The Impact and Benefit of Expanding the U.S. Core Autosomal STR Markers

Becky Hill, Margaret Kline, Dave Duewer and John Butler
National Institute of Standards and Technology

American Academy of Forensic Sciences
Washington, DC
February 21, 2013

Outline of Topics to Discuss

- Need for additional loci
  - Growth in U.S. and other national DNA databases
- Locus characteristics and population data
- New STR kits available with additional loci
  - GlobalFiler (Life Technologies)
  - PowerPlex Fusion (Promega)

Need for Additional Loci

Additional STR Loci in the Future?

- More loci will be needed for more complex kinship analyses and extended applications
  - Example: Y-STRs needed for familial searching
- Immigration testing needs more than 13 STRs
- Larger DNA databases will require more loci
  - U.S. national DNA database currently has 12 million profiles and it continues to quickly grow

Expanded U.S. Core Loci

Forensic Science International: Genetics 6 (2012) e52–e54

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Letter to the Editor

Expanding the CODIS core loci in the United States

Dear Editor:

After over a decade of operation, the National DNA Index System (NDIS) continues to grow in importance and size [1]. While the STR DNA technology has remained relatively consistent, other key aspects of the NDIS program have been reevaluated and revisions implemented. For example, based upon recommendations of the Scientific Working Group on DNA Analysis Methods, the Director of the Federal Bureau of Investigation (FBI) issued revised Quality Assurance Standards (QAS) for Forensic DNA

major reasons for expanding the CODIS core loci in the United States:

(1) To reduce the likelihood of adventitious matches [7] as the number of profiles stored at NDIS continues to increase each year (expected to total over 10 million profiles by the time of this publication). There are no signs that this trend will slow down as States expand the coverage of their DNA database programs and increase laboratory efficiency and capacity.

(2) To increase international compatibility to assist law enforcement data sharing efforts.

(3) To increase discrimination power to aid missing persons cases.

Hares, D.R. (2012a) Expanding the CODIS core loci in the United States. *Forensic Sci. Int. Genet.* 6(1), e52-4.

Forensic Science International: Genetics 6 (2012) e135

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

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

Letter to the Editor

Addendum to expanding the CODIS core loci in the United States

Dear Editor,

An important objective in proposing new CODIS core loci is to ensure that all loci would be available for all potential manufacturers. During the evaluation process, appropriate steps were taken to document access to all proposed core loci. Since

publication of the proposed list of core loci, additional information has come to our attention indicating that there may be outstanding issues with respect to some of the proposed loci. Consequently, to ensure the availability for all interested manufacturers in accordance with our stated objective, we are withdrawing Penta D and Penta E as proposed CODIS core loci and recommending the revised listing of core loci in Table 1. Manufacturers are still encouraged to attempt loci in Section B, in ranked order of preference, for inclusion in potential kits provided the impact on the kit's sensitivity and overall performance is negligible. Please

Hares, D.R. (2012b) Addendum to expanding the CODIS core loci in the United States. *Forensic Sci. Int. Genet.* 6(5), e135.

Required and Recommended CODIS Core Loci ("CODIS 20")

Table 1
Revised ranked list of CODIS core loci.

Locus
Section A (required) **Required Loci**
Amelogenin
D18S51
FGA
D21S11
D8S1179
vWA
D13S317
D16S539
D7S820
TH01
D3S1358
D5S818
CSF1PO
D2S1338
D19S433
D1S1656
D12S391
D2S441
D10S1248
DYS391
Section B (in order of preference) **Recommended Loci**
TPOX
D22S1045
SE33

Penta D and Penta E were removed from this list in the addendum

Y-STR to confirm Amelogenin null alleles

No longer required

NIST U.S. Population Samples

NIST U.S. Samples (>1450)

- **NIST U.S. population samples**
  - 260 African American, 260 Caucasian, 140 Hispanic, 3 Asian
- **U.S. father/son paired samples**
  - ~**100 fathers/100 sons for each group**: 200 African American, 200 Caucasian, 200 Hispanic, 200 Asian
- **NIST SRM 2391b**, PCR-based DNA Profiling Standard (highly characterized)
  - 10 genomic DNA samples, 2 cell line samples
  - Includes 9947A and 9948
- **NIST SRM 2391c**, PCR-based DNA Profiling Standard
  - 4 genomic DNA (one mixture)
  - 2 cell lines (903 and FTA paper)

Hill, C.R., et al. (2011) Concordance and population studies along with stutter and peak height ratio analysis for the PowerPlex® ESX 17 and ESI 17 Systems. *Forensic Sci. Int. Genet.* 5(4): 269-275.

Publications using NIST Population Samples

Data available at
http://www.cstl.nist.gov/strbase/NISTpop.htm

1. Butler et al. (2003) *J. Forensic Sci.* – Identifiler allele frequencies
2. Butler et al. (2003) *J. Forensic Sci.* – miniSTR assay development
3. Drabek et al. (2004) *J. Forensic Sci.* – miniSTR concordance
4. Schoske et al. (2004) *Forensic Sci. Int.* – Y-STR 20plex & 11plex
5. Vallone et al. (2004) *J. Forensic Sci.* – 50 Y-SNPs
6. Coble & Butler (2005) *J. Forensic Sci.* – NC01 & NC02 assay development
7. Butler et al. (2005) *J. Forensic Sci.* – PowerPlex Y with Y-STR duplications & triplications
8. Vallone et al. (2005) *Forensic Sci. Int.* – 70 autosomal SNPs
9. Butler et al. (2006) *Forensic Sci. Int.* – 27 Y-STR additional loci
10. Hill et al. (2007) *J. Forensic Sci.* – MiniFiler concordance
11. Decker et al. (2008) *FSI Genetics* - Yfiler mutation rates
12. Saunier et al. (2008) *FSI Genetics* – mtDNA control region sequencing (AFDIL)
13. Just et al. (2008) *FSI Genetics* – mtGenome analysis (AFDIL)
14. Hill et al. (2008) *J. Forensic Sci.* – NC01-NC09 miniSTR loci
15. Diegoli et al. (2009) *FSI Genetics* – mtDNA control region sequencing (AFDIL)
16. Hill et al. (2009) *J. Forensic Sci.* – NIST 26plex
17. Lao et al. (2010) *Human Mutation* – 24 ancestry SNPs, Y-SNPs, mtDNA
18. Hill et al. (2011) *FSI Genetics* – ESI 17 & ESX 17 concordance
19. Diegoli et al. (2011) *FSI Genetics Suppl. Ser.* – Argus X-12 X-STR loci
20. Fondevila et al. (2012) *Int. J. Legal Med*. – 68 InDel loci
21. Fondevila et al. (2012) *FSI Genetics* – 34 ancestry SNPs
22. Butler et al. (2012) *Profiles in DNA* – introduces NIST 1036 data set
23. Hill et al. (2013) *FSI Genetics (in press)* – 29 autosomal STRs in PowerPlex CS7 and other kits
24. Coble et al. (2013) *FSI Genetics (in press)* – 23 Y-STRs in PowerPlex Y23

Testing also completed with 16 X-STR loci and 14 rapidly mutating (RM) Y-STRs

NIST 1036 U.S. Population Samples

- 1032 males + 4 females
  - 361 Caucasians (2 female)
  - 342 African Americans (1 female)
  - 236 Hispanics
  - 97 Asians (1 female)

***Unrelated* samples**
All known or potential related individuals (based on autosomal & lineage marker testing) have been removed from the 1036 data set (e.g., only sons were used from father-son samples)

- Anonymous donors with self-identified ancestry
  - Interstate Blood Bank (Memphis, TN) – obtained in 2002
  - Millennium Biotech, Inc. (Ft. Lauderdale, FL) – obtained in 2001
  - DNA Diagnostics Center (Fairfield, OH) – obtained in 2007

- **Complete profiles with 29 autosomal STRs + PowerPlex Y23**
  - **Examined with multiple kits** and in-house primer sets enabling concordance

- Additional DNA results available on subsets of these samples
  - mtDNA control region/whole genome (AFDIL)
  - >100 SNPs (AIMs), 68 InDel markers, X-STRs (AFDIL)
  - NIST assays: miniSTRs, 26plex, >100 Y-STRs, 50 Y-SNPs

Data available on STRBase: **http://www.cstl.nist.gov/strbase/NISTpop.htm**

29 Autosomal STR Markers Present in Commercial STR Multiplex Kits

Locus	CODIS 13	CODIS 20	ESS 12	Other Kits
	Required loci			
D1S1656				
F13B				CS7
TPOX				
D2S441				
D2S1338				
D3S1358				
FGA				
CSF1PO				
D5S818				
F13A01				CS7
D6S1043				Sinofiler, PP21
SE33				PPESX17, PPESI17, NGM SElect, GlobalFiler
D7S820				
LPL				CS7
D8S1179				
Penta C				CS7
D10S1248				
TH01				
D12S391				
vWA				
D13S317				
FESFPS				CS7
Penta E				PP16, PP21, PP Fusion
D16S539				
D18S51				
D19S433				
D21S11				
Penta D				PP16, PP21, PP Fusion
D22S1045				
Amelogenin				
DYS391				PP Fusion, GlobalFiler

Benefits of NIST 1036 Data Set

- **Elimination of potential null alleles due to primer binding site mutations** through extensive concordance testing performed with different PCR primer sets from all available commercial STR kits

- **Ancestry testing performed** on DNA samples with autosomal SNPs, Y-SNPs, and mtDNA sequencing to verify self-declared ancestry categorization

- **Related individuals removed** based on Y-STR and mtDNA results so that allele frequencies are not potentially inflated by closely related individuals

Characterization of STR Loci

Available in Commercial Kits

The 10 STR Loci Beyond the CODIS 13

STR Locus	Location	Repeat Motif	Allele Range*	# Alleles*
D2S1338	2q35	TGCC/TTCC	10 to 31	40
D19S433	19q12	AAGG/TAGG	5.2 to 20	36
Penta D	21q22.3	AAAGA	1.1 to 19	50
Penta E	15q26.2	AAAGA	5 to 32	53
D1S1656	1q42	TAGA	8 to 20.3	25
D12S391	12p13.2	AGAT/AGAC	13 to 27.2	52
D2S441	2p14	TCTA/TCAA	8 to 17	22
D10S1248	10q26.3	GGAA	7 to 19	13
D22S1045	22q12.3	ATT	7 to 20	14
SE33	6q14	AAAG‡	3 to 49	**178**

5 new European loci

**Allele range and number of observed alleles from Appendix 1, J.M. Butler (2011) Advanced Topics in Forensic DNA Typing: Methodology; ‡SE33 alleles have complex repeat structure*

25 Alleles Reported in the Literature for **D1S1656**

15 NIST observed alleles circled in red

Allele (Repeat #)	Promega ESX 17	Promega ESI 17	ABI NGM	Repeat Structure $[TAGA]_4[TGA]_{0-1}[TAGA]_nTAGG[TG]_5$	Reference
8	133 bp	222 bp	171 bp	$[TAGA]_8[TG]_5$	Phillips *et al.* (2010)
9	137 bp	226 bp	175 bp	$[TAGA]_9[TG]_5$	Phillips *et al.* (2010)
10 (a)	141 bp	230 bp	179 bp	$[TAGA]_{10}[TG]_5$	Lareu *et al.* (1998)
10 (b)	141 bp	230 bp	179 bp	$[TAGA]_{10}TAGG[TG]_5$	Phillips *et al.* (2010)
11	145 bp	234 bp	183 bp	$[TAGA]_{11}[TG]_5$	Lareu *et al.* (1998)
12 (a)	149 bp	238 bp	187 bp	$[TAGA]_{12}[TG]_5$	Lareu *et al.* (1998)
12 (b)	149 bp	238 bp	187 bp	$[TAGA]_{11}TAGG[TG]_5$	Lareu *et al.* (1998)
13 (a)	153 bp	242 bp	191 bp	$[TAGA]_{12}TAGG[TG]_5$	Lareu *et al.* (1998)
13 (b)	153 bp	242 bp	191 bp	$[TAGA]_{13}[TG]_5$	Phillips *et al.* (2010)
13.3	156 bp	245 bp	194 bp	$[TAGA]_1TGA[TAGA]_{11}TAGG[TG]_5$	Phillips *et al.* (2010)
14 (a)	157 bp	246 bp	195 bp	$[TAGA]_{13}TAGG[TG]_5$	Lareu *et al.* (1998)
14 (b)	157 bp	246 bp	195 bp	$[TAGA]_{14}[TG]_5$	Phillips *et al.* (2010)
14.3	160 bp	249 bp	198 bp	$[TAGA]_4TGA[TAGA]_9TAGG[TG]_5$	Phillips *et al.* (2010)
15	161 bp	250 bp	199 bp	$[TAGA]_{14}TAGG[TG]_5$	Lareu *et al.* (1998)
15.3	164 bp	253 bp	202 bp	$[TAGA]_4TGA[TAGA]_{10}TAGG[TG]_5$	Lareu *et al.* (1998)
16	165 bp	254 bp	203 bp	$[TAGA]_{15}TAGG[TG]_5$	Lareu *et al.* (1998)
16.3	168 bp	257 bp	206 bp	$[TAGA]_4TGA[TAGA]_{11}TAGG[TG]_5$	Lareu *et al.* (1998)
17	169 bp	258 bp	207 bp	$[TAGA]_{16}TAGG[TG]_5$	Lareu *et al.* (1998)
17.1	170 bp	259 bp	208 bp	Not published	Schröer *et al.* (2000)
17.3	172 bp	261 bp	210 bp	$[TAGA]_4TGA[TAGA]_{12}TAGG[TG]_5$	Lareu *et al.* (1998)
18	173 bp	262 bp	211 bp	$[TAGA]_{17}TAGG[TG]_5$	Phillips *et al.* (2010)
18.3	176 bp	265 bp	214 bp	$[TAGA]_4TGA[TAGA]_{13}TAGG[TG]_5$	Lareu *et al.* (1998)
19	177 bp	266 bp	215 bp	Not published	Asamura *et al.* (2008)
19.3	180 bp	269 bp	218 bp	$[TAGA]_4TGA[TAGA]_{14}TAGG[TG]_5$	Lareu *et al.* (1998)
20.3	184 bp	273 bp	222 bp	Not published	Gamero *et al.* (2000)

from Appendix 1, J.M. Butler (2011) Advanced Topics in Forensic DNA Typing: Methodology

NIST **U.S. Population Allele Frequencies**

D1S1656 (15 different alleles)

15 different alleles

Allele	African American (n=342)	Asian (n=97)	Caucasian (n=361)	Hispanic (n=236)
10	0.0146	0.0000	0.0028	0.0064
11	0.0453	0.0309	0.0776	0.0275
12	0.0643	0.0464	0.1163	0.0890
13	0.1009	0.1340	0.0665	0.1144
14	**0.2573**	0.0619	0.0789	0.1165
14.3	0.0073	0.0000	0.0028	0.0042
15	0.1579	**0.2784**	**0.1496**	0.1377
15.3	0.0292	0.0000	0.0582	0.0508
16	0.1096	0.2010	0.1357	**0.1758**
16.3	0.1023	0.0155	0.0609	0.0508
17	0.0278	0.0722	0.0471	0.0424
17.3	0.0497	0.0876	0.1330	0.1483
18	0.0029	0.0155	0.0055	0.0064
18.3	0.0234	0.0515	0.0499	0.0254
19.3	0.0073	0.0052	0.0152	0.0042

N=1036

(only unrelated samples used; fathers removed from this sample set)

D1S1656 Characteristics

- **15 alleles** observed
- **93 genotypes** observed
- **>89% heterozygotes** (heterozygosity = 0.8890)
- **0.0224 Probability of Identity** (P_I)

$$P_I = \sum (genotype\ \ frequencies)^2$$

These values have been calculated for all 29 STR loci across the U.S. population samples examined

Loci sorted on Probability of Identity (P_I) values

Locus	Alleles Observed	Genotypes Observed	Het (obs)	P_I Value n=1036
SE33	52	304	0.9353	0.0066
Penta E	23	138	0.8996	0.0147
D2S1338	13	68	0.8793	0.0220
D1S1656	15	93	0.8890	0.0224
D18S51	22	93	0.8687	0.0258
D12S391	24	113	0.8813	0.0271
FGA	27	96	0.8745	0.0308
D6S1043	27	109	0.8494	0.0321
Penta D	16	74	0.8552	0.0382
D21S11	27	86	0.8330	0.0403
D8S1179	11	46	0.7992	0.0558
D19S433	16	78	0.8118	0.0559
vWA	11	39	0.8060	0.0611
F13A01	16	56	0.7809	0.0678
D7S820	11	32	0.7944	0.0726
D16S539	9	28	0.7761	0.0749
D13S317	8	29	0.7674	0.0765
TH01	8	24	0.7471	0.0766
Penta C	12	49	0.7732	0.0769
D2S441	15	43	0.7828	0.0841
D10S1248	12	39	0.7819	0.0845
D3S1358	11	30	0.7519	0.0915
D22S1045	11	44	0.7606	0.0921
F13B	7	20	0.6911	0.0973
CSF1PO	9	31	0.7558	0.1054
D5S818	9	34	0.7297	0.1104
FESFPS	12	36	0.7230	0.1128
LPL	9	27	0.7027	0.1336
TPOX	9	28	0.6902	0.1358

Better for mixtures (more alleles seen and less allele sharing)

Better for kinship (low mutation rate)

29 STR Loci present in STR kits

rank ordered by their variability

N=1036

(only unrelated samples used)

There are several loci more polymorphic than the **current CODIS 13 STRs**

361 Caucasians
342 African Americans
236 Hispanics
97 Asians

Probability of Identity Combinations
(assuming unrelated individuals)

STR Kit or Core Set of Loci	Total N=1036	Caucasians (n=361)	African Am. (n=342)	Hispanics (n=236)	Asians (n=97)
CODIS 13	5.02E-16	2.97E-15	1.14E-15	1.36E-15	1.71E-14
Identifiler	6.18E-19	6.87E-18	1.04E-18	2.73E-18	5.31E-17
PowerPlex 16	2.82E-19	4.24E-18	6.09E-19	1.26E-18	2.55E-17
PowerPlex 18D	3.47E-22	9.82E-21	5.60E-22	2.54E-21	7.92E-20
ESS 12	3.04E-16	9.66E-16	9.25E-16	2.60E-15	3.42E-14
ESI 16 / ESX 16 / NGM	2.80E-20	2.20E-19	6.23E-20	4.03E-19	9.83E-18
ESI 17 / ESX 17 / NGM SElect	1.85E-22	1.74E-21	6.71E-22	3.97E-21	1.87E-19
CODIS 20	9.35E-24	7.32E-23	6.12E-23	8.43E-23	4.22E-21
GlobalFiler	7.73E-28	1.30E-26	3.20E-27	2.27E-26	1.81E-24
PowerPlex Fusion	6.58E-29	2.35E-27	1.59E-28	2.12E-27	1.42E-25
All 29 autosomal STRs	2.24E-37	7.36E-35	3.16E-37	2.93E-35	4.02E-32
29 autoSTRs + DYS391	1.07E-37	3.26E-35	1.77E-37	1.29E-35	2.81E-32

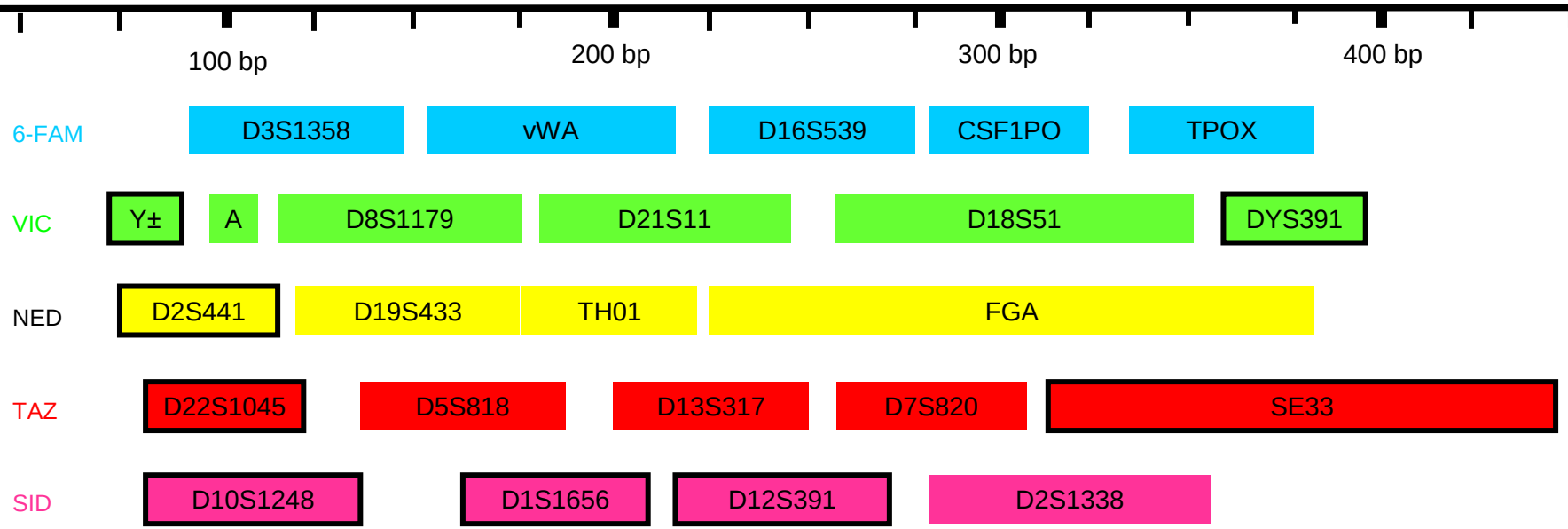
~10-13 orders of magnitude improvement for total P_I (n=1036)

New STR Multiplex Kits

Recently Launched

Life Technologies **GlobalFiler**

24plex

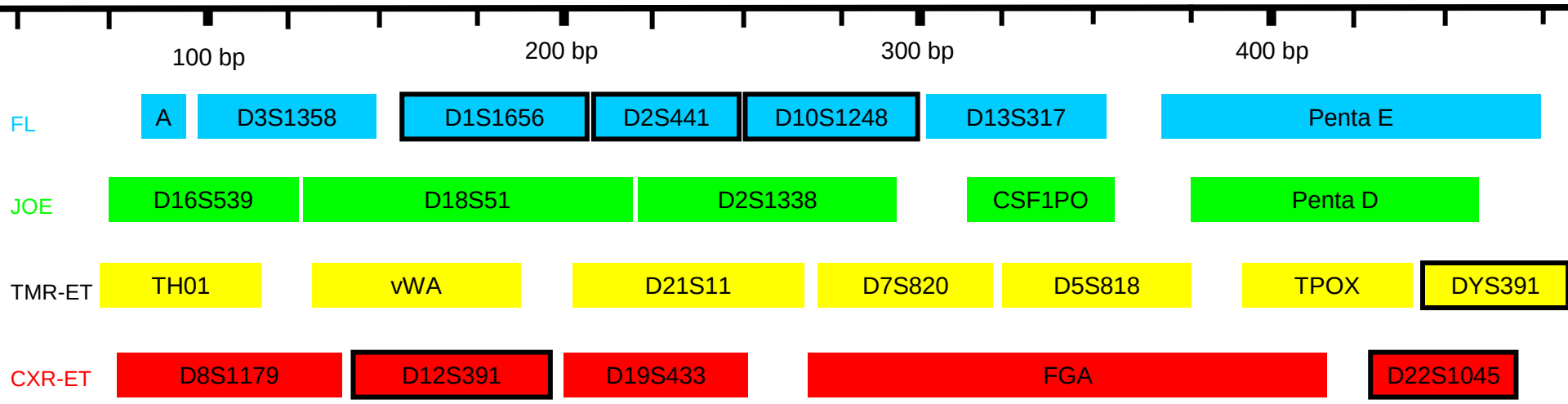


- 24 STR loci in 6 dyes (3500 and 3130 upgrade instrument use)
  - Includes SE33 and a Y-indel
- Direct amplification capabilities (two kits)
  - Single Source Samples: 40 min amplification
  - Casework Samples: 80 min amplification
- GlobalFiler gives ~12 orders of magnitude improvement using the NIST 1036 data set

http://www.invitrogen.com/site/us/en/home/Products-and-Services/Applications/Human-Identification/globalfiler_str_kit.html

Promega PowerPlex **FUSION**

24plex



- 24 STR loci in 5 dyes (3130 and 3500 instrument use)
  - Includes Penta D and E
- Direct amplification and casework capabilities: 85 min amp for both (one kit)
- PowerPlex Fusion gives ~13 orders of magnitude improvement using the NIST 1036 data set

http://www.promega.com/products/pm/genetic-identity/powerplex-fusion/

Summary

- Additional STR loci are important as DNA databases grow larger each year: the power of discrimination increases as new loci are added
  - Adding seven new loci (CODIS 13 vs CODIS 20) adds approximately 8 orders of magnitude improvement

- Commercial companies are continuing to release larger STR multiplex kits to meet the needs of the forensic community

- NIST has a set of 1036 unrelated U.S. population samples that have been used to fully characterize 29 autosomal STR loci available in commercial STR multiplex kits

Acknowledgments

NIST Funding: Interagency Agreement between the National Institute of Justice and NIST Office of Law Enforcement Standards

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 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.

Points of view are mine and do not necessarily represent the official position or policies of the US Department of Justice or the National Institute of Standards and Technology.

NIST Team for This Work


John Butler


Dave Duewer


Margaret Kline


Mike Coble

A special thanks to Applied Biosystems, Promega, and Qiagen for providing the kits used in this study

Contact Info: becky.hill@nist.gov, 301-975-4275