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

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

Carolyn R. (Becky) Hill, Michael D. Coble, Margaret C. Kline, John M. Butler, Peter M. Vallone

Email: becky.hill@nist.gov U.S. National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA

P-33

The original set of 13 Combined DNA Index System (CODIS) autosomal short tandem repeat (STR) loci are currently required for upload of DNA profiles to the U.S. national DNA database. As the number of profiles continues to increase each year, the likelihood of adventitious matches becomes greater. Expanding the core loci from 13 to 20 (including DYS391) is critical to reduce the potential of these types of matches occurring within the database, to increase international compatibility for data sharing (e.g. D1S1656, D2S441, D10S1248, D12S391, D22S1045), and to increase discrimination power in missing persons and complex kinship cases. Commercial companies have recently released next-generation STR multiplex kits (PowerPlex Fusion and GlobalFiler Express) that enable complete coverage of all of these additional loci for a total of 24 loci in each kit. These kits have been extensively tested at NIST using 3130xl and 3500 Genetic Analyzers, allowing the probability of identity calculations to be made with different sets of loci and population statistics determined with a standard set of unrelated U.S. population samples from 4 groups (n=1036). These loci have been characterized to determine the impact that this additional information will have on database searches. A variety of forensic genetic parameters have been analyzed with the NIST data set including allele frequencies, heterozygosities, peak height ratios, mutation rates, and stutter percentages.

Introduction: STR Loci and Multiplex Kits

Additional STR Loci and New Kits

Recently, Promega and Life Technologies released new STR multiplex kits to meet the needs of the planned U.S. core loci expansion (Hares D.R. 2012). PowerPlex Fusion (Promega) and GlobalFiler (Life Technologies) large multiplex kits both amplify 24 loci in a single reaction. With the launch of these new kits coverage of the previous CODIS 13 loci as well as the additional required (D1S1656, D2S441, D2S1338, D10S1248, D12S391, D19S433, and DYS391) and recommended (D22S1045) loci. PP Fusion also includes Penta D and Penta E, whereas GlobalFiler also includes SE33 and a Y-indel. At NIST, all of the loci present in commercial STR kits have been extensively tested with our population samples to assess the value of different combinations of loci present in these kits as well as their relative variability in these U.S. population samples.

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

Locus Characteristics (Example:D1S1656)

	STR Locus	Location	Repeat Motif	Allele Range*	# Alleles*
5 new European loci	**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**

*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

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

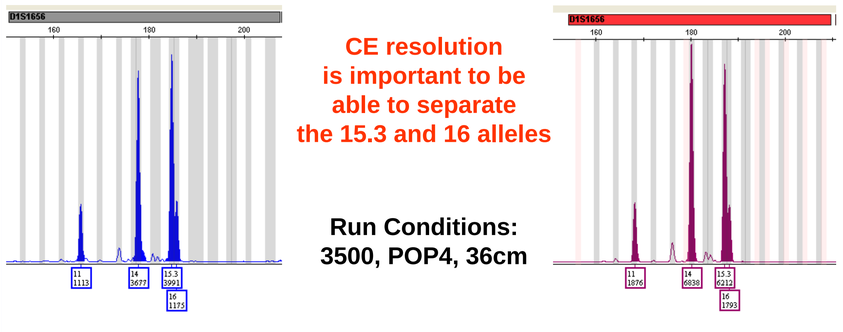
D1S1656 Allele Frequencies

15 alleles observed

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

The highest allele frequency for each population is in **bold** print

D1S1656 Mixture Profiles: 1:3 ratio, FTA spots



CE resolution is important to be able to separate the 15.3 and 16 alleles

Run Conditions: 3500, POP4, 36cm

PowerPlex Fusion 24plex STR Kit

GlobalFiler 24plex STR Kit

Materials and Methods: NIST U.S. Population Samples

Benefits of the NIST 1036 Population Samples

- **Null alleles due to primer binding site mutations were identified and removed** as a result of extensive concordance testing performed with different PCR primer sets from all available commercial STR kits
- **Ancestry testing was performed** on subsets of DNA samples with autosomal SNPs, Y-SNPs, and mtDNA sequencing to verify self-declared ancestry categorization
- **Related individuals were removed** based on autosomal STR, Y-STR and mtDNA results

NIST 1036 U.S. Population Samples

- The complete set of NIST population samples is comprised of ~1450 individuals with a subset of 1036 unrelated samples with full genotypes
- 1036 = 1032 males + 4 females
  - 361 Caucasians (2 female)
  - 342 African Americans (1 female)
  - 236 Hispanics
  - 97 Asians (1 female)

***Unrelated* samples**
All known or potentially related individuals have been removed from the 1036 subset

- Anonymous donors with self-identified ancestry
- **Complete profiles have been obtained 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)
  - SNPs (AIMs and Identity), 68 InDel markers, X-STRs (AFDIL)
  - NIST assays: miniSTRs, 26plex, >100 Y-STRs, 50 Y-SNPs

Example of identifying and eliminating related individuals

Hispanic samples ZT79994 and ZT79995

- **Full 23 Y-STR match** with PowerPlex Y23
- **Same mtDNA control region sequences**
- Out of 24 autosomal STR loci, these samples **share a total of 22 alleles at 22 loci** (only D12S391 and Penta D have non-overlapping heterozygous alleles)
- Kinship calculations
  - LR = 0 for parent-child
  - **LR = 56,300 for full-siblings (brothers)**
  - LR = 5,690 for half-siblings (or uncle-nephew, grandfather-grandson)
  - LR = 264 for first cousins
- ***Decision***: **Remove ZT79995 from final data set**
  - ZT79994 represents this individual's family in NIST 1036

NIST U.S. Population Data Summary

- The data from our 1036 U.S. population samples is currently available on STRBase:
  http://www.cstl.nist.gov/biotech/strbase/NISTpop.htm
- A summary of the NIST 1036 data set has been published in Profiles in DNA for autosomal and Y-STR loci
- The NIST U.S. population data is now included in PopStats within CODIS (COmbined DNA Index System)
- Population data announcements have been published in FSI: Genetics for
  - 29 autosomal STR loci (*Hill et al.*)
  - 23 Y-STR loci (*Coble et al.*)


Forensic Science International: Genetics 7 (2013) e82–e83

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

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

Letter to the Editor

U.S. population data for 29 autosomal STR loci

run and population statistics were confirmed using the Power-Marker v3.25 statistics program [10].

Conclusions and 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) increases random match probability (RMP) by approximately 8 orders of magnitude
  - Commercial companies have released larger STR multiplex kits to meet the needs of the forensic community
  - GlobalFiler (Life Technologies) – 24plex (including SE33 and Y-indel) gives ~12 orders of magnitude improvement
  - PowerPlex Fusion (Promega) – 24plex (including Penta D and E) gives ~13 orders of magnitude improvement
- **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**

Results: Characterization of STR Loci

Autosomal STR Locus Diversity with NIST 1036 Population Samples

Data analysis to determine individual locus diversity for each of the 29 STR loci present in commercial kits was performed with an Excel-based software tool (STR_Genotype) developed to calculate allele and genotype frequencies and heterozygosities observed from the NIST 1036 data set as well as the probability of identity values reported below.

Software programs available on STRBase:
http://www.cstl.nist.gov/biotech/strbase/software.htm

Probability of Identity

- The probability of identity (P_I), also referred to as the matching probability, is **the chance that two unrelated people selected at random will have the same genotype** (first described by George Sensabaugh in 1982). The P_I value of a single locus is determined by summing the square of the observed genotype frequencies.

$$\sum_{i=1}^{n} x_i^2 \quad where\ x_i\ is\ the\ genotype\ frequency$$

- **Lower P_I values indicate more variability** with the genetic marker in the measured population because there are more genotypes occurring at a lower frequency.
- P_I values from independently inherited loci can be multiplied together to produce an expected profile P_I

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

The CODIS 13 core STR loci are in black and additional loci are highlighted in blue

STR Loci Diversity

- SE33 is the most variable locus with the highest Het_{obs} (0.9353) and lowest P_I value (0.0066).
- SE33 exhibited the greatest number of alleles and genotypes (twice as many compared to the next highest ranked locus Penta E).
- TPOX is the least variable locus with the lowest Het_{obs} (0.6902) and highest P_I value (0.1358).
- Two of the new CODIS loci (D2S1338 and D1S1656) rank higher than the highest ranked CODIS 13 marker (D18S51).

Probability of Identity Loci Combinations (assuming locus independence)

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

~8-13 orders of magnitude improvement for total P_I (n=1036)

Funding This project was supported by an interagency agreement between NIJ and the NIST Law Enforcement Standards Office. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the US Department of Justice. 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.

Poster available for download from STRBase http://www.cstl.nist.gov/biotech/strbase/pub_pres/HillISFG2013poster.pdf