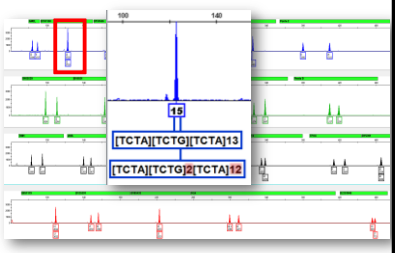
Forensic STRs

- Mainly looking at 4 base repeats
- Sequence motif varies by locus
- Number of repeats varies by locus (and by individual)
  - Highly polymorphic
- Repeat Sequence Classes
  - Simple
  - Compound
  - Complex
  - Hypervariable

Short Tandem Repeats **(TPOX – Simple) 6,8 Genotype**

Capillary Electrophoresis

- **Length based** STR results
- Fast turn around time from collection to results
- Accurate reproducible results
- Multiplexed Kits
  - ~20s Loci plus
- No Sequence Information



Platform and Kit Used for Sequencing

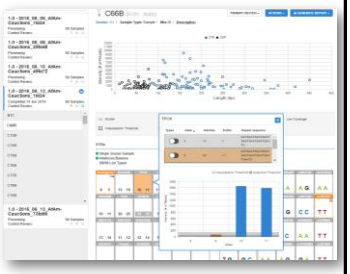
Platform: Illumina FGx System

Forensically relevant STR and SNP markers

- ForenSeq multiplex kit
  - 58 STRs, Amel, and 172 SNPs
    (Includes 27 Autosomal STRs)
- Data analysis
  - Universal Analysis Software (UAS)

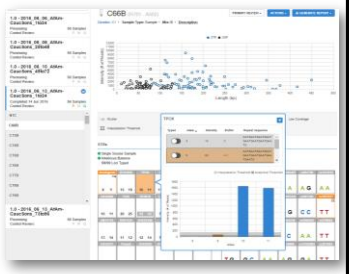
The UAS

- GUI for exploration of data
  - Easy to use
- Specific to ForenSeq workflow
- ~45 minutes to analyze a run of samples
- User-adjustable parameters
- Allows manual edits
- Exportable Excel report



The UAS

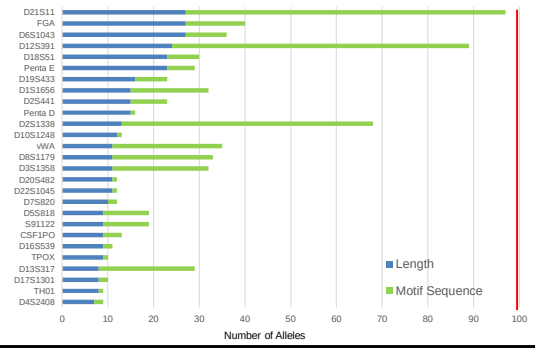
- Only includes sequences >10X DoC
- Limited flanking sequence evaluation
  - Focused on repeat region of STR
- **SE33** not visible in UAS
  - Raw data *is* present for SE33
  - There are other "hidden" loci not reported by the UAS
    - DYS456
    - DYS461
    - DXS10148
    - DXS3877



Autosomal STRs

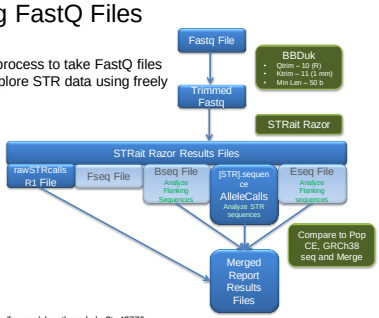
All loci show gains from sequencing

N = 1036



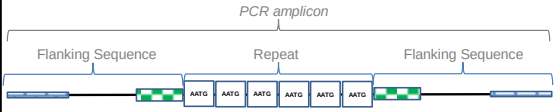
Processing FastQ Files

A example analysis process to take FastQ files and evaluate and explore STR data using freely available packages.



BBDuk - http://seqanswers.com/forums/showthread.php?t=42776
STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. Warshauer et al., Forensic Sci Int Genet. 2013 (7):409-17
STRait Razor v2.0: the improved STR Allele Identification Tool--Razor. Warshauer et al., Forensic Sci Int Genet. 2015 (14):182-6

STRait Razor



Software returns:
The length between the anchor sequences (= 24)
A reference table returns a "6" allele and the sequence between the anchor sites

PCR primers
Anchor site (~12 nt)
STR repeat region

[1] STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. Warshauer et al., Forensic Sci Int Genet. 2013 (7):409-17
[2] STRait Razor v2.0: the improved STR Allele Identification Tool--Razor. Warshauer et al., Forensic Sci Int Genet. 2015 (14):182-6
[3] *http://battelleexactid.org/*

Processing The STRait Razor Results

- Determine allele calls
  - Removing noise and stutter
- STR sequence Bracketing
  - Visually simplifies the sequence
  - Highlights the repetitive characteristics of the sequence
- Additional calculations
  - Allele Coverage Ratio (**ACR**)
  - Non Majority Allele (**NonMA**)

STRait Razor v2.0 – TPOX Raw Results

Locus:Allele	Length	Sequence	Coverage
TPOX:7	28 bases	AATGAATGAATGAATGAATGAATGAATG	9
TPOX:7	28 bases	AATGAATGAATGAGTGAATGAATGAATG	1
TPOX:8	32 bases	AATGAATGAATGAATGAATGAATGAATGAATG	332
TPOX:8	32 bases	AATGAATGAATGAATGAATGAATGAAAG	2
TPOX:8	32 bases	AATGAATGAATGAATGAAGGAATGAATG	1
TPOX:8	32 bases	AATGAATGAATGAATGACTGAATGAATGAATG	1
TPOX:8	32 bases	AACGAATGAATGAATGAATGAATGAATG	1
TPOX:8	32 bases	AATGAATGAATGAATGAGTGAATGACTGATTG	1
TPOX:8	32 bases	AATGAATGAATGAATGAATGAATGGATG	1
TPOX:8	32 bases	AATGAATGATTGAATGAATGAATGAATG	1
TPOX:8	32 bases	AATGAATGAATGAATGAATGAATGTATT	1
TPOX:8	32 bases	AATGAATGAATGGATGAATGAATGAATG	1
TPOX:8	32 bases	AATGAATGAACGAATGACTGAATGAATG	1
TPOX:9	36 bases	AATGAATGAATGAATGAATGAATGAATGAATG	17
TPOX:10	40 bases	AATGAATGAATGAATGAATGAATGAATGAATGAATG	321
TPOX:10	40 bases	ACTGAATGAATGAATGAATGAATGAATGAATGAATG	1
TPOX:10	40 bases	AATGAATGAATGAATGAATGAATGAAAGAATGAATG	1
TPOX:10	40 bases	AATGAATGAATGAGTGAATGAATGAATGAATGAATG	1
TPOX:10	40 bases	AATGAATGAATAAATGAATGAATGAATGAATGAATG	1
TPOX:10	40 bases	AATGAATGAATGAATGAATGAATGAATGATGAATG	1
TPOX:10	40 bases	AATGAATGAATGAATGAATGAATGAATGAATGAATG	1
TPOX:11	44 bases	AATGAATGAATGAATGAATGAATGAATGAATGAATGAATG	2

STRait Razor – TPOX Results Dissected

Locus	Length	Sequence	Coverage	Comment
TPOX:7	28 bases	AATGAATGAATGAATGAATGAATGAATG	9	n-4 stutter
TPOX:7	28 bases	AATGAATGAATGA**G**TGAATGAATGAATG	1	
TPOX:8	32 bases	AATGAATGAATGAATGAATGAATGAATGAATG	332	Allele
TPOX:8	32 bases	AATGAATGAATGAATGAATGAATGAA**A**G	2	
TPOX:8	32 bases	AATGAATGAATGAATGAATGAA**G**GAATGAATG	1	
TPOX:8	32 bases	AATGAATGAATGAATGA**C**TGAATGAATGAATG	1	
TPOX:8	32 bases	AA**C**GAATGAATGAATGAATGAATGAATG	1	
TPOX:8	32 bases	AATGAATGAATGAATGA**G**TGAATGACTGA**T**TG	1	
TPOX:8	32 bases	AATGAATGAATGAATGAATGAATG**G**ATG	1	
TPOX:8	32 bases	AATGAATGA**T**TGAATGAATGAATGAATGAATG	1	
TPOX:8	32 bases	AATGAATGAATGAATGAATGAATG**T**AT**T**	1	
TPOX:8	32 bases	AATGAATGAATG**G**ATGAATGAATGAATG	1	
TPOX:8	32 bases	AATGAATGAA**C**GAATGA**C**TGAATGAATGAATG	1	
TPOX:9	36 bases	AATGAATGAATGAATGAATGAATGAATGAATGAATG	17	n-4 stutter; n+4 sutter
TPOX:10	40 bases	AATGAATGAATGAATGAATGAATGAATGAATGAATG	321	Allele
TPOX:10	40 bases	A**C**TGAATGAATGAATGAATGAATGAATGAATGAATG	1	
TPOX:10	40 bases	AATGAATGAATGAATGAATGAATGAA**A**GAATGAATG	1	
TPOX:10	40 bases	AATGAATGAATGA**G**TGAATGAATGAATGAATGAATG	1	
TPOX:10	40 bases	AATGAATGAAT**A**AATGAATGAATGAATGAATGAATG	1	
TPOX:10	40 bases	AATGAATGAATGAATGAATGAATGAATGATG**T**TG	1	
TPOX:10	40 bases	AATGAATGAATG**C**ATGAATGAATGAATGAATGAATG	1	
TPOX:11	44 bases	AATGAATGAATGAATGAATGAATGAATGAATGAATGAATG	2	n+4 stutter?

STRait Razor – TPOX Results Dissected

Allele 8

Majority Allele Count = 332

TPOX:8	32 bases	AATGAATGAATGAATGAATGAATGAATG	332	Allele
TPOX:8	32 bases	AATGAATGAATGAATGAATGAATGAA**A**G	2	
TPOX:8	32 bases	AATGAATGAATGAATGAA**G**GAATGAATG	1	
TPOX:8	32 bases	AATGAATGAATGAATGA**C**TGAATGAATGAATG	1	
TPOX:8	32 bases	AA**C**GAATGAATGAATGAATGAATGAATG	1	Non Majority
TPOX:8	32 bases	AATGAATGAATGAATGA**G**TGAATGACTGA**T**TG	1	Allele Count = 11
TPOX:8	32 bases	AATGAATGAATGAATGAATGAATG**G**ATG	1	
TPOX:8	32 bases	AATGAATGA**T**TGAATGAATGAATGAATGAATG	1	
TPOX:8	32 bases	AATGAATGAATGAATGAATGAATG**T**AT**T**	1	
TPOX:8	32 bases	AATGAATGAATG**G**ATGAATGAATGAATG	1	
TPOX:8	32 bases	AATGAATGAA**C**GAATGA**C**TGAATGAATGAATG	1	

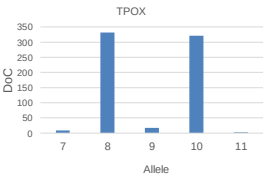
Total Coverage = 343

Non Majority Alleles are most likely caused by
- PCR Base Incorporation Errors
- Sequencing Errors

Majority Allele (MA) = 332/343 = 0.97
Non Majority Allele (NonMA) = 11/343 = 0.03

Allele Coverage Ratio (ACR)

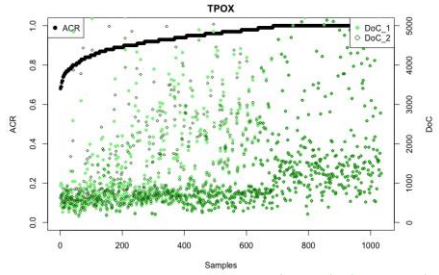
- For heterozygous genotypes
- ACR is calculated by dividing the **lower DoC** allele by the **higher DoC** allele
  - Example on the right is an 8,10 at the TPOX locus
  - ACR = DoC (allele 10)/DoC (allele 8)
  - 0.97 = 321 / 332



- Looking for "balanced alleles" ACR > 0.6
  - Commonly used threshold

- Similar to heterozygote balance or peak height ratio with CE data

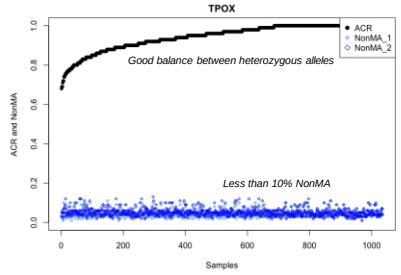
NIST Population Samples – TPOX



NIST 1036 population samples run using the ForenSeq kit

The Depth of Coverage (DoC) for TPOX is several hundred to several thousand sequences for both alleles in all samples

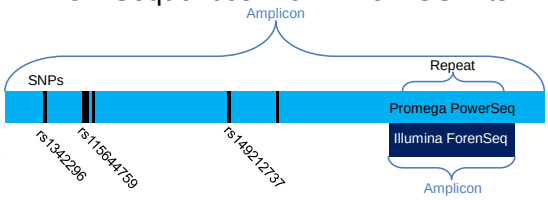
NIST Population Samples – TPOX



NIST 1036 population samples run using the ForenSeq kit

The non-majority allele (NonMA) is a small percentage of the alleles sequences in TPOX

TPOX Sequences From Two NGS Kits



- The ForenSeq kit targets the repeat area
- The PowerSeq kit includes a larger amount of flanking sequence
  - Includes several additional SNPs
    - 3 have been observed in the current NIST Promega dataset
      - rs1342296
      - rs115644759
      - rs149212737

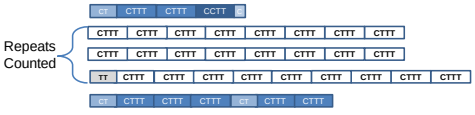
Conclusions for TPOX

- TPOX is a simple STR
- Lower levels of sequencing errors and noise
- NGS based results concordant with CE length-based methods
- TPOX has SNPs in the flanks that do not affect length but do result in unique sequences
- Sequencing TPOX results in more information about the locus

SE33 Basic Information

- One of the most polymorphic markers used by the forensic community
- 97% Heterozygosity by sequence
- In the NIST Population Set
- observed length-based range 6.3 to 36 tetranucleotides
- 52 unique alleles by length
- 264 unique alleles by sequence

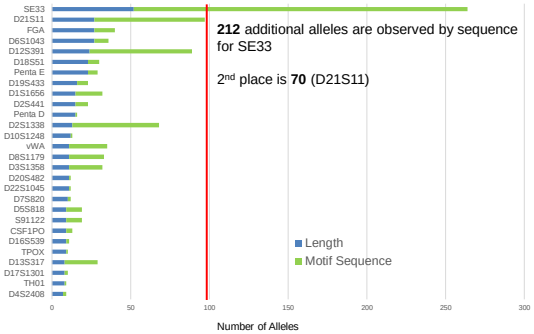
SE33 (Complex) - 25.2 Allele from the Reference Genome Sequence



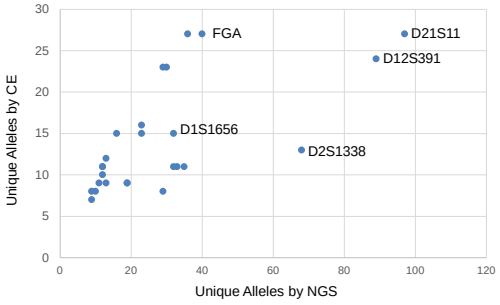
- CT [CTTT]2 CCTT C **[CTTT]16 TT [CTTT]9** CT [CTTT]3 CT [CTTT]2

Autosomal STRs

N = 1036

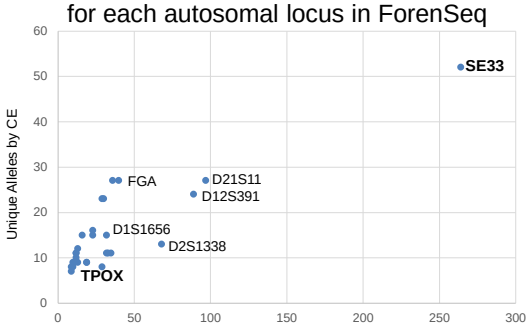


Numbers of Unique Alleles for CE vs. NGS for each autosomal locus in ForenSeq



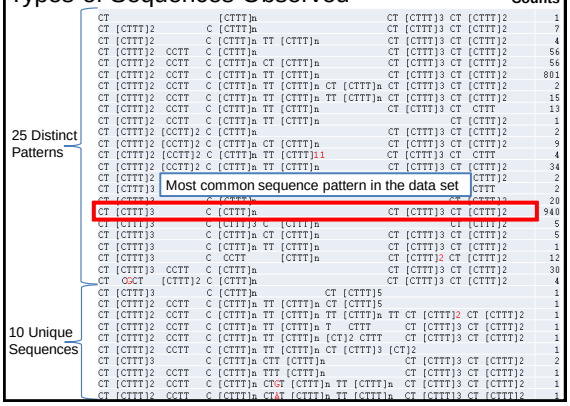
NIST 1036 Population Samples

Numbers of Unique Alleles for CE vs. NGS for each autosomal locus in ForenSeq



NIST 1036 Population Samples

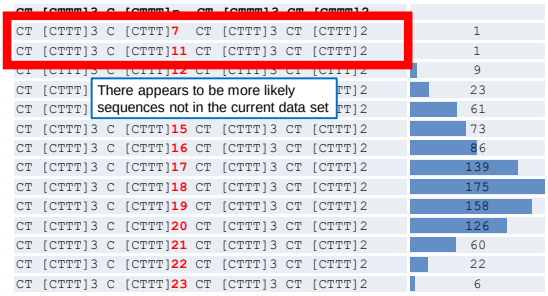
Types of Sequences Observed



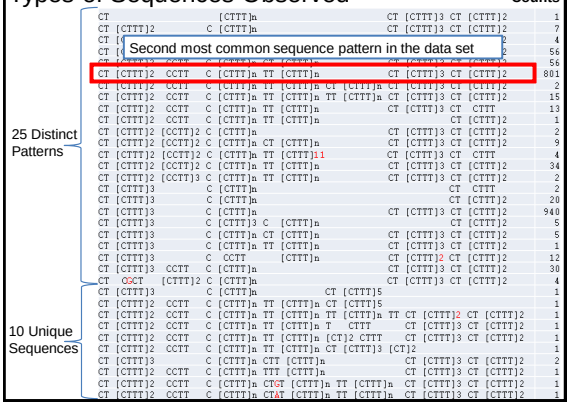
Most Common Pattern

14 Alleles

Counts in 1036



Types of Sequences Observed



Second Most Common Pattern

- **CT [CTTT]2 CCTT C [CTTT]m TT [CTTT]n CT [CTTT]3 CT [CTTT]2**
  - Observed range of m is between 8 – 24
  - Observed range of n is between 5 – 16
  - Observed range of allele length is between 19.2 – 32.2
- 109 Observed Alleles
  - There appears to be more sequences not in the current data set
- Examples

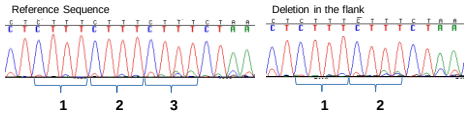
Concordance Check with CE

- 1036 samples typed with CE and NGS were compared
- **22** discordant samples
  - 21 resolved by examining NGS flanking sequences for deletions
    - Confirmed with Sanger
  - 1 resolved by examining **Sanger** flanking sequences for deletions

Example of discordance
- CE – 27.2,28.2
- NGS — 28.2*,28.2
- CTTT* deletion in the flank detected by NGS and confirmed by Sanger sequencing

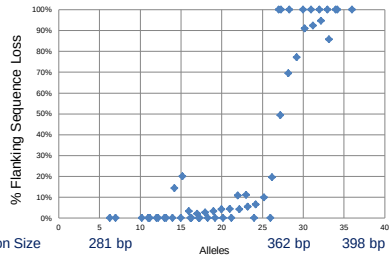
Sanger Sequencing Check

- Confirmed flanking sequence deletions



- Confirmed SE33 patterns
- Confirmed low coverage high noise sequences

Flanking Sequence

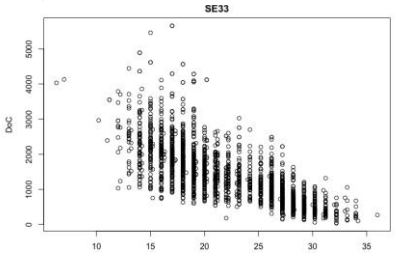


Amplicon Size 281 bp 362 bp 398 bp

Looking at the 75 bases beyond the repeat region as the flanking sequence

The flanking sequence becomes extremely noisy with many poor quality base calls in alleles larger then 27

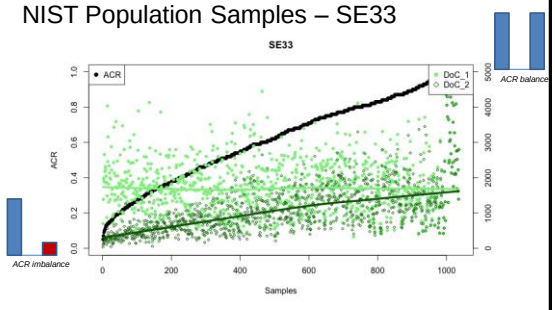
SE33 – Depth of Coverage decreases as length of allele increases



Largest observed SE33 allele is 49

Repeat Sequence Only

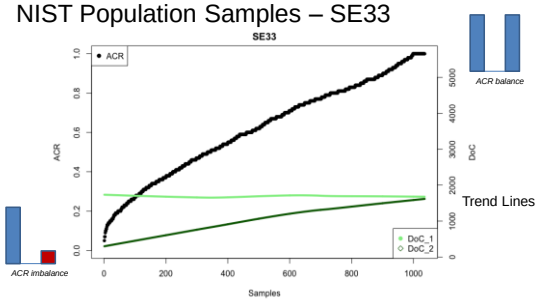
NIST Population Samples – SE33



The light green is the smaller allele from the sample – the dark green is the larger allele from the sample

The depth of coverage of the sequence is a problem with the larger allele

NIST Population Samples – SE33

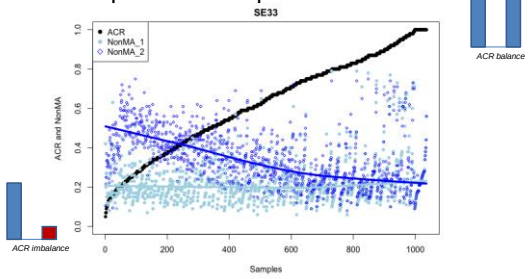


Trend Lines

The light green is the smaller allele from the sample – the dark green is the larger allele from the sample

The depth of coverage of the sequence is a problem with the larger allele

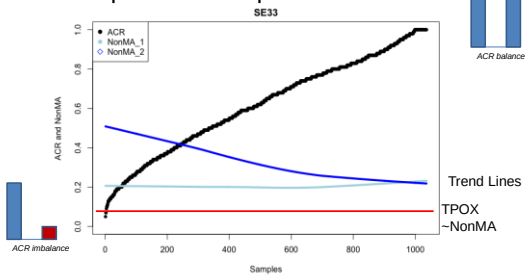
NIST Population Samples – SE33



The light blue is the smaller allele from the sample – the dark blue is the larger allele from the sample

There are more non-majority sequences present in the larger allele resulting in more noise in the sample

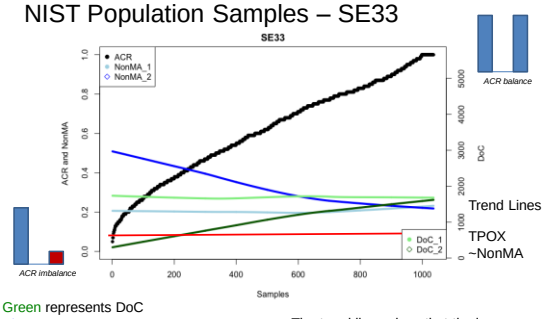
NIST Population Samples – SE33



The light blue is the smaller allele from the sample – the dark blue is the larger allele from the sample

There are more non-majority sequences present in the larger allele resulting in more noise in the sample

NIST Population Samples – SE33



Green represents DoC
Blue represents NonMA
The lighter color is the smaller allele in the sample. The darker color is the larger allele

The trend lines show that the larger allele has lower DoC and higher NonMA which is resulting in a low ACR

Conclusions of SE33

- SE33 genotypes for 1036 high quality single source samples were recovered from the ForenSeq data set
- Results concordant with CE length-based methods
  - Sanger sequencing provided further conformation
- SE33 is a complex STR to analyze
  - Required more manual data curation
  - Disparate depth of coverage for larger alleles resulted in allele coverage ratio imbalance and generally low depth of coverage
  - Longer alleles present a greater challenge for accurate sequencing – 49 is the largest Sanger sequence allele currently observed
- The SE33 sequence data set is a good resource for the community

Conclusions

- Sequencing adds alleles
  - Including flanking sequencing can add additional alleles
  - TPOX and other simple loci can gain from sequencing
- SE33 length and sequence complexity results in difficultly in sequencing
  - Other loci do demonstrate SE33 issues but not to the same degree

Acknowledgments

NIST
Dr. Peter Vallone
Dr. Katherine Gettings
Becky Steffen
Kevin Kiesler

Funding

Contact lisa.borsuk@nist.gov

Poster P01-53
Sequencing of the highly polymorphic STR locus SE33
Presentation Date: Wed Aug. 30 10:30 – 11:30

STR Nomenclature, STRSeq, and STRidER

Katherine Gettings PhD Research Biologist – NIST USA

WORKSHOP #10
International Society for Forensic Genetics
Tuesday, August 29, 2017

Autosomal STR Markers and Interpretation

Disclaimers

Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Commerce. Certain commercial equipment, instruments, and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by NIST, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose.

Funding:
NIJ Inter-Agency Agreement: Forensic DNA Application of Next Generation Sequencing.

A copy of this presentation is available at: http://strbase.nist.gov/NISTpub.htm#Presentations

STR Nomenclature

Focusing on

- Autosomal STR Loci Sequencing
- Work over the past two years
- Foundation for NGS/MPS
- CE Back-compatibility

STR Nomenclature

Forensic Science International: Genetics 18 (2015) 118–130

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

STR allele sequence variation: Current knowledge and future issues

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

[a] U.S National Institute of Standards and Technology, Biomolecular Measurement Division, 100 Bureau Drive, Gaithersburg, MD 20899, USA
[b] The George Washington University, Department of Forensic Sciences, 2100 Foxhall Road NW, Washington, DC 20007, USA
[c] U.S. National Institute of Standards and Technology, Special Programs Office, 100 Bureau Drive, Gaithersburg, MD 20899, USA

STR Review Article

Table 2
Genomic information on the twenty-four most commonly used autosomal STR loci.

STR locus	Chromosome position	GRCh38 (Dec 2013)			
		Location	Strand	Allele designation	Repeat region sequence[a]
D1S1656	1q42	230,769,616–230,769,683	R	17	[TAGA]$_{16}$[TAGG] **[TG]$_5$**
TPOX	2p25.3	1,489,653–1,489,684	F	8	[AATG]$_8$
D2S441	2p14	68,011,947–68,011,994	F	12	[TCTA]$_{12}$
D2S1338	2q35	218,014,859–218,014,950	R	23	[TGCC]$_7$ [TTCC]$_{13}$ [GTCC] [TTCC]$_2$
D3S1358	3p21.31	45,540,739–45,540,802	F	16	[TCTA] [TCTG] [TCTA]$_{14}$
FGA	4q31.3	154,587,736–154,587,823	R	22	[TTTC]$_3$ [TTTT] [TTCT] [CTTT]$_{14}$ [CTCC] [TTCC]$_2$
D5S818	5q23.2	123,775,556–123,775,599	R	11	[AGAT]$_{11}$
CSF1PO	5q33.1	150,076,324–150,076,375	F	13	[ATCT]$_{13}$
SE33	6q14	88,277,144–88,277,245	R	25.2	**[AAAG]$_2$ AG [AAAG]$_3$**AG [AAAG]$_9$ AAAAAG [AAAG]$_{15}$ **G AAGG [AAAG]$_2$ AG**
D6S1043	6q15	91,740,225–91,740,272	R	12	[AGAT]$_{12}$
D7S820	7q21.11	84,160,226–84,160,277	R	13	[GATA]$_{13}$
D8S1179	8q24.13	124,894,865–124,894,916	F	13	[TCTA] [TCTG] [TCTA]$_{11}$
D10S1248	10q26.3	129,294,244–129,294,295	F	13	[GGAA]$_{13}$
TH01	11p15.5	2,171,088–2,171,115	F	7	[AATG]$_7$
vWA	12p13.31	5,983,977–5,984,044	R	17	[TCTA] [TCTG]$_5$[TCTA]$_{11}$ **TCCA TCTA**
D12S391	12p13.2	12,297,020–12,297,095	F	19	[AGAT]$_{11}$ [AGAC]$_7$ [AGAT]
D13S317	13q31.1	82,148,025–82,148,068	F	11	[TATC]$_{11}$ **[AATC]$_2$**
Penta E	15q26.2	96,831,015–96,831,039	R	5	[AAAGA]$_5$
D16S539	16q24.1	86,352,702–86,352,745	F	11	[GATA]$_{11}$
D18S51	18q21.33	63,281,667–63,281,738	F	18	[AGAA]$_{18}$
D19S433	19q12	29,926,235–29,926,298	R	14	[AAGG] **AAAG**[AAGG] **TAGG** [AAGG]$_{12}$
D21S11	21q21.1	19,181,973–19,182,099	F	29	[TCTA]$_4$[TCTG]$_6$[TCTA]$_3$**TA**[TCTA]$_3$**TCA**[TCAT]$_2$**TCCATA**[TCTA]$_{11}$
Penta D	21q22.3	43,636,205–43,636,269	F	13	[AAAGA]$_{13}$
D22S1045	22q12.3	37,140,287–37,140,337	F	17	[ATT]$_{14}$ ACT [ATT]$_2$

[a] Nucleotides in bold are not counted toward the allele designation.

STR Review Article

Descriptions for 24 autosomal STR loci

D5S818 is a simple tetranucleotide repeat found on the long arm of chromosome 5. The AGAT repeat unit corresponds to the reverse strand of the GenBank reference sequence (GRCh38, Chromosome 5, location 123775556–123775599, reverse strand, allele designation 11) with sequenced alleles ranging in size from 7 to 18 repeat units. Length-based alleles have been reported from 4 to 20 repeats [26,46], while the largest observed "allele" measures 29 by length but contains only 17 repeats by sequence due to a 48 base insertion within the repeat region [8]. Sequence data also shows a rare, non-consensus ACAT tetranucleotide that breaks up the repeat motif: $[AGAT]_n$ [ACAT] $[AGAT]_n$ [14,33]. Microvariant alleles have also been reported including a trinucleotide within the repeat region that results in an x.3 allele (e.g., $[AGAT]_n$ GAT $[AGAT]_n$) [33] and a single adenine base adjacent to the repeat unit on the 5′ end (e.g., A $[AGAT]_n$) [8]. One high frequency SNP has been reported within 150 bp of the repeat region: rs25768 is located 13 bp from the 5′ end of the repeat region, and has a frequency of 16%. This SNP was the subject of a study that showed the potential of flanking region polymorphisms in lineage attribution of STR mutations in parentage cases [29].

- Orient the locus on GRCh38
- Allele ranges by length and sequence
- Sequence motifs and microvariants
- Flanking region polymorphisms

STR Review Article

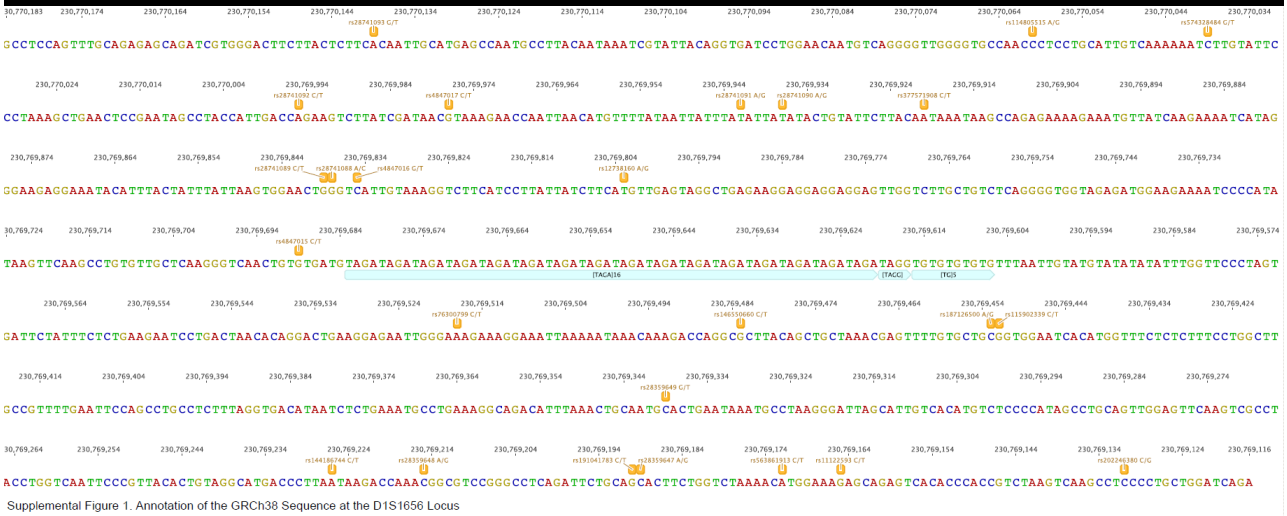
Supplemental Table 1

- Observed alleles for 24 autosomal STR loci
- Excel file with a tab for each locus
- Sequences organized by motif
- References given
- Sequencing method noted
- Length-only observations also noted

D1S1656			
Allele	Repeat Structure	Reference	Platform
	[TAGA]9-14[TG]5		
8	[TAGA]8 [TG]5	Phillips et al. (2011)	Sanger
9	[TAGA]9 [TG]5	Phillips et al. (2011)	Sanger
10	[TAGA]10[TG]5	Lareu et al. (1998)	Sanger
11	[TAGA]11[TG]5	Lareu et al. (1998), 2391c Components B/C/E	Sanger
12	[TAGA]12[TG]5	Lareu et al. (1998)	Sanger
13	[TAGA]13[TG]5	Phillips et al. (2011)	Sanger
14	[TAGA]14[TG]5	Phillips et al. (2011), 2391c Component B	Sanger
16	[TAGA]16[TG]5	Gettings et al. (2015)	MiSeq
	[TAGA]9-17[TAGG][TG]5		
10	[TAGA]9 [TAGG][TG]5	Phillips et al. (2011)	Sanger
12	[TAGA]11[TAGG][TG]5	Lareu et al. (1998)	Sanger
13	[TAGA]12[TAGG][TG]5	Lareu et al. (1998)	Sanger
14	[TAGA]13[TAGG][TG]5	Lareu et al. (1998)	Sanger
15	[TAGA]14[TAGG][TG]5	Lareu et al. (1998), 2391c Component C	Sanger
16	[TAGA]15[TAGG][TG]5	Lareu et al. (1998)	Sanger
17	[TAGA]16[TAGG][TG]5	Lareu et al. (1998)	Sanger
18	[TAGA]17[TAGG][TG]5	Phillips et al. (2011)	Sanger
	[TAGA]1-4[TGA][TAGA]9-14[TAGG][TG]5		
13.3	[TAGA]1[TGA][TAGA]11[TAGG][TG]5	Phillips et al. (2011)	Sanger
14.3	[TAGA]4[TGA][TAGA]9 [TAGG][TG]5	Phillips et al. (2011)	Sanger
15.3	[TAGA]3[TGA][TAGA]11[TAGG][TG]5	Gettings et al. (2015)	MiSeq
15.3	[TAGA]4[TGA][TAGA]10[TAGG][TG]5	Lareu et al. (1998)	Sanger
16.3	[TAGA]4[TGA][TAGA]11[TAGG][TG]5	Lareu et al. (1998), 2391c Component E	Sanger
17.3	[TAGA]4[TGA][TAGA]12[TAGG][TG]5	Lareu et al. (1998), 2391c Components A/F	Sanger
18.3	[TAGA]4[TGA][TAGA]13[TAGG][TG]5	Lareu et al. (1998)	Sanger
19.3	[TAGA]4[TGA][TAGA]14[TAGG][TG]5	Lareu et al. (1998)	Sanger
Other Repeat Region Variants			
13	[TAGA]11[TAGC][TAGA][TG]5	Gettings et al. (2015)	MiSeq
16	[TAGA]15[TAAG][TG]5	Gettings et al. (2015)	MiSeq
Genotyped by Length Only			
17.1		Schröer et al. (2000)	
19		Asamura et al. (2008)	
20.3		Gamero et al. (2000)	

STR Review Article

Supplemental Figures 1-24



Supplemental Figure 1. Annotation of the GRCh38 Sequence at the D1S1656 Locus

STR Review Article

Supplemental Table 2

Excel File with a tab for each locus, details flanking region polymorphisms

D7S820

Reference SNP	Chromosome	Chromosome Position	Distance from STR repeat	RefSNP alleles	Minor Allele	Minor Allele Frequency	Minor Allele Count
		GRCh 38	Blue text if < 150 bp, bolded if > 5 % frequency	Forward strand unless otherwise noted	Second most frequent allele	Bolded if > 5 % frequency	
rs540349249	7	84159732	494	C/T	T	0.0004	2
rs149464212	7	84159838-57	388	-/ATGTGAACAATTGTGTTCTA	-	0.0104	52
rs58675984	7	84160017	**209**	G/T	G	**0.0802**	401
rs59186128	7	84160110	**116**	C/T	T	**0.0758**	379
rs7786079	7	84160161	**65**	A/C	C	**0.0798**	399
rs7789995	7	84160204	**22**	A/T	T	**0.0698**	349
Repeat Region	7	84160226-84160277 (REV)	*13 repeats*				
rs16887642	7	84160286	**9**	A/G	A	**0.1406**	704
rs141022647	7	84160382	105	G/T	G	0.0074	37
rs150246249	7	84160452	175	A/G	A	0.0006	3
rs554238483	7	84160498	221	A/G	A	0.0006	3
rs192610146	7	84160504	227	A/G	A	0.0044	22
rs533853989	7	84160553	276	A/T	A	0.0006	3
rs563661578	7	84160565	288	C/T	C	0.0004	2
rs544030261	7	84160606	329	C/T	T	0.0004	2
rs7806601	7	84160645	**368**	A/G	A	**0.0802**	401
rs188794547	7	84160704	427	C/T	T	0.0014	7

STR Nomenclature

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

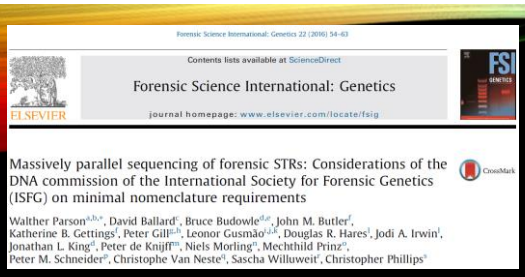
Contents lists available at ScienceDirect

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]



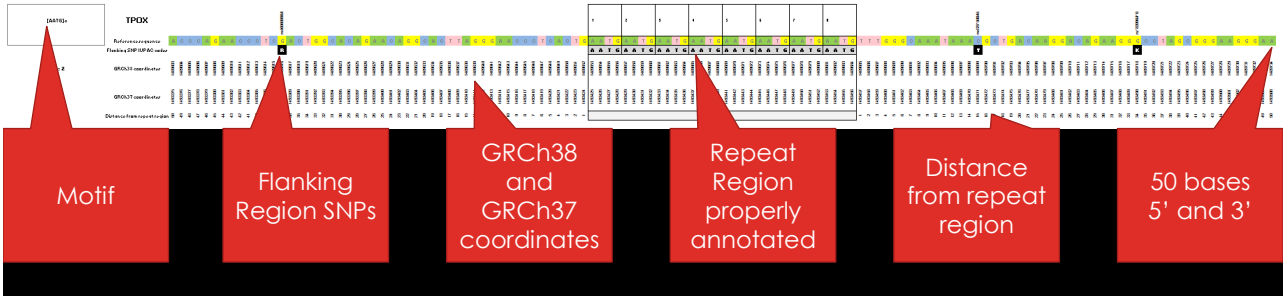
STR Nomenclature

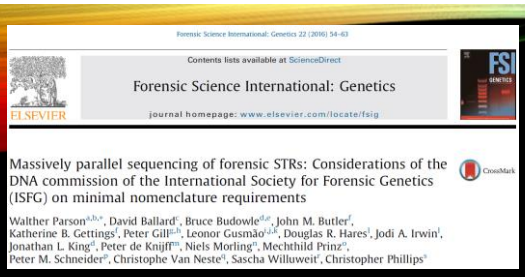
Principal guidelines:

1. STR sequences aligned to the genome reference sequence.
2. Variant annotation (systematic description of genome sequence differences between individuals), use locus identifiers and variant reporting methods applied in 1000 Genomes and dbSNP databases.

STR sequence template file summarised STR sequence alignments and annotations.

STR Nomenclature




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]

STR Nomenclature

It was recognised at the time of publication that sequence variation in STRs presents particular challenges, requiring care and a period of time to compile sufficiently detailed sequencing data.

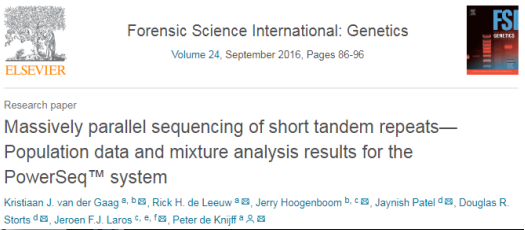
NGS Population Studies 2016-2017


Forensic Science International: Genetics
Volume 21, March 2016, Pages 15-21

Research paper

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

Katherine Butler Gettings [a], Kevin M. Kiesler [a], Seth A. Faith [b], Elizabeth Montano [c], Christine H. Baker [c], Brian A. Young [c], Richard A. Guerrieri [c], Peter M. Vallone [a]


Forensic Science International: Genetics
Volume 24, September 2016, Pages 86-96

Research paper

Massively parallel sequencing of short tandem repeats—Population data and mixture analysis results for the PowerSeq™ system

Kristiaan J. van der Gaag [a, b], Rick H. de Leeuw [a], Jerry Hoogenboom [b, c], Jaynish Patel [d], Douglas R. Storts [d], Jeroen F.J. Laros [c, e, f], Peter de Knijff [a]


Forensic Science International: Genetics
Volume 21, March 2016, Pages 68-75

Research paper

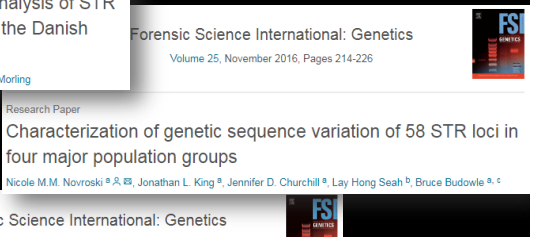
Introduction of the Python script STRinNGS for analysis of STR regions in FASTQ or BAM files and expansion of the Danish STR sequence database to 11 STRs

Susanne L. Friis [1], Anders Buchard [1], Eszter Rockenbauer, Claus Børsting, Niels Morling


Forensic Science International: Genetics
Volume 24, September 2016, Pages 18-23

Genetic analysis of the Yavapai Native Americans from West-Central Arizona using the Illumina MiSeq FGx™ forensic genomics system

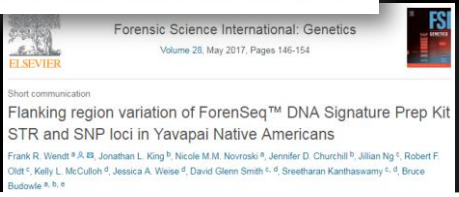
Frank R. Wendt [a], Jennifer D. Churchill [a], Nicole M.M. Novroski [a], Jonathan L. King [a], Jillian Ng [b], Robert F. Oldt [b], Kelly L. McCulloh [c], Jessica A. Weise [c], David Glenn Smith [b, c], Sreetharan Kanthaswamy [b, c], Bruce Budowle [a, d]


Forensic Science International: Genetics
Volume 25, November 2016, Pages 214-226

Research Paper

Characterization of genetic sequence variation of 58 STR loci in four major population groups

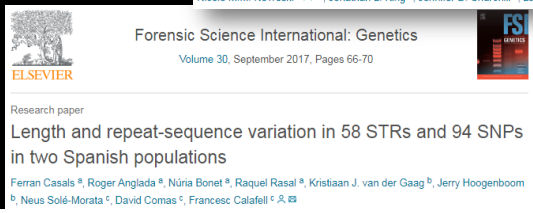
Nicole M.M. Novroski [a], Jonathan L. King [a], Jennifer D. Churchill [a], Lay Hong Seah [b], Bruce Budowle [a, c]


Forensic Science International: Genetics
Volume 28, May 2017, Pages 146-154

Short communication

Flanking region variation of ForenSeq™ DNA Signature Prep Kit STR and SNP loci in Yavapai Native Americans

Frank R. Wendt [a], Jonathan L. King [b], Nicole M.M. Novroski [a], Jennifer D. Churchill [b], Jillian Ng [c], Robert F. Oldt [c], Kelly L. McCulloh [d], Jessica A. Weise [d], David Glenn Smith [c, d], Sreetharan Kanthaswamy [c, d], Bruce Budowle [a, b, e]

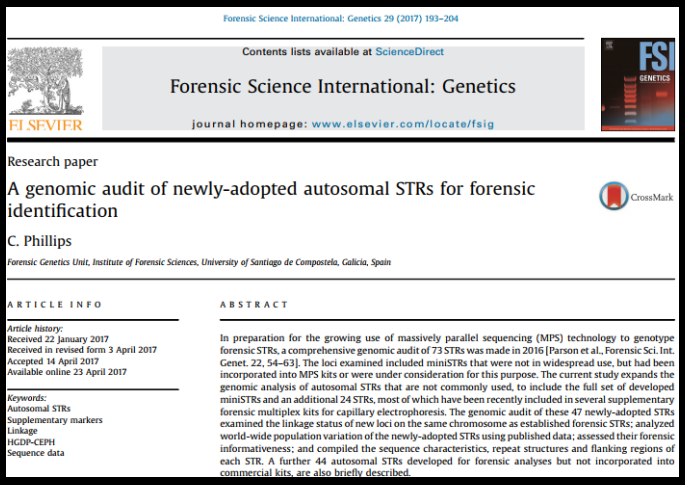

Forensic Science International: Genetics
Volume 30, September 2017, Pages 66-70

Research paper

Length and repeat-sequence variation in 58 STRs and 94 SNPs in two Spanish populations

Ferran Casals [a], Roger Anglada [a], Núria Bonet [a], Raquel Rasal [a], Kristiaan J. van der Gaag [b], Jerry Hoogenboom [b], Neus Solé-Morata [c], David Comas [c], Francesc Calafell [c]

STR Nomenclature



Forensic Science International: Genetics 29 (2017) 193–204

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

A genomic audit of newly-adopted autosomal STRs for forensic identification

C. Phillips

Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Galicia, Spain

ARTICLE INFO

Article history:
Received 22 January 2017
Received in revised form 3 April 2017
Accepted 14 April 2017
Available online 23 April 2017

Keywords:
Autosomal STRs
Supplementary markers
Linkage
HGDP-CEPH
Sequence data

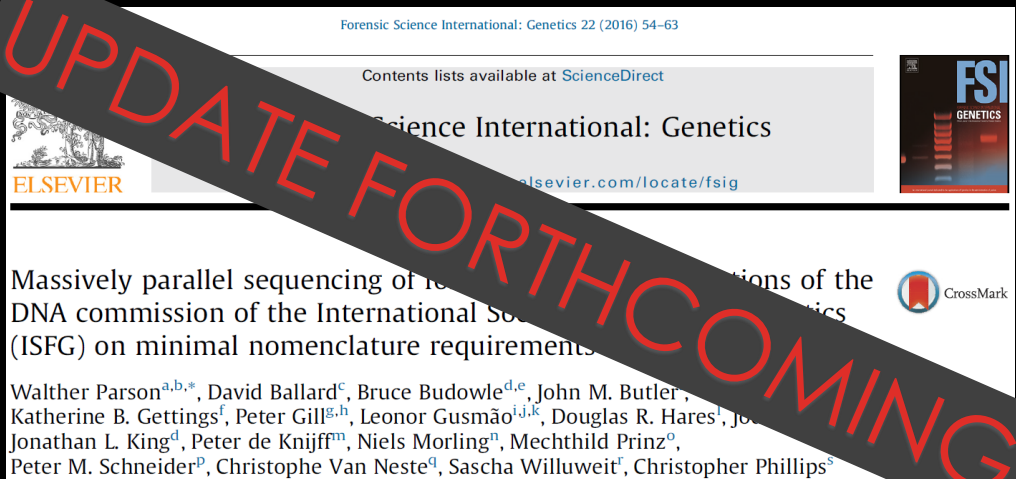
ABSTRACT

In preparation for the growing use of massively parallel sequencing (MPS) technology to genotype forensic STRs, a comprehensive genomic audit of 73 STRs was made in 2016 [Parson et al., Forensic Sci. Int. Genet. 22, 54–63]. The loci examined included miniSTRs that were not in widespread use, but had been incorporated into MPS kits or were under consideration for this purpose. The current study expands the genomic analysis of autosomal STRs that are not commonly used, to include the full set of developed miniSTRs and an additional 24 STRs, most of which have been recently included in several supplementary forensic multiplex kits for capillary electrophoresis. The genomic audit of these 47 newly-adopted STRs examined the linkage status of new loci on the same chromosome as established forensic STRs; analyzed world-wide population variation of the newly-adopted STRs using published data; assessed their forensic informativeness; and compiled the sequence characteristics, repeat structures and flanking regions of each STR. A further 44 autosomal STRs developed for forensic analyses but not incorporated into commercial kits, are also briefly described.

- Characterizes 47 newly adopted autosomal STRs and examines linkage status
- Briefly describes 44 additional autosomal STRs
- Eliminate ambiguity prior to assay development/publication

STR Nomenclature

UPDATE FORTHCOMING



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

Contents lists available at ScienceDirect

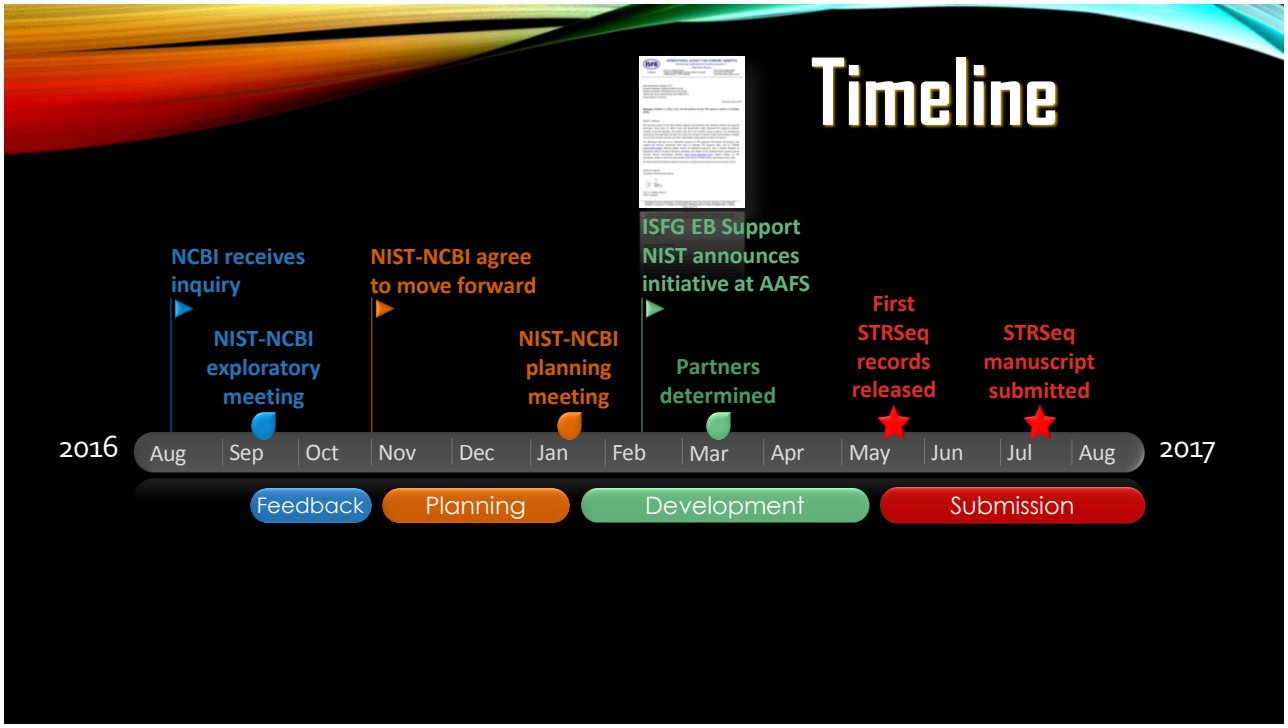
ience International: Genetics

lsevier.com/locate/fsig

Massively parallel sequencing of ... ions of the DNA commission of the International So... ics (ISFG) on minimal nomenclature requirement...

Walther Parson[a,b,*], David Ballard[c], Bruce Budowle[d,e], John M. Butler[...], Katherine B. Gettings[f], Peter Gill[g,h], Leonor Gusmão[i,j,k], Douglas R. Hares[l], Jo... 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]


STRSeq 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.

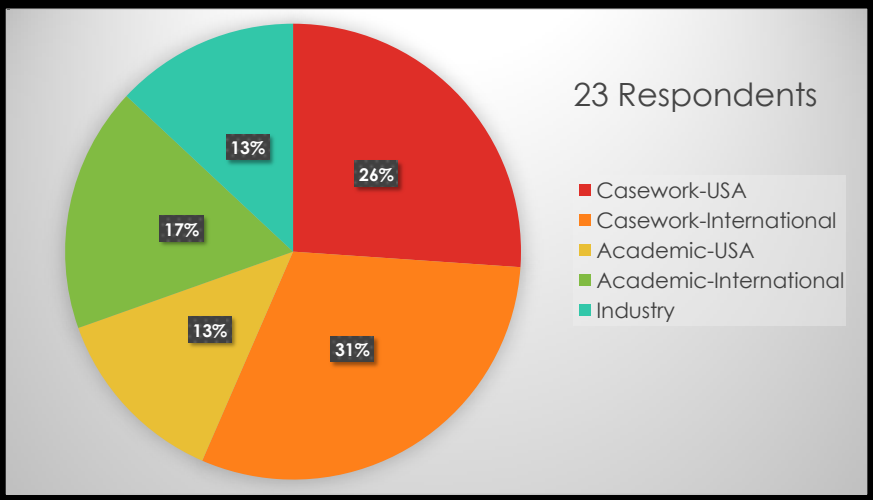

Timeline
NCBI receives inquiry
NIST-NCBI exploratory meeting
NIST-NCBI agree to move forward
NIST-NCBI planning meeting
ISFG EB Support NIST announces initiative at AAFS
Partners determined
First STRSeq records released
STRSeq manuscript submitted
2016
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
2017
Feedback
Planning
Development
Submission

Feedback

Three questions sent to 40 laboratories interested in STR sequencing:

1. Are you sequencing forensic STRs? For what purpose?
2. Would you use forensic STR reference sequences? How?
3. How would you want to access this information?

Feedback



Feedback

Are you sequencing forensic STRs? For what purpose?

Casework
- Validating for casework and/or missing persons
- Interlaboratory studies and/or beta testing

Academia
- Population data
- Degraded DNA and/or mixture studies

Industry
- Assay and/or software development

Feedback

Would you use forensic STR reference sequences? How?

Yes
- Standardize reporting
- International databasing
- Flanking region variants
  - imputation
- Bioinformatics
- Searchable repository
- ~~Frequency data~~
- ~~Certified controls~~

No
- Direct comparisons
- Local databasing
- In-house nomenclature

Feedback

How would you want to access this information?

- **Manufacturer software**
- Database query by sequence, allele (size), or locus
- Public, official database
- Download FASTA
- Online naming tool
- N/A already using an in-house database
- ~~Database with frequencies~~

Planning

1. NIST creates record for each unique sequence
2. Initial data are NIST population samples
3. Non redundant records; number of records per locus varies
4. Records include flanking regions with high confidence sequence
5. Records include length-based allele designations determined by CE
6. Records can expand for future additional flank
7. Records organized into BioProject for improved access

Development

BioProject hierarchy

Record format

Partner labs

Partners-Roles

UNT HEALTH SCIENCE CENTER

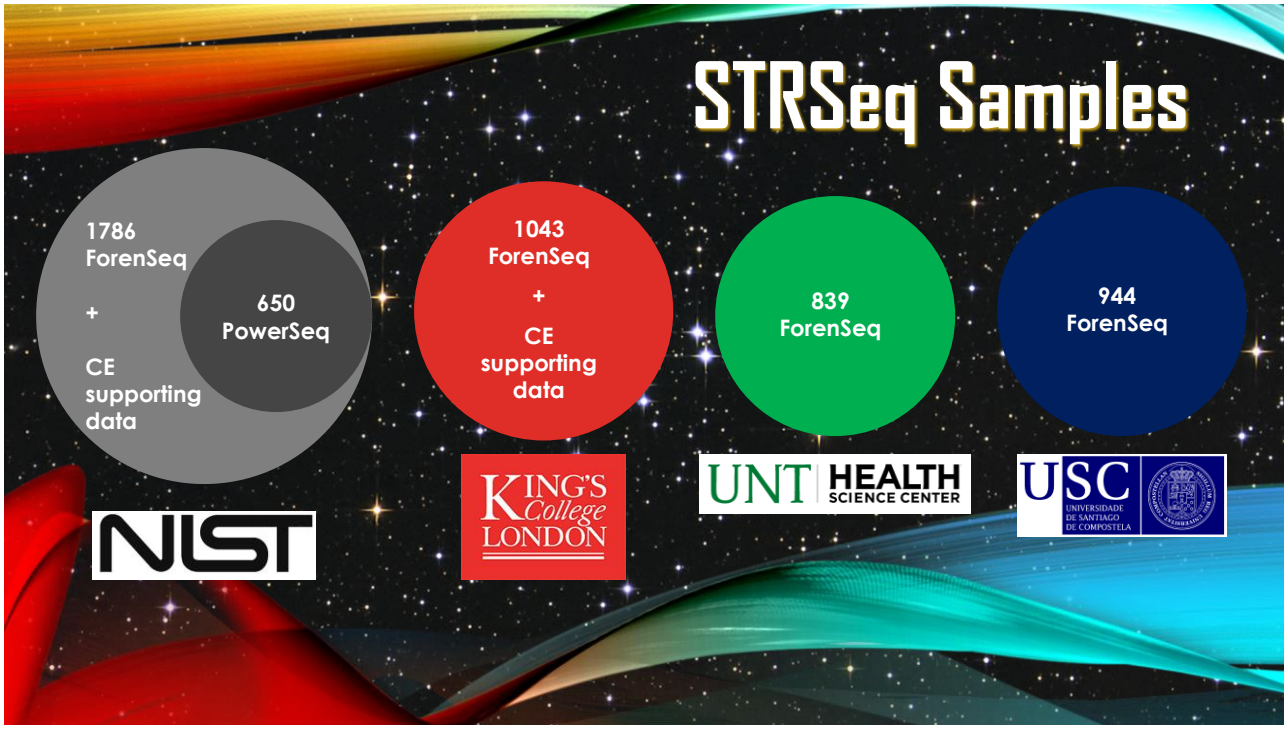
NIST
Population Samples
Project Coordination
Record Submission

KING'S College LONDON

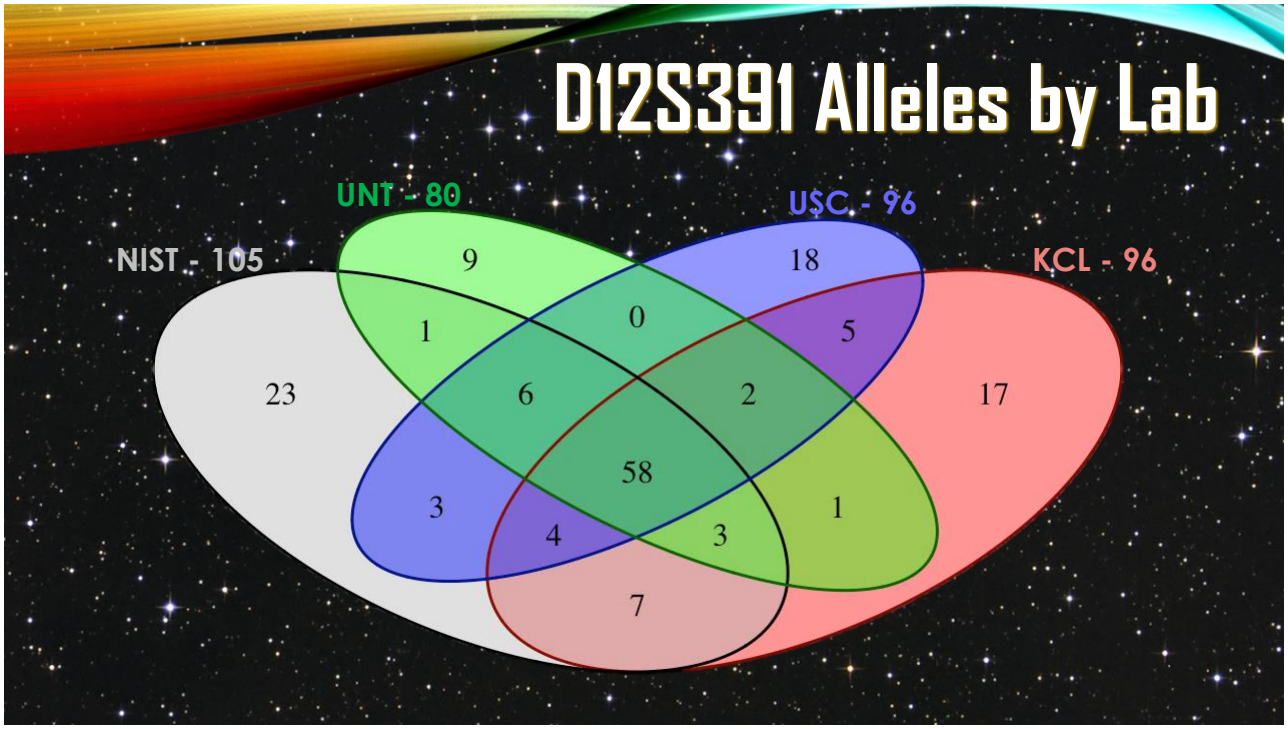
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
Population Data
Project Input

GMI TIROL
Project Input
STRidER Integration

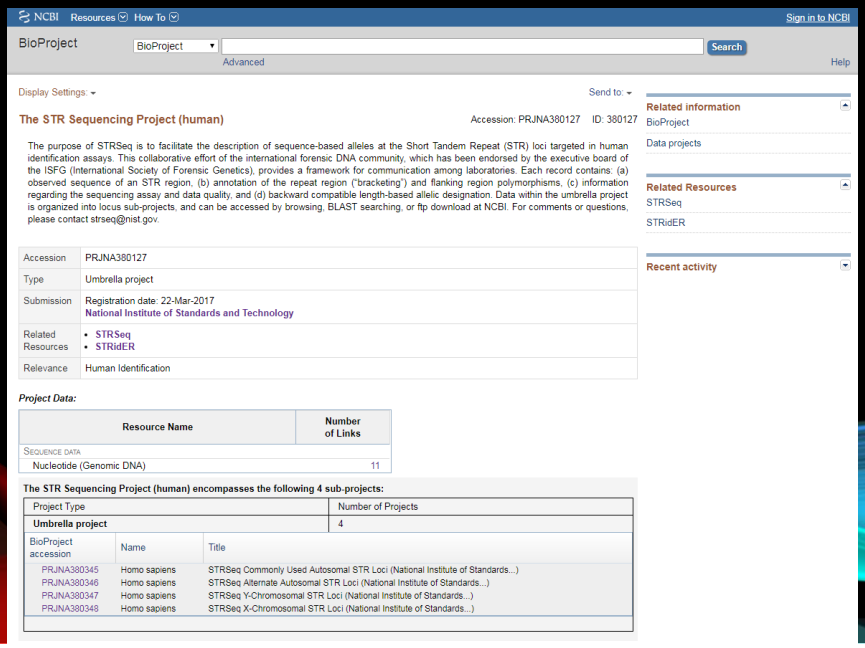
NCBI
Project Input
Hosting

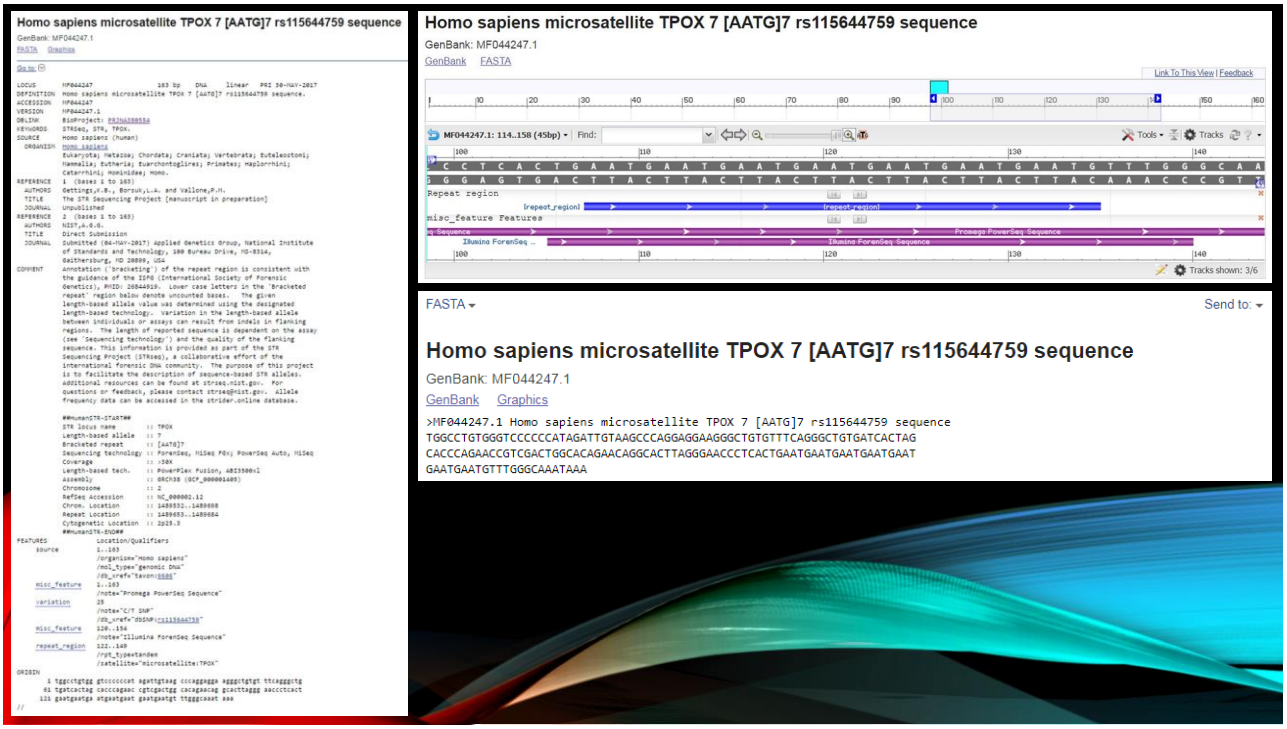

STRSeq Samples
1786 ForenSeq + CE supporting data
650 PowerSeq
1043 ForenSeq + CE supporting data
839 ForenSeq
944 ForenSeq
NIST
KING'S College LONDON
UNT HEALTH SCIENCE CENTER
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

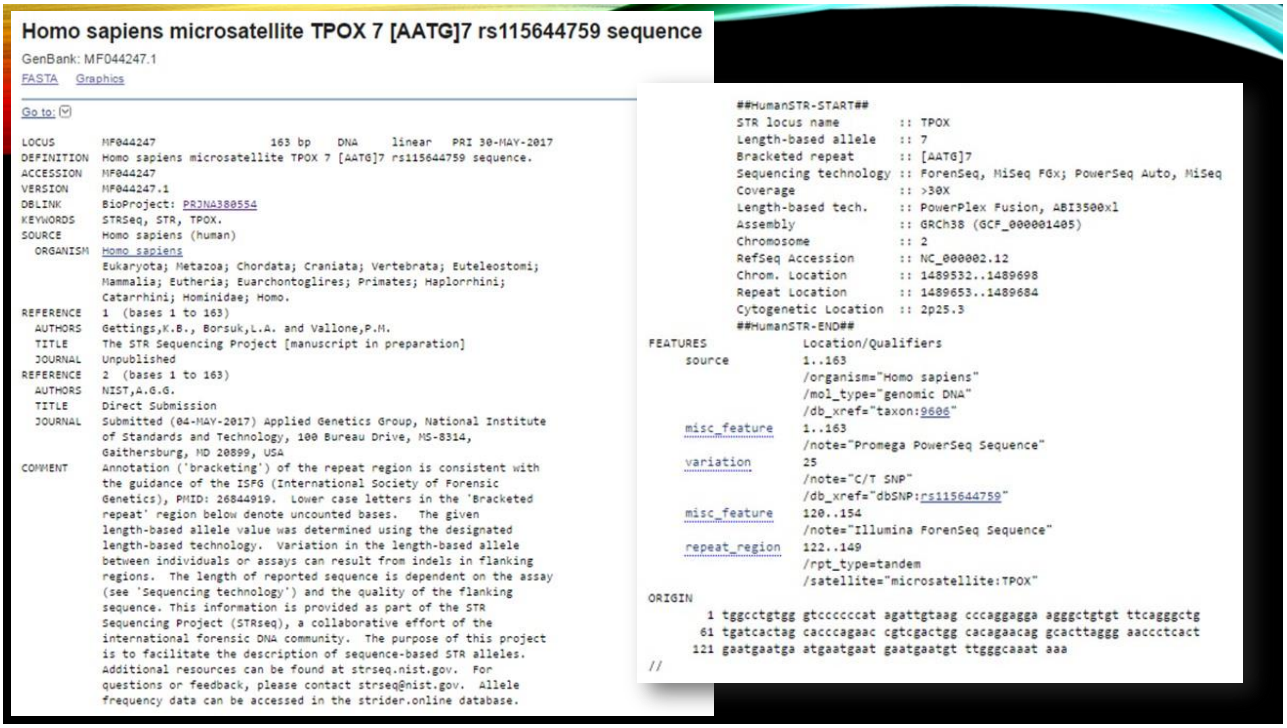

STRSeq Samples
Aggregate alleles from 4612 samples

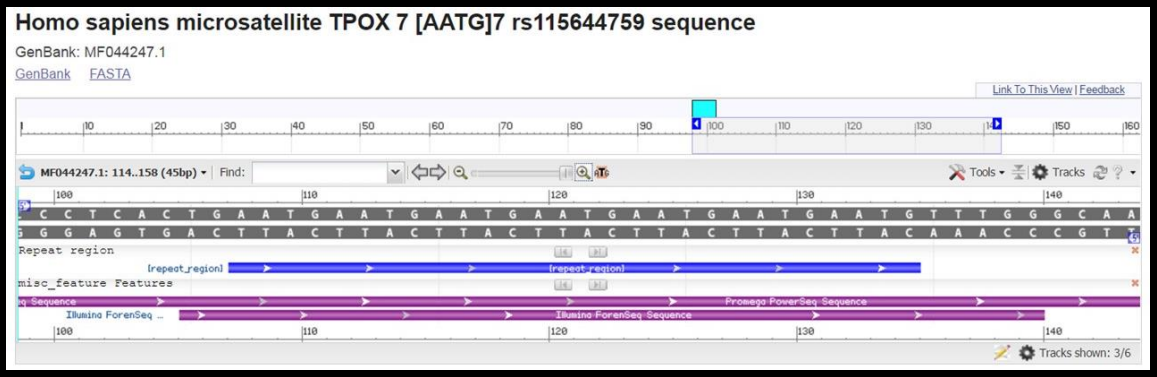


https://www.ncbi.nlm.nih.gov/bioproject/380127




Homo sapiens microsatellite TPOX 7 [AATG]7 rs115644759 sequence
GenBank: MF044247.1
GenBank FASTA
Link To This View | Feedback
Repeat region
misc_feature Features
Illumina ForenSeq
Promega PowerSeq Sequence
Tracks shown: 3/6
FASTA
Send to:
Homo sapiens microsatellite TPOX 7 [AATG]7 rs115644759 sequence
GenBank: MF044247.1
GenBank Graphics
>MF044247.1 Homo sapiens microsatellite TPOX 7 [AATG]7 rs115644759 sequence
TGGCCTGTGGGTCCCCCCATAGATTGTAAGCCCAGGAGGAAGGGCTGTGTTTCAGGGCTGTGATCACTAG
CACCCAGAACCGTCGACTGGCACAGAACAGGCACTTAGGGAACCCTCACTGAATGAATGAATGAATGAAT
GAATGAATGTTTGGGCAAATAAA


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
//



STRSeq in Bioinformatics

Standalone and API BLAST (Application Programming Interface)

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

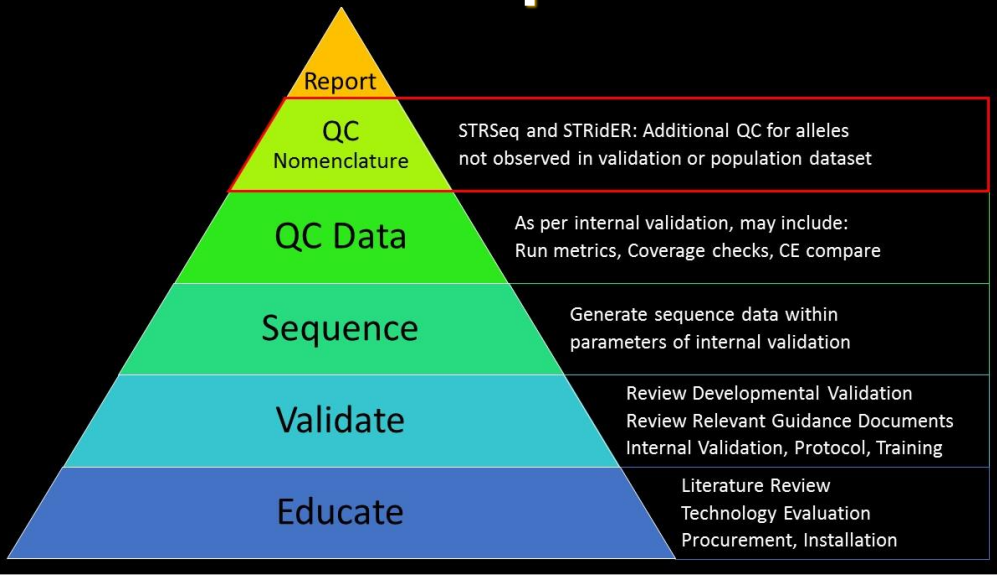
Introduction

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
Link To This View | Feedback
MF044247.1: 114..158 (45bp)
Find:
Tools
Tracks
Repeat region
misc_feature Features

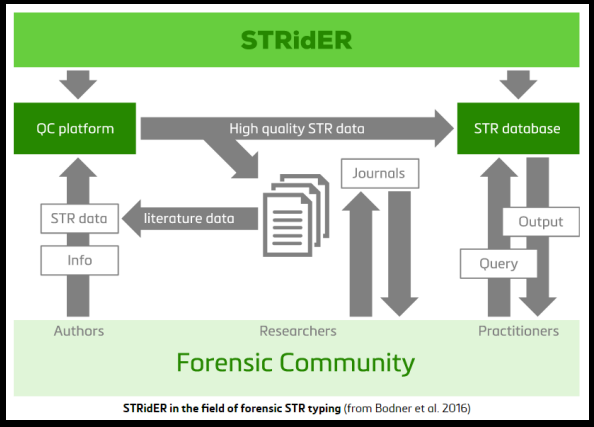
STRSeq in Casework



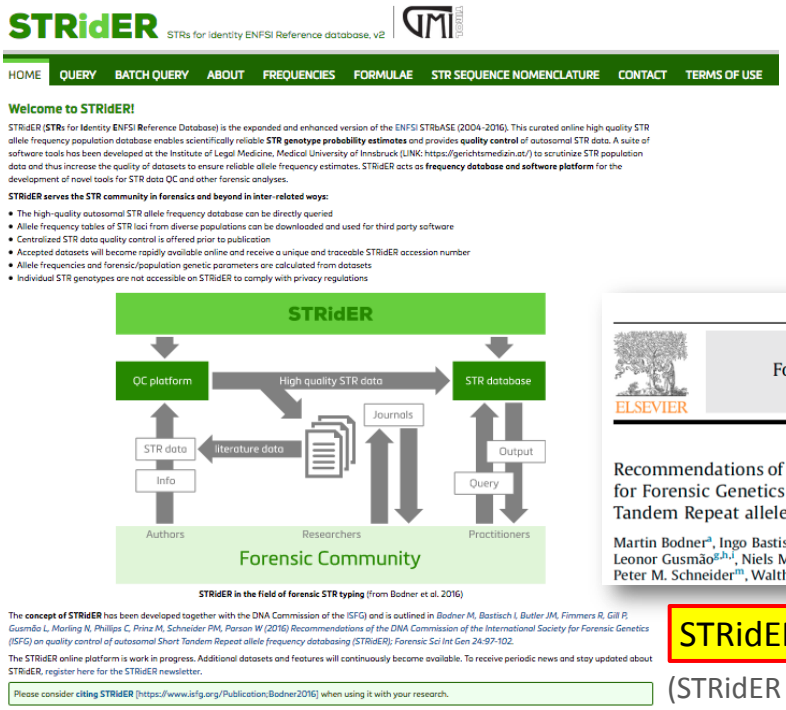
STRSeq in Population Data

STRSeq

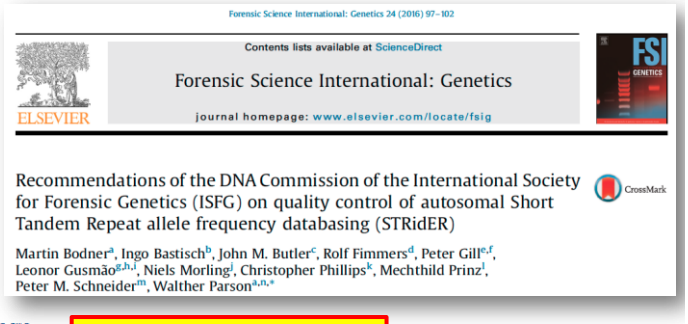
Collaboration in QC and exchange of data



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



new URL
https://strider.online/



STRidER newsletter

(STRidER slides courtesy of Dr. Walther Parson)

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)

CrossMark

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, pop STR, popAffiliator, ALLST*R); **important element of QC**

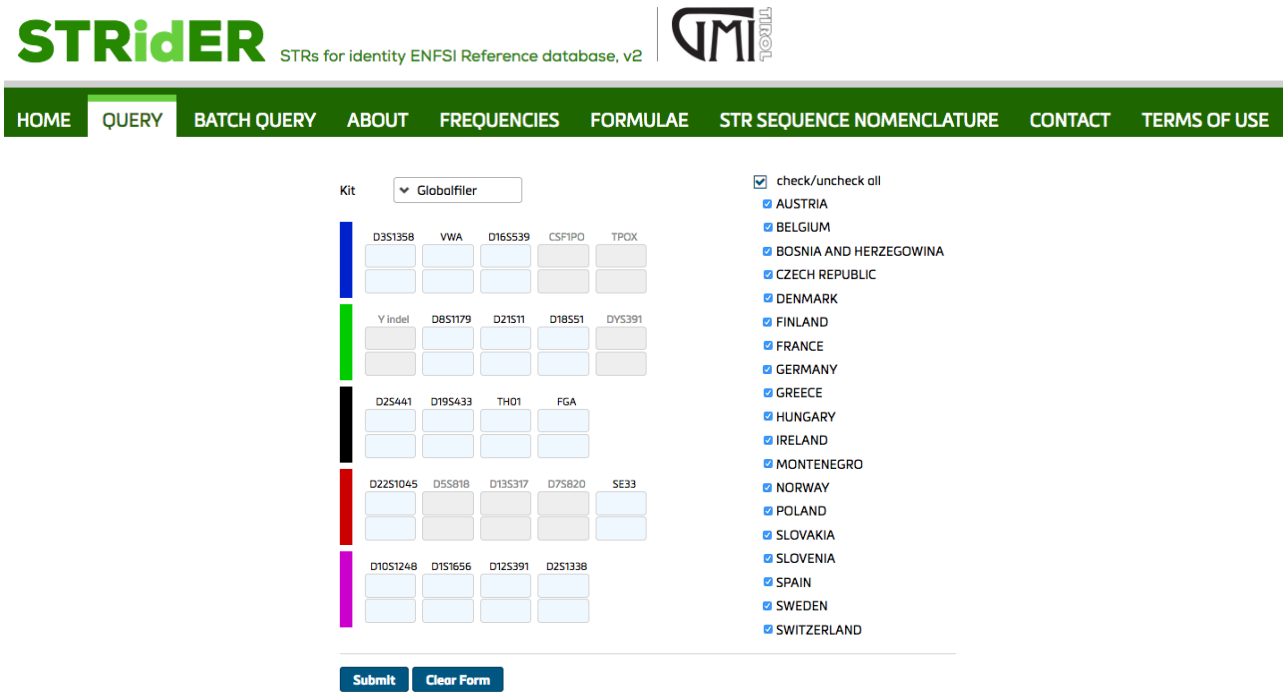
II) Rationale, concept and workflow of **QC** via STRidER

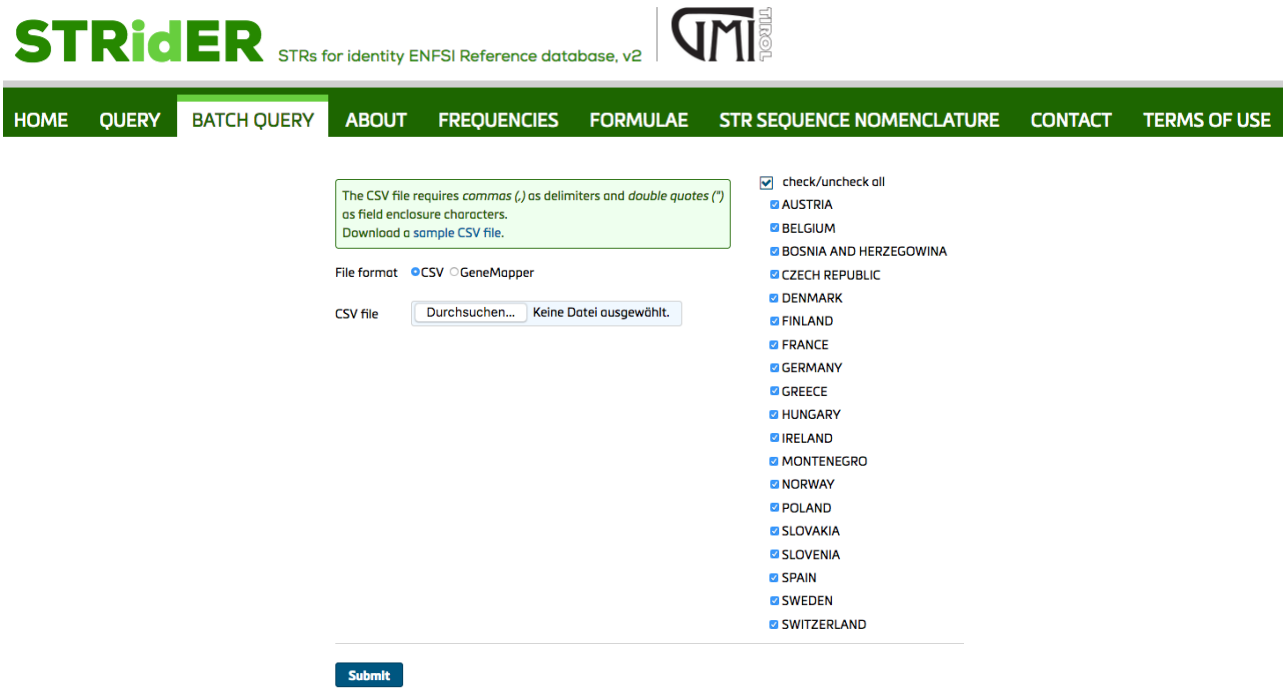
III) Benefits to forensic and other scientific community

IV) Transparency, traceability and protection of data

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

STRidER


STRidER
STRs for identity ENFSI Reference database, v2
HOME
QUERY
BATCH QUERY
ABOUT
FREQUENCIES
FORMULAE
STR SEQUENCE NOMENCLATURE
CONTACT
TERMS OF USE
Kit
Globalfiler
D3S1358
VWA
D16S539
CSF1PO
TPOX
Y indel
D8S1179
D21S11
D18S51
DYS391
D2S441
D19S433
TH01
FGA
D22S1045
D5S818
D13S317
D7S820
SE33
D10S1248
D1S1656
D12S391
D2S1338
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
Submit
Clear Form


STRidER
STRs for identity ENFSI Reference database, v2
HOME
QUERY
BATCH QUERY
ABOUT
FREQUENCIES
FORMULAE
STR SEQUENCE NOMENCLATURE
CONTACT
TERMS OF USE
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
Submit

STRidER STRs for identity ENFSI Reference database, v2

HOME | QUERY | BATCH QUERY | ABOUT | FREQUENCIES | FORMULAE | STR SEQUENCE NOMENCLATURE | CONTACT | TERMS OF USE

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
	222	206	171	200	200	230	208	662	208	224	304	200	202	206	247	207	449	424
11								7.5529e-4										
12									4.8077e-3									
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
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
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
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
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
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
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
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
21					2.5000e-3	6.5217e-3		7.5529e-4	2.4038e-3	2.2321e-3						4.8309e-3		2.3585e-3

TH01

Allele	AUSTRIA	BELGIUM	BOSNIA AND HERZEGOWINA	CZECH REPUBLIC	DENMARK	FINLAND	FRANCE	GERMANY	GREECE	HUNGARY	IRELAND	MONTENEGRO	NORWAY	POLAND	SLOVAKIA	SLOVENIA	SPAIN	SWEDEN
	222	206	171	200	200	230	208	662	208	224	304	200	202	206	247	207	454	425

STRidER STRs for identity ENFSI Reference database, v2

HOME | QUERY | BATCH QUERY | ABOUT | FREQUENCIES | FORMULAE | STR SEQUENCE NOMENCLATURE | CONTACT | TERMS OF USE

Formulae

Actual matching probability

$P_m = 2p_ip_j$ 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.

STRidER STRs for identity ENFSI Reference database, v2

HOME | QUERY | BATCH QUERY | ABOUT | FREQUENCIES | FORMULAE | STR SEQUENCE NOMENCLATURE | CONTACT | TERMS OF USE

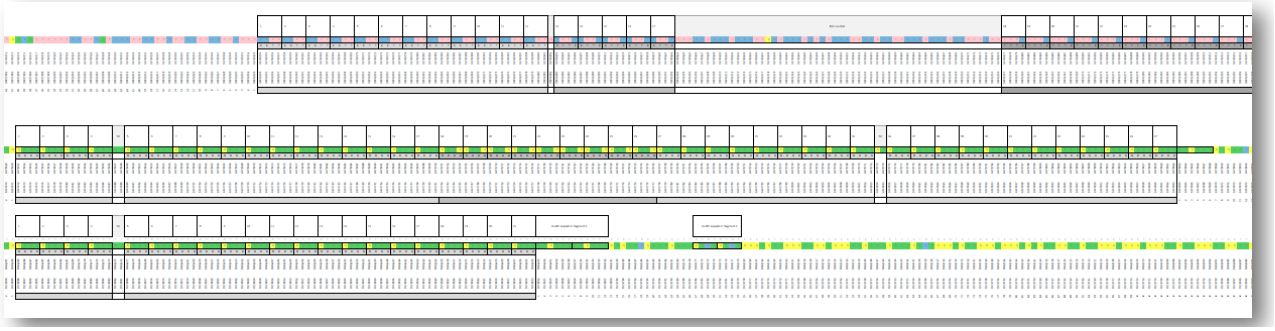
STR Sequence Nomenclature

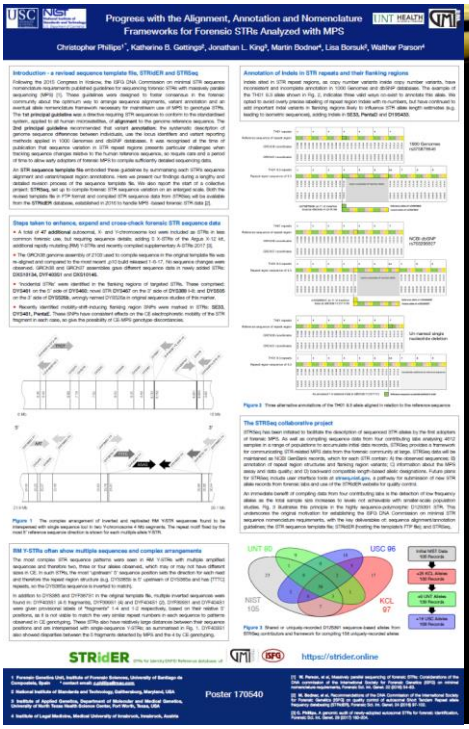
The file 'STR Sequence Structure (Supplementary File S1)' is an updated set of forensic STR sequences that accompany the article:

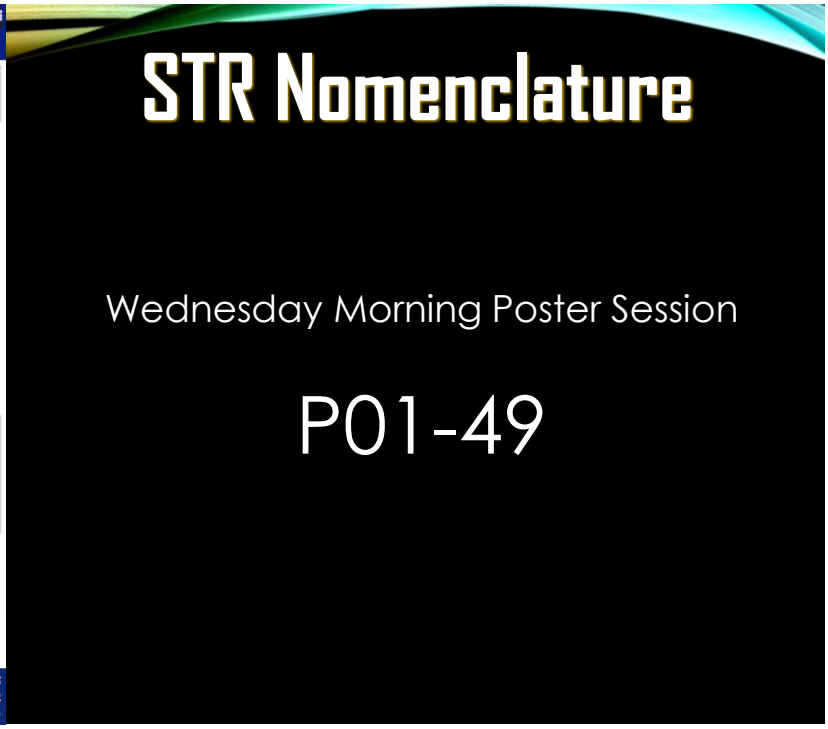
Parson W, Ballard D, Budowle B, Butler JM, Gettings KB, Gill P, Gusmão L, Hares DR, Irwin JA, King JL, de Knijff P, Morling N, Prinz M, Schneider PM, Van Neste C, Willuweit S, Phillips C: **Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements.** Forensic Science International Genetics 2016, 22: 54-63 (doi: 10.1016/j.fsigen.2016.01.009; available at http://www.isfg.org/Publication;Parson2016).

Major changes to this file are currently ongoing. When the review process is finished, a new version of Supplementary File S1 containing updated information will be available for download here. To stay updated about STRidER, register here for the STRidER newsletter.

The updates since the last version are:








Somewhere something incredible is waiting to be known.
Carl Sagan
Acknowledgements
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
Dr. Chris Phillips
NIST
Lisa Borsuk
Dr. Peter Vallone
Dr. John Butler
UNT HEALTH SCIENCE CENTER
Jonathan King
Dr. Bruce Budowle
UMIT TIROL
Drs. Martin Bodner, Walther Parson
KING'S College LONDON
Dr. David Ballard
Laurence Devesse
NCBI
Drs. Lori Black, Melissa Landrum, Ilene Mizrachi, Kim Pruitt, George Riley, Steve Sherry
Thanks to Labs who provided feedback for STRSeq,
the ISFG Commission on MPS of STRs, and the ISFG Commission on STRidER

29 August 2017

Workshop #10

STRBase & Revisions Planned

John M. Butler, Ph.D.
Special Programs Office
U.S. National Institute of Standards and Technology

Lisa Borsuk, M.S.
Applied Genetics Group
U.S. National Institute of Standards and Technology

Information Gathering and Sharing

- **We live in the information age and need to share what we learn as scientists with others**
- Sharing information impacts validation of techniques, which impact court use of the technique
- DNA is often referred to as the "gold standard" in forensic science because of the scientific studies performed and information sharing that has occurred
- **You need a good library (information collection) to be successful in developing any scientific discipline**
- Knowing the literature provides a solid foundation for research and future work

Forensic DNA Library

Books Located in John Butler's NIST Office

Collected >300 books on topics related to forensic DNA analysis and forensic science in general

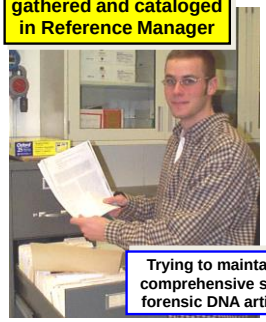
Initially funded from 2002-2007 by some award money received

Tracking Updates in the Literature is Time Consuming

Creating a Reference Collection

>9,000 references
gathered and cataloged in Reference Manager

Numbered by order of entry in database



Trying to maintain a comprehensive set of forensic DNA articles

- My forensic DNA reference collection began while I was in graduate school
  - Continued over the years with the help of student interns like Christian Ruitberg shown here
- **Mostly printed copies of articles are stored**
  - has increasing become digital (this part is not as well organized)

Organization of Digital Files

- Downloaded pdf files are stored in multiple folders
- Files are named by first author last name, year of publication, journal name, and title or subject summary

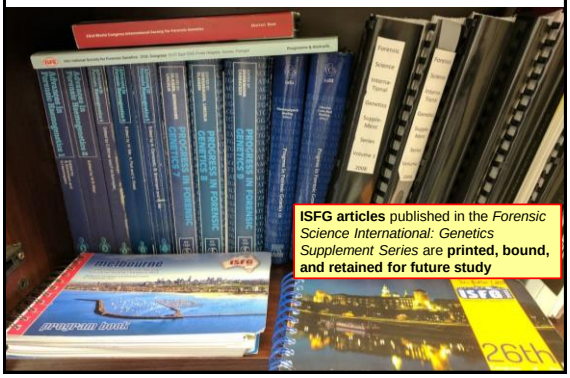
Some example file names

Gertner 2012 Criminal Justice - NAS report challenge to the courts.pdf
Gill 1996 ULM - STR interpretation using inferential logic.pdf
Gjertson 2007 FSI Genetics - ISFG recommendations on paternity biostatistics.pdf
Grochau 2017 Accred Qual Assur - current ISO 17025 accreditation use.pdf
Grover 2017 ULM - FlexPlex27 ANDE rapid DNA multiplex.pdf
Hanson 2005 Legal Med - Y-STRs annotated.pdf
Hansson 2017 Int J Legal Med - degradation and het balance by simulation.pdf
Hares 2012a FSI Gen - expansion of new CODIS loci.pdf
Hares 2012b FSI Gen - addendum to expansion of new CODIS loci.pdf
Hares 2015 FSI Gen - selection of new CODIS loci.pdf
Harralson 2015 - survey of forensic handwriting examination research.pdf
Hebrard 2013 Encyc Forensic Sci 2nd ed - history of forensic sciences.pdf
Helsel 2005 Environ Sci & Tech - more than obvious - interpreting nondetect data.pdf
Henke 2007 Clin Lab - validation of new STR multiplex.pdf

Physical Collection of Notes from Meetings Attended

Have retained detailed notes from every meeting attended in the past 25 years

My Copies of the ISFG Meeting Proceedings



ISFG articles published in the *Forensic Science International: Genetics Supplement Series* are **printed, bound, and retained for future study**

Good information input improves output…
Some Fruits of a Good Literature Collection

Review Articles

J Forensic Sci, March 2006, Vol. 51, No. 2
doi:10.1111/j.1556-4029.2006.00046.x
Available online at: www.blackwell-synergy.com

John M. Butler,[1] Ph.D.

Genetics and Genomics of Core Short Tandem Repeat Loci Used in Human Identity Testing

Anal. Chem. **2007**, *79*, 4305–4304

Analytical Chemistry **(June 15, 2007 issue)**

Forensic Science

T. A. Brettell*

Department of Chemical and Physical Sciences, Cedar Crest College, 100 College Drive, Allentown, Pennsylvania 18104-6196

J. M. Butler

Biochemical Science Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8311

J. R. Almirall

Department of Chemistry and Biochemistry and International Forensic Research Institute, Florida International University, University Park, Miami, Florida 33199

Textbooks

And a Useful Reference Website…
NIST **STRBase** Website

Serving the Forensic DNA Community for 20 Years

Short Tandem Repeat DNA Internet DataBase

NIST Standard Reference Database SRD 130 [Recent Updates]

Serving the forensic DNA and human identity testing communities for 20 years... **These data are intended to benefit research and application of short tandem repeat DNA markers to human identity testing. The authors are solely responsible for the information herein.**

This database has been accessed >500,000 times since 10/02/97.

Created by ***John M. Butler*** *and Dennis J. Reeder (NIST), with invaluable help from Jan Redman, Christian Ruitberg and Michael Tung*
Site creators' curriculum vitaes available using links above.

Partial support for the design and maintenance of this website was previously provided by The National Institute of Justice through the NIST Forensic Science Program Office.

General Information

- Purpose of STRBase/NAR 2001 Paper describing STRBase/Overview Presentation
- Publications and Presentations from NIST Human Identity Project Team
- NIJ-Funded Projects
- Training Materials
- Links to other web sites
- Glossary of commonly used terms

http://www.cstl.nist.gov/strbase/

New Version of NIST STRBase

Server moved (from cstl.nist.gov) in Spring 2017

NIST National Institute of Standards and Technology U.S. Department of Commerce — STRBase (SRD-130)

HOME | UPDATES | INFO | Human STRs | Other Markers | Resources

Short Tandem Repeat DNA Internet DataBase

NIST Standard Reference Database SRD 130 [Recent Updates]

Serving the forensic DNA and human identity testing communities for 20 years... **These data are intended to benefit research and application of short tandem repeat DNA markers to human identity testing. The authors are solely responsible for the information herein.**

This database has been accessed >500,000 times since 10/02/97.

Created by ***John M. Butler*** *and Dennis J. Reeder (NIST), with invaluable help from Jan Redman, Christian Ruitberg and Michael Tung*
Site creators' curriculum vitaes available using links above.

Partial support for the design and maintenance of this website was previously provided by The National Institute of Justice through the NIST Forensic Science Program Office.

General Information

- Purpose of STRBase/NAR 2001 Paper describing STRBase/Overview Presentation
- Publications and Presentations from NIST Human Identity Project Team
- NIJ-Funded Projects
- Training Materials
- Links to other web sites
- Glossary of commonly used terms

Additional changes are planned besides a new URL …

http://strbase.nist.gov

Publications regarding STRBase

CURRENT EVENTS

STRBase: A Short Tandem Repeat DNA Database

By John M. Butler and Dennis J. Reeder*
National Institute of Standards and Technology, Biotechnology Division, Gaithersburg, MD

Profiles in DNA **(Promega) 1997; 1(2): 10**

320–322 *Nucleic Acids Research*, 2001, Vol. 29, No. 1

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

Received August 31, 2000; Accepted September 17, 2000

Nucleic Acids Research **(database issue) 2001; 29(1): 320-322**

ELSEVIER

Available online at www.sciencedirect.com
ScienceDirect

Forensic Science International: Genetics Supplement Series 1 (2008) 97–99

FSI GENETICS

www.elsevier.com/locate/FSIGSS

Research Articles

New resources for the forensic genetics community available on the NIST STRBase website☆

John M. Butler*

National Institute of Standards and Technology, Biochemical Science Division, 100 Bureau Dr, MS 8311, Gaithersburg, MD 20899-8311, USA

Received 16 August 2007; accepted 7 October 2007

Forensic Science International: Genetics Supplement Series **(ISFG 2007 Meeting Proceedings) 2008; 1: 97-99**

A Brief History of the STRBase Website

- Initial information was collected on STR markers while working on my PhD dissertation in 1993-1995
- Started a review article in 1996 while a NIST postdoc but wanted to create a dynamic rather than an out-of-date resource
- Created hundreds of individual web pages that were hyperlinked together
- **Website launched in July 1997** (discussed at ISHI 1997)
- Became a NIST Standard Reference Database (SRD 130) because of its high visibility
- **Until early 2017, a single person (JMB) used an HTML editor to update the website**
- **URL was recently changed to http://strbase.nist.gov**

Core STRBase Information

Forensic STR Information

- STRs101: Brief Introduction to STRs
- *Core Loci*: FBI CODIS Core STR Loci and European Core Loci
- STR Fact Sheets (observed alleles and PCR product sizes)
- **Multiplex STR kits**
- Sequence Information (annotated)
- Variant Allele Reports
- Tri-Allelic Patterns
- Mutation Rates for Common Loci
- Published PCR primers
- Y-chromosome STRs
- Low-template DNA Information
- Mixture Interpretation
- Kinship Analysis
- miniSTRs (short amplicons)
- Null Alleles - discordance observed between STR kits
- STR Reference List - *now 3687 references*

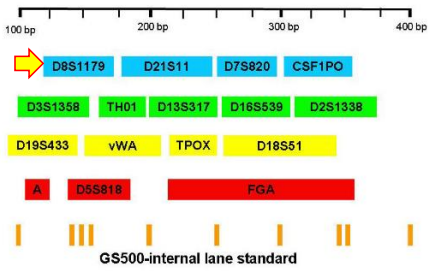
Multiplex STR Kit Information

STR Kits from Applied Biosystems (Foster City, CA)

- AmpFℓSTR **Identifiler**: D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, D2S1338, D19S433, VWA, TPOX, D18S51, D5S818, FGA, Amelogenin
- AmpFℓSTR **Identifiler Direct**: D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, D2S1338, D19S433, VWA, TPOX, D18S51, D5S818, FGA, Amelogenin
- AmpFℓSTR **Identifiler Plus**: D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, D2S1338, D19S433, VWA, TPOX, D18S51, D5S818, FGA, Amelogenin
- AmpFℓSTR **NGM**: D10S1248, VWA, D16S539, D2S1338, Amelogenin, D8S1179, D21S11, D18S51, D22S1045, D19S433, TH01, FGA, D2S441
- AmpFℓSTR **NGM SElect**: D10S1248, VWA, D16S539, D2S1338, Amelogenin, D8S1179, D21S11, D18S51, D22S1045, D19S433, TH01, FGA, D2S441, D3S1358
- AmpFℓSTR **GlobalFiler**: D3S1358, VWA, D16S539, CSF1PO, TPOX, Yindel, Amelogenin,D8S1179, D21S11, D18S51, DYS391, D2S441, D3S1358
- AmpFℓSTR **VeriFiler**: D10S1248, D1S1656, Amelogenin, D2S1338, D22S1045, D19S433, TH01, D2S441, D6S1043, D12S391
- AmpFℓSTR **MiniFiler**: D13S317, D7S820, Amelogenin, D2S1338, D21S11, D16S539, D18S51, CSF1PO, FGA
- AmpFℓSTR **Yfiler**: DYS456, DYS389I, DYS389II, DYS390, DYS458, DYS19, DYS385a/b, DYS393, DYS391, DYS439, DYS635, DYS392, Y-GATA-H4
- AmpFℓSTR **SGM Plus**: D3S1358, VWA, D16S539, D2S1338, Amelogenin, D8S1179, D21S11, D18S51, D19S433, TH01, FGA
- AmpFℓSTR **Profiler Plus**: D3S1358, VWA, FGA, Amelogenin, D8S1179, D21S11, D18S51, D5S818, D13S317, D7S820
- AmpFℓSTR **Profiler Plus ID**: D3S1358, VWA, FGA, Amelogenin, D8S1179, D21S11, D18S51, D5S818, D13S317, D7S820
- AmpFℓSTR **COfiler** : D3S1358, D16S539, Amelogenin, TH01, TPOX, CSF1PO, D7S820
- AmpFℓSTR **Sinofiler** *(available only in China)*: D8S1179, D21S11, D7S820, CSF1PO, D3S1358, D5S818, D13S317, D16S539, D2S1338, Amelogenin
- AmpFℓSTR **Profiler**: D3S1358, VWA, FGA, Amelogenin, TH01, TPOX, CSF1PO, D5S818, D13S317, D7S820
- AmpFℓSTR **SEfiler**: D3S1358, VWA, D16S539, D2S1338, Amelogenin, D8S1179, SE33, D19S433, TH01, FGA, D21S11, D18S51
- AmpFℓSTR **SEfiler Plus**: D3S1358, VWA, D16S539, D2S1338, Amelogenin, D8S1179, SE33, D19S433, TH01, FGA, D21S11, D18S51
- AmpFℓSTR **Green I**: Amelogenin, TH01, TPOX, CSF1PO
- AmpFℓSTR **Blue**: D3S1358, VWA, FGA

http://www.cstl.nist.gov/strbase/multiplx.htm

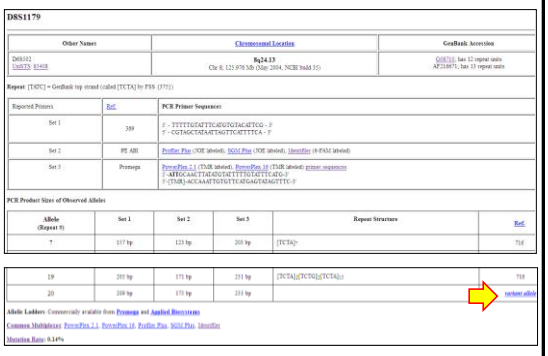
AmpFℓSTR® Identifiler™



The schematic diagram illustrates the fluorescent dye label color and relative PCR product size ranges for the various STR loci present in this particular kit. *Click on the locus name to learn more about the STR marker of interest.*

http://www.cstl.nist.gov/strbase/kits/Identifiler.htm

STR Fact Sheet for D8S1179

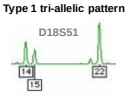


http://www.cstl.nist.gov/strbase/str_D8S1179.htm

Information on Variant Alleles

http://strbase.nist.gov/var_tab.htm

- We collect contributions from all over the world where unusual results have been observed with STR data
- Enables laboratories to check if others have seen a specific variant allele or tri-allelic pattern
- Currently (as of 23 Feb & 3 March 2017 updates)
  808 variants at 43 loci
  401 tri-allelic patterns at 38 loci



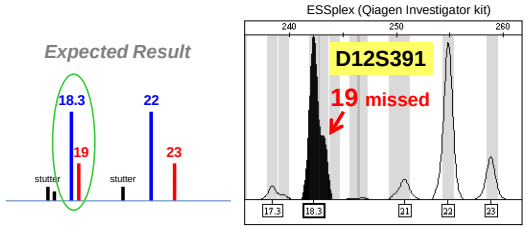
There is a backlog of >70 variant and tri-allele submissions for upload

Why the Planned Changes to STRBase?

- More information to upload than one person (i.e., John Butler) can easily manage
- The website system design and maintenance is out-of-date having been developed 20 years ago
- Some portions of the website are extensively used and updated (e.g., variant alleles) and other sections have fallen **significantly out-of-date**
  - STRBase is not up-to-date with many new autosomal STRs now being used
- **Would like to enable search capabilities** to aid future research investigations and answer specific questions
  - For example, D12S391 single-base variants exist and could potential impact accurate DNA mixture resolution

NIST SRM 2391c Component D

Provides a Single Base Resolution Challenge



Resolution challenges exist with D12S391 alleles 18.3 and 19, which differ by a single nucleotide; resolution can be impacted by the size of the PCR products in the specific STR kit and electrophoresis conditions (especially run voltage and polymer concentration)

STR Locus **D12S391 "Variant" Alleles**

What is defined as a variant (or off-ladder) allele by a laboratory is typically based on alleles present in STR kit allelic ladder

Allelic Ladder Alleles

Fusion	14	15	16	17	17.3	18	18.3	19		20	21	22	23	24	25	26	27
GlobalFiler	14	15	16	17		18		19	19.3	20	21	22	23	24	25	26	27

(NGM SElect = red dye)

D12S391 variant alleles (126 total) reported so far in STRBase
(data provided based on **123 NGM SElect**, 1 ESI16, 1 NGM, and 1 PP21)

As of Feb 2013

Variant	# times	Variant	# times
16.1	1x	19.1	1x
17.1	2x	20.1	2x
17.3	**43x**	20.3	2x
18.1	3x	21.3	1x
18.3	**66x**	28	1x

1 tri-allele reported
17,19,20
Sinofiler (China)

From NIST 1036 data set (Butler et al. 2012 *Profiles in DNA*)

Allele	Total #	Total %	Populations, % AfAm	Asian	Cauc	Hisp
14	1	0.0	0.1			
15	105	5.1	7.7	4.1	3.2	4.4
16	84	4.1	6.7	1.0	2.2	4.2
17	258	12.5	16.7	8.2	12.7	7.6
17.1	3	0.1	0.4			
17.3	**26**	1.3	0.4		**2.1**	1.7
18	**432**	20.8	25.3	26.3	17.2	17.8
18.1	1	0.0	0.1			
18.3	**27**	1.3	0.4		**2.5**	1.3
19	**314**	15.2	14.8	17.5	12.5	18.9
19.1	7	0.3	0.9			0.2
19.3	**10**	0.5	0.4	0.5	0.4	0.6
20	**262**	12.6	10.4	19.6	11.1	15.5
20.1	2	0.1	0.3			
20.3	1	0.0				0.2
21	209	10.1	6.4	9.8	12.9	11.2
22	137	6.6	3.7	5.7	9.6	6.8
22.2	1	0.0				0.2
23	102	4.9	2.9	2.6	6.9	5.7
24	53	2.6	1.3	1.0	4.7	1.7
24.3	1	0.0		0.5		
25	24	1.2	0.9	1.5	1.7	0.6
26	7	0.3		1.0	0.3	0.6
27	5	0.2		0.5	0.1	0.6

Hil, C.R., et al. (2013) U.S. population data for 29 autosomal STR loci. *Forensic Sci. Int. Genet.* 7: 682-683

D12S391 NIST U.S. Allele Frequencies

Theoretical heterozygotes (2pq)

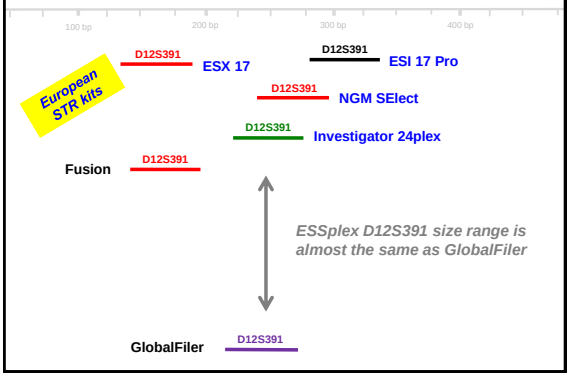
2 x 0.013 x 0.208 = **0.54%** (17.3,18)
2 x 0.013 x 0.152 = **0.40%** (18.3, 19)

Observed heterozygotes with a single nucleotide difference

9 out of 1036 = 0.87%

17, 17.1
17.3, 18 (3x)
18, 18.1
18.3, 19 (2x)
19, 19.1
19.3, 20

Relative Positions of D12S391



Benefits of a Website like STRBase

Now **http://strbase.nist.gov**

- Develops expertise when collecting information
- Requires NIST to stay up-to-date with field
- Provides transparency to our team's work
- Training tool and resource for the world
- Respected resource for >20 years
- **>10,000 pages of information available now**
- Widely used (>500,000 hits cumulative)
- **Method for sharing information (presentations, population data, etc.)**

STRBase could be a model for other forensic disciplines in sharing information with the forensic science community

Mixture Section of STRBase

http://strbase.nist.gov/mixture.htm

- **Training workshop slides** (thousands of slides of training materials available from >10 workshops)
- **SWGDAM Mixture Committee resource page** (contains worked mixture examples by Bruce Heidebrecht, Maryland State Police DNA Technical Leader)
- **Links to mixture interpretation software** (currently 17 links)
- **Literature references**
  - currently 150 articles listed
  - needs to be updated

Literature listing by topic for 150 articles

Topic category	# References
Mixture Principles & Recommendations	13
Setting Thresholds	12
Stutter Products & Peak Height Ratios	20
Stochastic Effects & Allele Dropout	18
Estimating the Number of Contributors	15
Mixture Ratios	9
Statistical Approaches	23
Low Template DNA Mixtures	10
Separating Cells to Avoid Mixtures	3
Software (plus 17 websites links)	7
Probabilistic Genotyping Approach	13
General Information on Mixtures	7

Additional Information Needed/Planned

- **Mutation rate information** to aid kinship analysis
  - More father/son studies are needed with D12S391, D1S1656, D2S441, D10S1248, and D22S1045
- A complete summary of **flanking region variation** and null alleles produced from primer binding site mutations
- Future plans for STRBase: listing of **full sequences for detected STR alleles** (repeats and flanking regions) to aid next-generation sequencing efforts
  - Will enable nomenclature and classification of sub-allele variation for STR markers

Revisions Planned

Lisa

Redesigning STRBase - Goals

- For the public
  - Make STRBase easier to navigate and use
    - Find what you are looking for faster
    - Ability to download more types of useful information
  - Make submitting information to STRBase simpler
- For the curators of STRBase
  - Make maintaining STRBase simpler
  - Simplify adding new information to STRBase
  - Simplify reviewing submitting information

Past and Future formats of STRBase

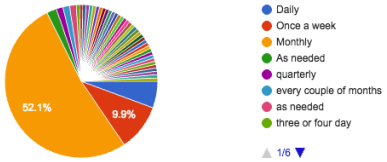
- Currently STRBase is over 2,500 individual files
  - HTML
- The new STRBase in development
  - ASP .NET Core
  - MySQL database

STRBase Questionnaire – an informal survey

147 Responses Total - 13 Questions

How often do you visit STRBase?

142 responses

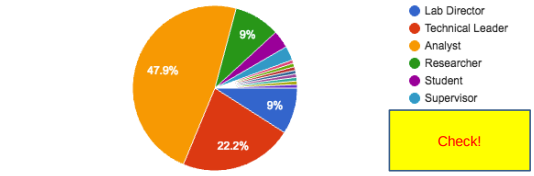


STRBase Questionnaire – an informal survey

147 Responses Total - 13 Questions

Which of these best describes your role?

144 responses

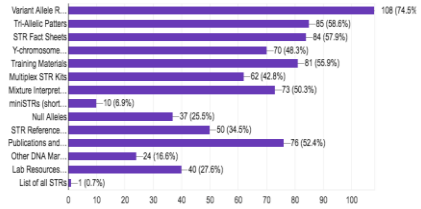


STRBase Questionnaire – an informal survey

147 Responses Total

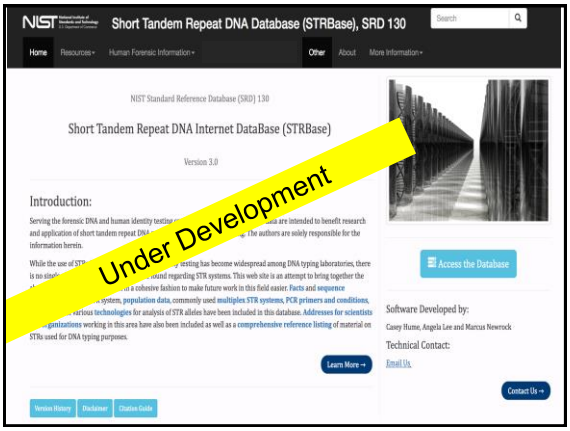
What informational pages do you use? (Select all that apply)

145 responses



Current designs for STRBase

- Pulling all information about a locus into one place
- Cleaning up the site
  - Removing duplicate data
  - Organizing
  - Simplifying
- General house cleaning



Some Examples of What Is in the Works for STRBase 3.0

Example D1S1656 – Variant Alleles

Search for specific criteria
Locus = D1S1656
Allele = 13.3

Allele Designation	Allele Size	Instrument	Amp Kit*	Contributor	Verification/ Confirmation Method(s)	Notes	Frequency
13.3	192.5	ABI 3500	NGM SElect	ABC Lab	Reamplified and reanalyzed	Convicted offender sample	
13.3	192.61	ABI 3500	NGM SElect	ABC Lab	Reamplified and reanalyzed	Convicted offender sample	
13.3	150.29	ABI 3500xl	ESX 16	XYZ Group	Re-amplified and re-electrophoresed	Reference sample	1
13.3	176.65	ABI 3500	PP21	LMN Department	Re-amplified & re-analyzed	Immigration case	2 in 2730

Sort Tables by columns

Example D1S1656 – Observed Alleles

Search for specific criteria
Sort Tables by columns

Allele	Bracket	Reference	Platform
7	[ca]5[TCTA]7/8	KCL	FGx/MiSeq
8	[ca]5[TCTA]8	Phillips[et al.][2011]	Sanger
9	[ca]5[TCTA]9	Phillips[et al.][2011]	Sanger
10	[ca]5[TCTA]10	Lareu[et al.][1998]	Sanger
10	[ca]5[CTA][TCTA]9	Phillips[et al.][2011]	Sanger
10.3		Variant Allele/STRBase	CE
11	[ca]5[TCTA]11	Lareu[et al.][1998]	Sanger
11	[ca]5[CTA][TCTA]10	Lareu[et al.][1998]	Sanger
11.1		Variant Allele/STRBase	CE
11.3		Variant Allele/STRBase	CE
12	[ca]5[TCTA]12	Lareu[et al.][1998]	Sanger
12	[ca]5[CTA][TCTA]11	Lareu[et al.][1998]	Sanger
12.1		Variant Allele/STRBase	CE
12.3		Variant Allele/STRBase	CE
13	[ca]5[TCTA]13	Phillips[et al.][2011]	Sanger
13	[ca]5[CTA][TCTA]12	Lareu[et al.][1998]	Sanger
13	[ca]5[TCTA]3[CTA][TCTA]11	Gettings[et al.][2015]	MiSeq
13.1		Variant Allele/STRBase	CE
13.3	[ca]5[CTA][TCTA]3[TCA][TCTA]4	Novroski[et al.][2016]	FGx/MiSeq
13.3	[ca]5[CTA][TCTA]11[TCA][TCTA	Phillips[et al.][2011]	Sanger

Example D1S1656 – Visualization of the Sequence

Identify surrounding sequence and provide observed SNPs

Other Updates

- Additional STRs
- Additional general information about the STRs
- Additions of new kits
- Visualization of new kits

STRBase Group for Website Upgrade

- John Butler
- Peter Vallone
- Katherine Gettings
- Lisa Borsuk

- Arlin Stoltzfus
- Casey Hume
- Angela Lee
- Marcus Newrock

What would you like to see in STRBase?

Thank you for your attention

Acknowledgments:
Peter Vallone

Contact Information

John M. Butler
NIST Fellow & Special Assistant
to the Director for Forensic Science
Special Programs Office
john.butler@nist.gov
+1-301-975-4049

Lisa Borsuk
Research Scientist, Bioinformatics
Applied Genetics Group
lisa.borsuk@nist.gov
+1-301-975-5405

Katherine Gettings
Research Biologist
Applied Genetics Group
katherine.gettings@nist.gov
+1-301-975-6401

A copy of this presentation is available at:
http://strbase.nist.gov/NISTpub.htm

29 August 2017

Workshop #10

Other Uses & Potential Privacy Concerns

John M. Butler, Ph.D.
Special Programs Office
U.S. National Institute of Standards and Technology

STRs are the Dominant Genetic Markers Used in Paternity Testing

Latest available report (2013):
http://www.aabb.org/sa/facilities/Documents/2013-relationship-testing-summary-report.pdf

Table 4. The Technology Used in Cases Reported in 2013

Technology	Number of Cases	Utilization (%)
STR	368357	99.84
RFLP	48	0.01
HLA Class II Molecular	1	0.0003
Y Chromosome	489	0.13
HLA Class I Molecular	46	0.01
SNP	0	0.00
Total of All Technologies	368941	100

*Note that some cases used more than one technology.

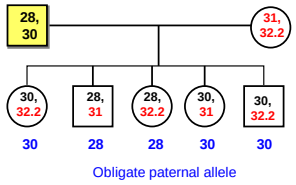
Table 5. Sample Source in 2013.

Sample	Number	Percent
Buccal Swabs	892256	99.5051
Blood	1458	0.1626
Blood Spot Cards	1356	0.1512
Amniotic Fluid	463	0.0516
Misc Tissues	314	0.0350
Paraffin Blocks	67	0.0075
Hair	364	0.0406
CVS	385	0.0429
Bone	31	0.0035
Total	896694	100.00

http://www.aabb.org/sa/facilities/Pages/relationshipreports.aspx

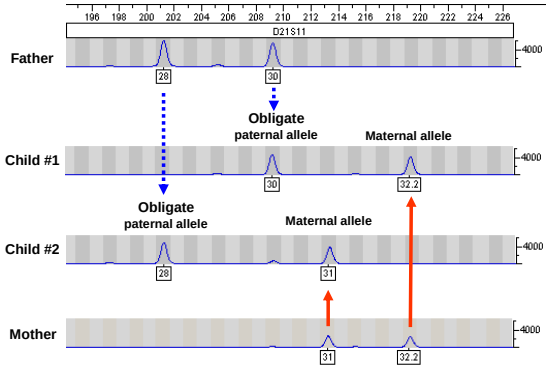
Parentage or Relationship Testing

Example Pedigree: STR Marker D21S11



Obligate paternal allele

From Figure 14.2(b) in Butler, J.M. (2015) *Advanced Topics in Forensic DNA Typing: Interpretation* Elsevier Academic Press: San Diego



From Figure 14.2(c) in Butler, J.M. (2015) *Advanced Topics in Forensic DNA Typing: Interpretation* Elsevier Academic Press: San Diego

Aiding Cell Line Authentication

Katsnelson, A. (2010) *Nature News*, 465: 537 (3 June 2010)

Biologists tackle cells' identity crisis

DNA fingerprinting scheme aims to make sure researchers are working on the right cells.

Ever since biologists learned how to grow human cells in culture half a century ago, the cells have been plagued by a problem of identity: many commonly used cell lines are not actually what researchers think they are.

Cell-line misidentification has led to mistakes in the literature, misguided research based on those results and millions wasted in grant money. Last year, *Nature* described the situation as a scandal[1].

But a universal system for determining the identity of cell lines may now be in view. Next month, a working group led by the American Type Culture Collection (ATCC), a nonprofit biological repository based in Manassas, Virginia, that stores 3,600 cell lines from more than 150 species, plans to unveil standard-

ATCC
LEADING BIOLOGICAL STANDARDS
ATCC® Standards Development Organization

Designation: ASN-0002

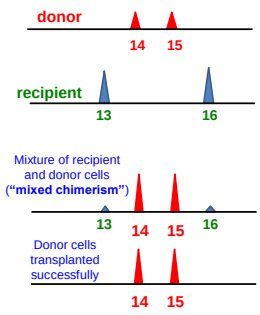
Authentication of Human Cell Lines: Standardization of STR Profiling

The working group, composed of representatives from academia, government and industry, a universally accepted approach will allow different facilities to compare their cell lines with each other, he adds.

Fingerprinting has its limits, cautions Michael Johnson, a cancer researcher at Georgetown University in Washington DC. "Just because a cell fingerprints out as the same [as another cell] doesn't mean they will behave the same," he says, noting that a cell's properties can also be affected by the way it has been grown, the number of times it has been cultured anew and small genetic changes that wouldn't show up in a fingerprint test. One classic example, he notes, is an immortalized breast cell line called MCF10A, which can form organized hollow structures similar to those found in mammary tissue; MCF10A cells currently distributed by

http://www.nature.com/news/2010/100602/pdf/465537a.pdf

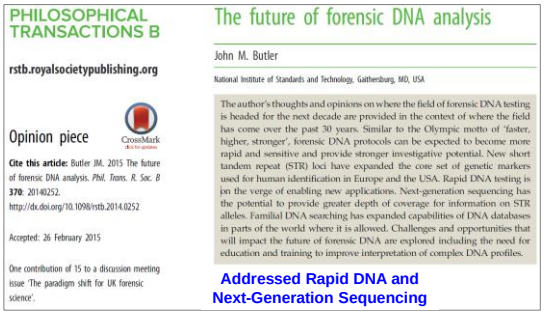
Bone Marrow Transplant Monitoring:
Easier with STRs that possess many alleles



- When there are different STR genotypes between donors and recipients at the tested loci, it is possible to evaluate the degree of donor transplantation
- STR analysis enables monitoring the persistence of recipient cells

Antin, J.H. et al. (2001) Establishment of complete and mixed donor chimerism after allogeneic lymphohematopoietic transplantation: recommendations from a workshop at the 2001 tandem meetings. *Biology of Blood and Marrow Transplantation* 7: 473-485

Thoughts on the Future of Forensic DNA Published in 2015



PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Opinion piece

Cite this article: Butler JM. 2015 The future of forensic DNA analysis. *Phil. Trans. R. Soc. B* **370**: 20140252. http://dx.doi.org/10.1098/rstb.2014.0252

Accepted: 26 February 2015

One contribution of 15 to a discussion meeting issue 'The paradigm shift for UK forensic science'.

The future of forensic DNA analysis

John M. Butler

National Institute of Standards and Technology, Gaithersburg, MD, USA

The author's thoughts and opinions on where the field of forensic DNA testing is headed for the next decade are provided in the context of where the field has come over the past 30 years. Similar to the Olympic motto of 'faster, higher, stronger', forensic DNA protocols can be expected to become more rapid and sensitive and provide stronger investigative potential. New short tandem repeat (STR) loci have expanded the core set of genetic markers used for human identification in Europe and the USA. Rapid DNA testing is on the verge of enabling new applications. Next-generation sequencing has the potential to provide greater depth of coverage for information on STR alleles. Familial DNA searching has expanded capabilities of DNA databases in parts of the world where it is allowed. Challenges and opportunities that will impact the future of forensic DNA are explored including the need for education and training to improve interpretation of complex DNA profiles.

Addressed Rapid DNA and Next-Generation Sequencing

Butler, J.M. (2015) The future of forensic DNA analysis. *Phil. Trans. R. Soc.* B 370: 20140252

Current Trends in Forensic DNA

- ***Faster** results:* Rapid DNA capabilities and new sample-to-answer integrated instruments
- ***Higher** sensitivity:* New assays lowering the limits of detection, which makes interpretation more challenging
- ***Higher** information content:* Next-generation sequencing (NGS) for more markers & STR allele information
- ***Stronger** conclusions:* Mixture interpretation with probabilistic genotyping models

Butler, J.M. (2015) The future of forensic DNA analysis. *Phil. Trans. R. Soc.* B 370: 20140252

Stages of Forensic DNA Progression

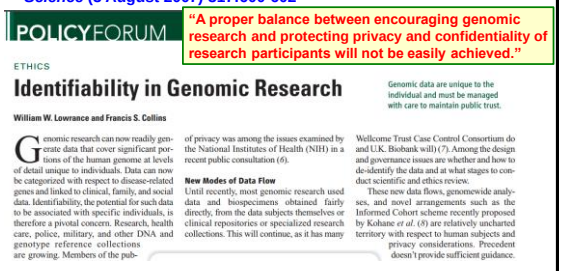
Stages	Time Frame	Description
Exploration	1985 - 1995	Beginnings, different methods tried (RFLP and early PCR)
Stabilization	1995 - 2005	Standardization to STRs, selection of core loci, implementation of Quality Assurance Standards
Growth	2005 - 2015	Rapid growth of DNA databases, extended applications pursued
Sophistication	*2015 to 2025 and beyond*	*Expanding tools available, **confronting privacy concerns***

Table 1 from J.M. Butler (2015) The future of forensic DNA analysis. *Phil. Trans. R. Soc. B* 370: 20140252

Genomic Research Enables Identity Testing and Potentially Impacts Privacy

Lowrance, W.W., & Collins, F.S. Identifiability in genomic research. *Science* (3 August 2007) 317:600-602

"A proper balance between encouraging genomic research and protecting privacy and confidentiality of research participants will not be easily achieved."



POLICYFORUM

ETHICS

Identifiability in Genomic Research

Genomic data are unique to the individual and must be managed with care to maintain public trust.

William W. Lowrance and Francis S. Collins

Genomic research can now readily generate data that cover significant portions of the human genome at levels of detail unique to individuals. Data can now be categorized with respect to disease-related genes and linked to clinical, family, and social data. Identifiability, the potential for such data to be associated with specific individuals, is therefore a pivotal concern. Research, health care, police, military, and other DNA and genotype reference collections are growing. Members of the public of privacy was among the issues examined by the National Institutes of Health (NIH) in a recent public consultation (*6*).

New Modes of Data Flow

Until recently, most genomic research used data and biospecimens obtained fairly directly, from the data subjects themselves or clinical repositories or specialized research collections. This will continue, as it has many

Wellcome Trust Case Control Consortium do and U.K. Biobank will) (*7*). Among the design and governance issues are whether and how to de-identify the data and at what stages to conduct scientific and ethics review.

These new data flows, genomewide analyses, and novel arrangements such as the Informed Cohort scheme recently proposed by Kohane *et al.* (*8*) are relatively uncharted territory with respect to human subjects and privacy considerations. Precedent doesn't provide sufficient guidance.

An Attempt to Link Forensic STR Markers to Clinical SNP Assays

Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets

Michael D. Edge[a], Bridget F. B. Algee-Hewitt[a], Trevor J. Pemberton[b], Jun Z. Li[c], and Noah A. Rosenberg[a,1]

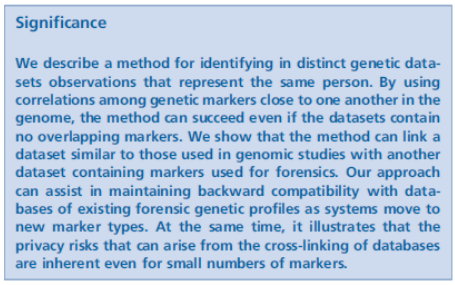
[a]Department of Biology, Stanford University, Stanford, CA 94305; [b]Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, MB, Canada R3E0J9; and [c]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109

Edited by Andrew G. Clark, Cornell University, Ithaca, NY, and approved April 10, 2017 (received for review December 6, 2016)

"Using two datasets for the same 872 people—one with 642,563 genome-wide SNPs and the other with 13 short tandem repeats (STRs) used in forensic applications—we find that 90–98% of forensic STR records can be connected to corresponding SNP records and vice versa. Accuracy increases to 99–100%when ~30 STRs are used. Our method expands the potential of data aggregation, but it also suggests privacy risks intrinsic in maintenance of databases containing even small numbers of markers—including databases of forensic significance."

Edge, M.D., et al. (2017) Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets. *PNAS (Proceedings of the National Academy of Sciences USA)* 114: 5671-5676

This Work Has Raised the Potential for Perceived Privacy Risks



Significance

We describe a method for identifying in distinct genetic datasets observations that represent the same person. By using correlations among genetic markers close to one another in the genome, the method can succeed even if the datasets contain no overlapping markers. We show that the method can link a dataset similar to those used in genomic studies with another dataset containing markers used for forensics. Our approach can assist in maintaining backward compatibility with databases of existing forensic genetic profiles as systems move to new marker types. At the same time, it illustrates that the privacy risks that can arise from the cross-linking of databases are inherent even for small numbers of markers.

Edge, M.D., et al. (2017) Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets. *PNAS (Proceedings of the National Academy of Sciences USA)* 114: 5671-5676

This Concern is Not New…

Kimpton, C.P., et al. (1995). Report on the second EDNAP collaborative STR exercise. *Forensic Science International*, 71, 137-152.

"…it is likely that many or possibly most STRs will eventually be shown to be useful in following a genetic disease or other genetic trait *within a family* and therefore this possibility must be recognized at the outset of the use of such systems" (emphasis added)

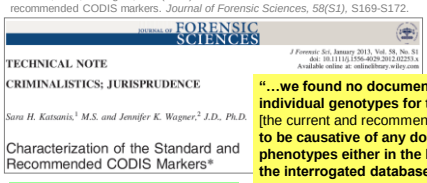
Laird, R., et al. (2007). Forensic STRs as potential disease markers: a study of VWA and von Willebrand's disease. *Forensic Science International: Genetics, 1,* 253-261

Abstract

"In recent years it has been established that non-coding variants may be in linkage disequilibrium (LD) with coding variants up to several thousand base pairs away forming haplotype blocks. These non-coding markers may be haplotype specific and, therefore, informative regarding the surrounding coding sequence. In this study, we chose to study the VWA short tandem repeat (STR) as it is targeted in all major commercial kits utilized in routine forensic DNA profiling and is located in the von Willebrand Factor (vWF) gene; a gene associated with von Willebrand's Disease (vWD)… [T]**here appeared to be no evidence of LD blocks surrounding the VWA STR and evidence for recombination within 3 kb of VWA, hence, it is unlikely that VWA STR alleles could be used to predict haplotypes within the vWF gene that are associated with different forms of vWD."**

Forensic STR loci are not linked to disease…

Katsanis, S.H., & Wagner, J.K. (2013) Characterization of the standard and recommended CODIS markers. *Journal of Forensic Sciences, 58(S1),* S169-S172.


JOURNAL OF FORENSIC SCIENCES

TECHNICAL NOTE

CRIMINALISTICS; JURISPRUDENCE

Sara H. Katsanis,[1] M.S. and Jennifer K. Wagner,[2] J.D., Ph.D.

Characterization of the Standard and Recommended CODIS Markers*

"…we found no documentation of individual genotypes for the 24 STRs [the current and recommended CODIS loci] to be causative of any documented phenotypes either in the literature or in the interrogated databases."

See also on http://www.swgdam.org/

Open SWGDAM Letter Regarding the Claims Raised in State v. Abernathy that the CODIS Core Loci are Associated with Medical Conditions/Disease States

"The utility of the CODIS profile … is limited to identification purposes at this time."

"…we can affirm that individual genotypes are not at present revealing information beyond identification."

SWGDAM statement on *Abernathy* ruling

See **http://www.swgdam.org/**

SCIENTIFIC WORKING GROUP — DNA ANALYSIS METHODS

SWGDAM[1] Considerations for Claims that the CODIS Core Loci are 'Associated' with Medical Conditions/Diseases

In a June 2012 ruling, *State v. Abernathy*, No. 3599-9-11, a Vermont Court adopted the testimony from a defense expert on the CODIS core loci and found that "the analogy between DNA testing and fingerprinting is no longer valid, because a DNA profile consisting of the thirteen CODIS loci contains information beyond mere identity." In

See https://www.swgdam.org/publications

From J.M. Butler (2012) *Advanced Topics in Forensic DNA Typing: Methodology*, p. 228

"…[U]se of STRs for family linkage studies is different than associations of specific alleles in a general population with a disease state. Colin Kimpton and coworkers from the European DNA Profiling Group (EDNAP) recognized early on in the application of STRs for human identity testing that 'it is likely that many or possibly most STRs will eventually be shown to be useful in following a genetic disease or other genetic trait *within a family* and therefore this possibility must be recognized at the outset of the use of such systems' (Kimpton et al. 1995; emphasis added). **Family pedigree studies that track a few specific loci and alleles are different than equating a specific allele in the population with some kind of phenotypic correlation**…"

Kimpton, C.P., et al. (1995). Report on the second EDNAP collaborative STR exercise. *Forensic Science International*, 71, 137-152.

From J.M. Butler (2012) *Advanced Topics in Forensic DNA Typing: Methodology*, p. 228

"In 2005, an infrequently used X-chromosome STR marker named HumARA was removed from future consideration in human identity testing (Szibor et al. 2005) since it was located in an exon. Some of the longer CAG repeat alleles with HumARA have been shown to be the cause of a genetic disease, which is why this STR locus was removed from use. **All of the 23 commonly used STR markers described throughout this book and present in current commercial STR kits are located in between genes ('junk DNA' regions) or in introns. Thus, by definition they are non-coding**."

Szibor, R., et al. (2005). Letter to the editor: the HumARA genotype is linked to spinal and bulbar muscular dystrophy and some further disease risks and should no longer be used as a DNA marker for forensic purposes. *International Journal of Legal Medicine*, *119*, 179-180.

From J.M. Butler (2012) *Advanced Topics in Forensic DNA Typing: Methodology*, p. 228

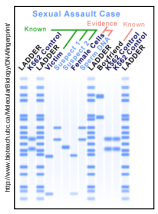
"[T]he relatively high mutation rate of STRs means that even if any linkage existed at one time between a specific allele and a genetic disease state, this linkage would likely not last beyond a few generations before mutation altered the allele length and effectively broke any linkage of an allele or genotype state to that specific phenotype state."

Summary

- STR markers have proven to be valuable in forensic evidence examinations for almost two decades (the U.S. has recently moved from 13 to ~20 core STR loci)
- Genetic disease linkage studies often involve STR markers, some of which may be core forensic loci
- **The high mutation rate of forensic STR markers means that any potential allele associations with disease phenotypes will not hold over time in the general population**

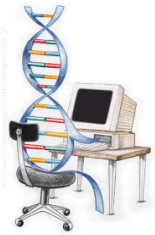
The DNA Field Moves Forward…

The Past | **The Present** | **The Future**

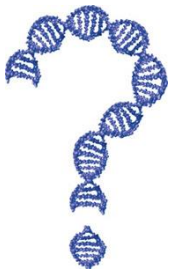


RFLP

STRs



Thank you for your attention!



Contact info:
john.butler@nist.gov
301-975-4049

A copy of this presentation is available at:
http://www.cstl.nist.gov/strbase/NISTpub.htm

Reference List for ISFG 2017 Workshop #10 on Autosomal STR Markers & Interpretation

STR Kits and Measurement Techniques

Butler, J.M., et al. (2004). Forensic DNA typing by capillary electrophoresis: using the ABI Prism 310 and 3100 Genetic Analyzers for STR analysis. *Electrophoresis*, 25: 1397-1412.

Butler, J.M. (2012). *Advanced Topics in Forensic DNA Typing: Methodology*. Elsevier Academic Press: San Diego.

Butler, J.M. (2015). *Advanced Topics in Forensic DNA Typing: Interpretation.* Elsevier Academic Press: San Diego.

Butler, J.M. (2015) The future of forensic DNA analysis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 370,* 20140252.

STR Markers and Alleles

Butler, J.M. (2006). Genetics and genomics of core STR loci used in human identity testing. *Journal of Forensic Sciences, 51(2),* 253-265.

Butler, J.M., et al. (2009). The single most polymorphic STR locus: SE33 performance in U.S. populations. *Forensic Science International: Genetics Supplement Series (Progress in Forensic Genetics 13*), *2*, 23-24. [available at http://www.fsigeneticssup.com]

Butler, J.M., et al. (2011). SE33 variant alleles: sequences and implications. *Forensic Science International: Genetics Supplement Series, 3*, e502-e503. [available at http://www.fsigeneticssup.com]

Butler, J.M., et al. (2012). Variability of new STR loci and kits in U.S. population groups. *Profiles in DNA*. Available at https://www.promega.com/resources/profiles-in-dna/2012/variability-of-new-str-loci-and-kits-in-us-population-groups/.

Butler, J.M., & Hill, C.R. (2012). Biology and genetics of new autosomal STR loci useful for forensic DNA analysis. *Forensic Science Reviews, 24(1),* 15-26.

Gill, P., et al. (2006a). The evolution of DNA databases-Recommendations for new European STR loci. *Forensic Science International, 156*, 242-244.

Gill, P., et al. (2006b). New multiplexes for Europe-amendments and clarification of strategic development. *Forensic Science International, 163*, 155-157.

Hares, D.R. (2015) Selection and implementation of expanded CODIS core loci in the United States. *Forensic Sci. Int. Genet.* 17): 33-34.

Phillips, C., et al. (2014). "New turns from old STaRs": enhancing the capabilities of forensic short tandem repeat analysis. *Electrophoresis, 35,* 3173-3187.

Phillips, C., et al. (2016) D5S2500 is an ambiguously characterized STR: identification and description of forensic microsatellites in the genomics age. *Forensic Science International*: *Genetics, 23,* 19-24.

Interpretation

Cadenas, A.M., et al. (2007). Male amelogenin dropouts: phylogenetic context, origins and implications. *Forensic Science International, 166*, 155-163.

Coble, M.D., et al. (2015). Uncertainty in the number of contributors in the proposed new CODIS set. *Forensic Science International*: *Genetics, 19,* 207-211.

Gill, P., et al. (2015). Genotyping and interpretation of STR-DNA: Low-template, mixtures and database matches-Twenty years of research and development. *Forensic Science International*: *Genetics, 18,* 100-117.

Santos, F.R., et al. (1998). Reliability of DNA-based sex tests. *Nature Genetics, 18(2)*, 103-103.

SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories (see swgdam.org)

Length vs Sequence Information and STR Nomenclature

Bodner, M., et al. (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 Science International*: *Genetics, 24,* 97-102.

Casals, F., et al. (2017). Length and repeat-sequence variation in 58 STRs and 94 SNPs in two Spanish populations. *Forensic Science International*: *Genetics, 30*, 66-70.

Friis, S.L., et al. (2016). Introduction of the Python script STRinNGS for analysis of STR regions in FASTQ or BAM files and expansion of the Danish STR sequence database to 11 STRs. *Forensic Science International*: *Genetics, 21*, 68-75.

Gettings, K.B., et al. (2015). STR allele sequence variation: current knowledge and future issues. *Forensic Science International*: *Genetics, 18,* 118-130.

Gettings, K.B., et al. (2016). Sequence variation of 22 autosomal STR loci detected by next generation sequencing. *Forensic Science International: Genetics, 21,* 15-21.

Novroski, N.M., et al. (2016). Characterization of genetic sequence variation of 58 STR loci in four major population groups. *Forensic Science International*: *Genetics, 25,* 214-226.

Parson, W., et al. (2016). Massively parallel sequencing of forensic STRs: considerations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. *Forensic Science International*: *Genetics, 22,* 54-63.

Phillips, C. (2017). A genomic audit of newly-adopted autosomal STRs for forensic identification. *Forensic Science International: Genetics, 29,* 193-204.

STRidER (STRs for Identity ENFSI Reference Database) available at https://strider.online/

STRSeq (The STR Sequencing Project) available at https://www.ncbi.nlm.nih.gov/bioproject/380127

van der Gaag, K.J., et al. (2016). Massively parallel sequencing of short tandem repeats-Population data and mixture analysis results for the PowerSeq system. *Forensic Science International*: *Genetics, 24,* 86-96.

Wendt, F.R., et al. (2016). Genetic analysis of the Yavapai Native Americans from West-Central Arizona using the Illumina MiSeq FGx™ forensic genomics system. *Forensic Science International*: *Genetics, 24,* 18-23.

Wendt, F.R., et al. (2017). Flanking region variation of ForenSeq DNA Signature Prep Kit STR and SNP loci in Yavapai Native Americans. *Forensic Science International*: *Genetics, 28,* 146-154.

STRBase

Ruitberg, C.M., et al. (2001). STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Research, 29,* 320-322.

STRBase available at http://strbase.nist.gov/ (URL changed from http://www.cstl.nist.gov/strbase)

Other Uses, Potential Disease Linkage, and Privacy Concerns

Katsanis, S.H., & Wagner, J.K. (2013) Characterization of the standard and recommended CODIS markers. *Journal of Forensic Sciences, 58(S1),* S169-S172.

Kimpton, C.P., et al. (1995). Report on the second EDNAP collaborative STR exercise. *Forensic Science International*, *71,* 137-152.

Kloosterman, A., Sjerps, M. and Quak, A. (2014) Error rates in forensic DNA analysis: definition, numbers, impact and communication. *Forensic Science International: Genetics, 12,* 77-85.

Laird, R., et al. (2007). Forensic STRs as potential disease markers: a study of VWA and von Willebrand's disease. *Forensic Science International: Genetics, 1,* 253-261.

SWGDAM (2012). SWGDAM Executive Board Considerations for Claims that the CODIS Core Loci are 'Associated' with Medical Conditions/Diseases. Available at https://www.swgdam.org/publications.