Population Statistics on the Proposed Expanded U.S. Core Loci

Carolyn R. (Becky) Hill, Margaret C. Kline, David L. Duewer, Michael D. Coble and John M. Butler

National Institute of Standards and Technology (NIST), 100 Bureau Drive, Gaithersburg, MD 20899-8314

Email: becky.hill@nist.gov
Phone: 301-975-4275

Poster # 84 at:
23rd Int Symposium on Human ID (ISHI), Nashville, TN (October 15-18, 2012)

Poster available for download from STRBase:
http://www.cstl.nist.gov/strbase/pub_pres/Hill-ISHI2012-STRloci.pdf

The original core set of 13 Combined DNA Index System (CODIS) autosomal short tandem repeat (STR) loci were selected in November 1997 and are required by the FBI for upload of DNA profiles to the national DNA database [1,2]. As the number of profiles stored in the National DNA Index System (NDIS) continues to increase each year (>11.5 million total profiles), the likelihood of adventitious matches becomes greater [3,4]. Expanding beyond the 13 core loci is critical to reduce the potential of these types of matches occurring within the database, to increase international compatibility for data sharing, and to increase discrimination power in missing persons cases [3,4]. In November 2009, the European Union adopted five new autosomal STR loci as part of their expanded European Standard Set (ESS), including D12S391, D1S1656, D2S441, D10S1248, and D22S1045 [5,6]. All five of these loci are being considered for the expansion of the U.S. core set to provide greater capabilities for international comparisons when necessary. Also, D2S1338 and D19S443 are recommended as two new additions to the original 13 core loci because almost half of the U.S. national database already contains data for these loci. Finally, it has been suggested that the DYS391 locus be added to confirm amelogenin null alleles sometimes present in DNA profiles [3,4]. In the past few years, Promega Corporation and Life Technologies have released several new next generation STR multiplex kits that enable complete coverage of all of these additional loci. These multiplex kits have been extensively tested at National Institute of Standards and Technology (NIST), allowing the probability of identity (PI) calculations to be made with different sets of loci and population statistics, including allele frequencies and heterozygosity values for each locus, determined with our set of unrelated U.S. population samples (n=1036). With this information, it has been possible to thoroughly characterize these new STR loci beyond the original 13 CODIS core loci to determine the impact that this additional information will have on database searches. A summary of these results, including STR locus population statistics for the new STR loci, are shown in order to help assess the benefits of adding additional loci to the current 13 CODIS core loci.

References:
[1] Budowle, B., et al. (1998). CODIS and PCR-based short tandem repeat loci: law enforcement tools. *Proceedings of the Second European Symposium on Human Identification*, pp. 73-88. Madison, Wisconsin: Promega Corporation. Available at http://www.promega.com/geneticidproc/eusymp2proc/17.pdf.
[2] Butler, J.M. (2006) Genetics and genomics of core short tandem repeat loci used in human identity testing. *J. Forensic Sci., 51*, 253-265.
[3] Hares, D.R. (2012a) Expanding the CODIS core loci in the United States. *Forensic Sci. Int. Genet.* 6(1):e52-4.
[4] Hares, D.R. (2012b) Addendum to expanding the CODIS core loci in the United States. *Forensic Sci. Int. Genet.* 6(5):e135.
[5] Gill, P., et al. (2006) The evolution of DNA databases-Recommendations for new European STR loci. *Forensic Sci. Int.* 156: 242-244.
[6] Gill, P., et al. (2006) New multiplexes for Europe-amendments and clarification of strategic development. *Forensic Sci. Int.* 163: 155-157.
[7] Hill, C.R., et al. (2010) Strategies for concordance testing. *Profiles in DNA (Promega)*, 13(1).
[8] Hill, C.R., et al (2011) Concordance testing comparing STR multiplex kits with a standard data set. *Forensic Sci. Int. Genet.: Suppl. Ser.* 3: e188-e189.
[9] Butler, J.M., et al. (2012) Variability of new STR loci and kits in U.S. population groups. *Profiles in DNA. (in press).*

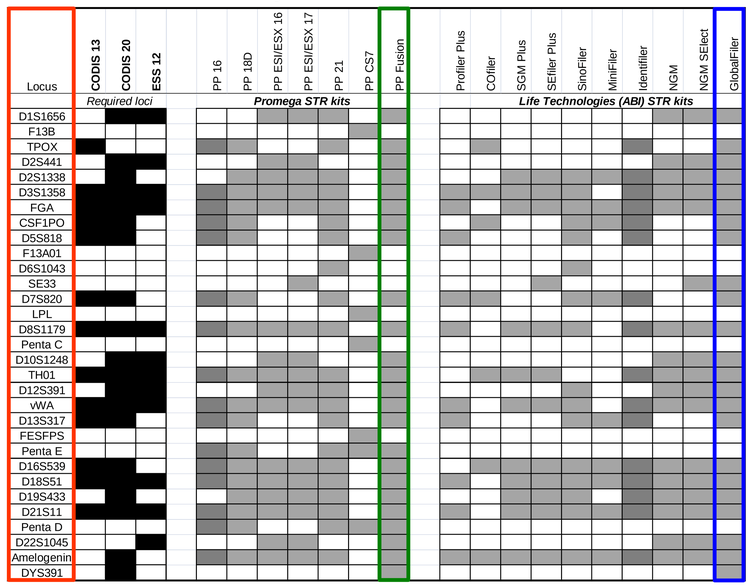
NIST Funding: Interagency Agreement 2010-DN-R-7121 between the National Institute of Justice and NIST Office of Law Enforcement Standards.

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 [3,4]. 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 sample set 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.

Commonly Used STR kits World-Wide



29 autosomal STR loci — **PowerPlex Fusion 24plex STR Kit** — **GlobalFiler 24plex STR Kit**

Locus Characteristics (Example:D1S1656)

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
5 new European loci: 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**

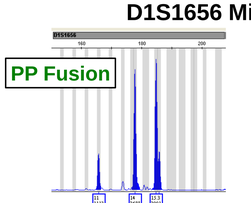
*Allele range and number of observed alleles from Appendix 1, J.M. Butler (2011) Advanced Topics in Forensic DNA Typing: Methodology; ‡SE33 alleles have complex repeat structure

D1S1656 AlleleFrequencies

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

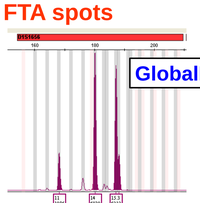
15 alleles observed

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



Materials and Methods: NIST US Population Samples

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

NIST 1036 Unrelated US Population Samples

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

***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)

YSTR data available at Poster #100

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

Example of identifying and eliminating related samples

Hispanic samples ZT79994 and ZT79995

- 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)
- **Full 23 Y-STR match** with PowerPlex Y23
- **Same mtDNA control region sequences**
- 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

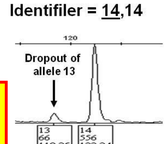
Concordance Testing with Samples [7,8]

- Many of these STR kits have different primer sequences for amplifying the same STR locus
- Need to analyze the same DNA samples with different STR typing kits looking for differences
- In some rare cases, allele dropout may occur due to mutations in primer binding regions

Current Total Concordance %

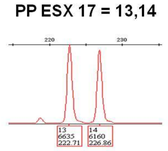
1,176,994 allele comparisons
1,225 total differences
99.90% concordance

Identifiler = 14,14



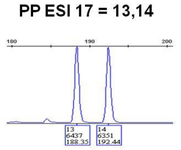
PHR = 11.9%

PP ESX 17 = 13,14



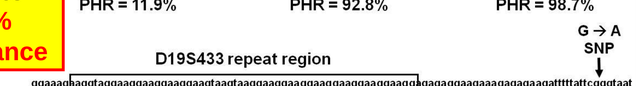
PHR = 92.8%

PP ESI 17 = 13,14



PHR = 98.7%

D19S433 repeat region



Results and Conclusions: Characterization of STR Loci

Autosomal STR Locus Diversity with 1036 NIST 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 developed by Dave Duewer at NIST 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 [9]

- 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

STR Loci Diversity

- SE33 is the most variable locus with the highest Het_{obs} (0.9353), lowest P_I value (0.0066), and most amount of alleles and genotypes observed by over double as compared to the next highest ranked locus Penta E.
- TPOX is the least variable locus with the lowest Het_{obs} (0.0.6902) and highest P_I value (0.1358),
- Two of the new CODIS required 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

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

The genotypes and allele frequencies are available for download on STRBase: http://www.cstl.nist.gov/strbase/NIST pop.htm

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) 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
  - 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