Characteristics of 24 Commonly Used Autosomal STR Loci
and U.S. Population Data with the Recently Announced Expanded CODIS Core Loci

John M. Butler, Carolyn R. (Becky) Hill, David L. Duewer, Margaret C. Kline, and Kristen L. O'Connor*

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

*Current Address: Booz Allen Hamilton; klewisoconnor@gmail.com

Email: john.butler@nist.gov
Phone: 301-975-4049

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

Poster #38 at 22nd International Symposium on Human Identification, National Harbor, MD, October 4-5, 2011

Acknowledgments: Funding support from the National Institute of Justice through interagency agreement 2008-DN-R-121 to the NIST Office of Law Enforcement Standards. We thank Applied Biosystems, Promega Corporation, and Qiagen for supplying STR typing kits used for concordance testing purposes. Feedback on this work from Doug Hares of the FBI Laboratory's CODIS Unit is also appreciated.

Current commercially available short tandem repeat (STR) typing kits provide capabilities for analysis of 24 different autosomal loci including the 13 CODIS (Combined DNA Index System) STR loci [1,2] plus D2S1338, D19S433, Penta D, Penta E, D12S391, D1S1656, D2S441, D10S1248, D22S1045, D6S1043, and SE33. Many of these additional STR loci are part of the expanded European Standard Set adopted in 2009 [3]. In the past two years, Promega Corporation and Applied Biosystems have released new STR kits to enable coverage of these additional loci. In 2010, Qiagen also began supplying STR typing kits in some parts of the world. Earlier this year the National DNA Index System (NDIS), which currently contains the 13 CODIS STR loci, exceeded 10 million DNA profiles [4]. The large amount of legacy data in the U.S. makes it difficult to move to completely different typing systems, such as single nucleotide polymorphisms or a non-overlapping set of STR loci, because previously collected samples would have to be re-genotyped with the new markers. In April 2011, the NDIS Custodian announced a plan to expand the CODIS core loci in the United States to 20 required STR loci plus amelogenin with 4 optional loci [5]. The additional required loci include amelogenin, D2S1338, D19S433, D12S391, D1S1656, D2S441, D10S1248, Penta E, and the Y-chromosome locus DYS391. TPOX, the least polymorphic STR locus of the current CODIS core loci, has been made optional. Data from over 500 U.S. population samples were evaluated across the 24 STR loci using Applied Biosystems (Identifiler, MiniFiler, NGM, NGM SElect), Promega (PowerPlex 16, PowerPlex ESI 17, PowerPlex ESX 17, PowerPlex 18D), and Qiagen (ESSplex, ESSplex SE, IDplex) kits. Allele ranges and locus characteristics for the 24 common STR loci are discussed. The probability of identity (P_I) with different sets of loci are illustrated to help assess the benefits of adding loci to the current 13 CODIS core loci.

References

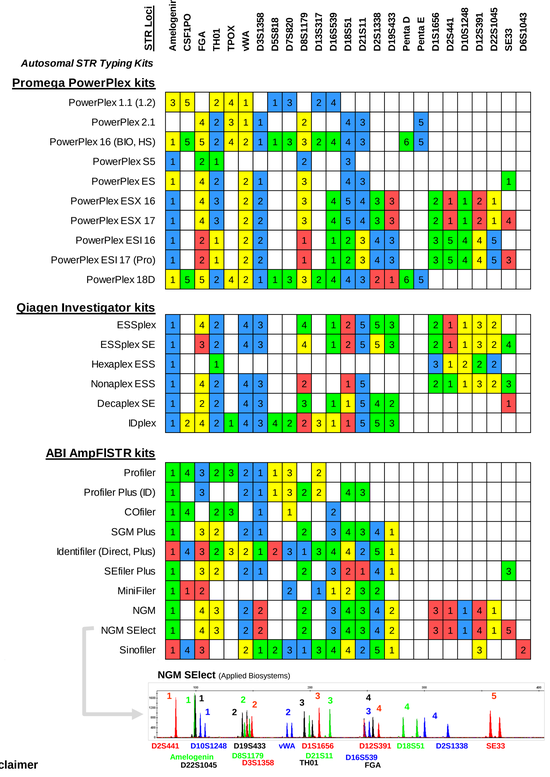
[1] Budowle, B., et al. (1998). CODIS and PCR-based short tandem repeat loci: law enforcement tools. *Proceedings of the Second European Symposium on Human Identification*, pp. 73-88. Madison, Wisconsin: Promega Corporation. Available at http://www.promega.com/products/pm/genetic-identity/ishi-conference-proceedings/2nd-eshi-oral-presentations/.
[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] Schneider, P. (2009). Expansion of the European Standard Set of DNA database loci – the current situation. *Profiles in DNA* 12(1): 6-7.
[4] http://www.fbi.gov/about-us/lab/codis/ndis-statistics
[5] Hares, D.R. (2011). Expanding the CODIS core loci in the United States. *Forensic Sci. Int. Genet.* (in press), doi:10.1016/j.fsigen.2011.04.012
[6] Butler, J.M. (2012). *Advanced Topics in Forensic DNA Typing: Methodology.* Elsevier Academic Press: San Diego.
[7] Thanakiatkrai, P. & Welch, L. (2011). Evaluation of nucleosome forming potentials (NFPs) of forensically important STRs. *Forensic Sci. Int. Genet* 5: 285-290.
[8] Y Chromosome Haplotype Reference Database. Available at http://www.yhrd.org/.
[9] 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.

United States / Europe

Proposed April 2011 Extended U.S. Core	PP16	Identifiler	MiniFiler	ESSplex SE ESX/ESI17	NGM (SElect)	SEfiler Plus	SGM Plus
	TPOX	TPOX					
CSF1PO	CSF1PO	CSF1PO	CSF1PO				
D5S818	D5S818	D5S818					
D7S820	D7S820	D7S820	D7S820				
D13S317	D13S317	D13S317	D13S317				
FGA	FGA	FGA	FGA	FGA	FGA	FGA	FGA
vWA	vWA	vWA		vWA	vWA	vWA	vWA
D3S1358	D3S1358	D3S1358		D3S1358	D3S1358	D3S1358	D3S1358
D8S1179	D8S1179	D8S1179		D8S1179	D8S1179	D8S1179	D8S1179
D18S51	D18S51	D18S51	D18S51	D18S51	D18S51	D18S51	D18S51
D21S11	D21S11	D21S11	D21S11	D21S11	D21S11	D21S11	D21S11
TH01	TH01	TH01		TH01	TH01	TH01	TH01
D16S539	D16S539	D16S539	D16S539	D16S539	D16S539	D16S539	D16S539
D2S1338		D2S1338	D2S1338	D2S1338	D2S1338	D2S1338	D2S1338
D19S433		D19S433		D19S433	D19S433	D19S433	D19S433
D12S391				D12S391	D12S391		
D1S1656				D1S1656	D1S1656		
D2S441				D2S441	D2S441		
D10S1248				D10S1248	D10S1248		
				D22S1045	D22S1045		
				SE33	(SE33)	SE33	
	Penta D						
Penta E	Penta E						
DYS391							
Amelogenin	Amelogenin	Amelogenin	Amelogenin	Amelogenin	Amelogenin	Amelogenin	Amelogenin

Loci Present in Commercial STR Kits

33 commercially available STR typing kits (left side) are shown with the 24 autosomal STR loci and amelogenin (across the top) with dye label colors and size positions (grid positions)



Disclaimer

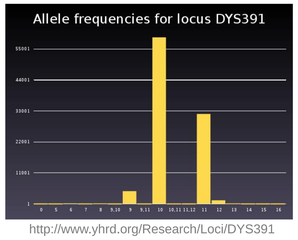
Points of view are those of the authors and do not necessarily represent the official position or policies of the US Department of Justice. Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such 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. While John Butler is a member of the FBI CODIS Core Loci Working Group, the information on this poster in no way reveals confidential information he received as a member of that group nor does it reflect or is it intended to reflect the opinions of the CODIS Core Loci Working Group.

STR Locus Characteristics

Current 13 CODIS STR loci are in bold font and shaded

STR Locus	Chromosome Location	Physical Position GRCh37 assembly (from ref. [7])	GenBank Accession (allele repeat #)	*Repeat Category* Repeat Motif	Allele Range (from ref. [6])	# Observed Alleles (from ref. [6])
D1S1656	1q42	Chr 1 230.905 Mb	G07820 (15.3)	*compound* TAGA	8 to 20.3	25
TPOX	2p25.3 thyroid peroxidase, 10th intron	Chr 2 1.493 Mb	M68651 (11)	*simple* AATG	4 to 16	19
D2S441	2p14	Chr 2 68.239 Mb	AC079112 (12)	*compound* TCTA/TCAA	8 to 17	22
D2S1338	2q35	Chr 2 218.879 Mb	AC010136 (23)	*compound* TGCC/TTCC	10 to 31	40
D3S1358	3p21.31	Chr 3 45.582 Mb	AC099539 (16)	*compound* TCTA/TCTG	6 to 26	44
FGA	4q31.3 alpha fibrinogen, 3rd intron	Chr 4 155.509 Mb	M64982 (21)	*compound* CTTT/TTCC	12.2 to 51.2	95
D5S818	5q23.2	Chr 5 123.111 Mb	AC008512 (11)	*simple* AGAT	4 to 29	24
CSF1PO	5q33.1 c-fms proto-oncogene, 6th intron	Chr 5 149.455 Mb	X14720 (12)	*simple* AGAT	5 to 17	29
SE33 (ACTBP2)	6q14 beta-actin related pseudogene	Chr 6 88.987 Mb	V00481 (26.2)	*complex* AAAG	3 to 49	**178**
D6S1043	6q15	Chr 6 92.450 Mb	G08539 (11)	*compound* AGAT/AGAC	8 to 25	25
D7S820	7q21.11	Chr 7 83.789 Mb	AC004848 (13)	*simple* GATA	5 to 16	33
D8S1179	8q24.13	Chr 8 125.907 Mb	AF216671 (13)	*compound* TCTA/TCTG	6 to 20	34
D10S1248	10q26.3	Chr 10 131.093 Mb	AL391869 (13)	*simple* GGAA	7 to 19	13
TH01	11p15.5 tyrosine hydroxylase, 1st intron	Chr 11 2.192 Mb	D00269 (9)	*simple* TCAT	3 to 14	24
VWA	12p13.31 von Willebrand Factor, 40th intron	Chr 12 6.093 Mb	M25858 (18)	*compound* TCTA/TCTG	10 to 25	36
D12S391	12p13.2	Chr 12 12.450 Mb	G08921(20)	*compound* AGAT/AGAC	13 to 27.2	52
D13S317	13q31.1	Chr 13 82.692 Mb	AL353628 (11)	*simple* TATC	5 to 17	22
Penta E	15q26.2	Chr 15 97.374 Mb	AC027004 (5)	*simple* AAAGA	5 to 32	53
D16S539	16q24.1	Chr 16 86.386 Mb	AC024591 (11)	*simple* GATA	4 to 17	29
D18S51	18q21.33	Chr 18 60.949 Mb	AP001534 (18)	simple AGAA	5.3 to 40	73
D19S433	19q12	Chr 19 30.416 Mb	AC008507 (14)	*compound* AAGG/TAGG	5.2 to 20	36
D21S11	21q21.1	Chr 21 20.554 Mb	AP000433 (29.1)	*complex* TCTA/TCTG	12 to 43.2	90
Penta D	21q22.3	Chr 21 45.056 Mb	AP001752 (13)	*simple* AAAGA	1.1 to 19	50
D22S1045	22q12.3	Chr 22 37.536	AL022314 (17)	*simple* ATT	7 to 20	14
DYS391	Yq11.21	Chr Y 14.103 Mb	AC011302 (11)	*simple* TCTA	5 to 16	17
Amelogenin	Xp22.2 Yp11.2	Chr X: 11.315 Mb Chr Y: 6.738 Mb	M55418 M55419		*Not applicable*	

YHRD (Y-chromosome Haplotype Reference Database) **data from 97,575 samples**



http://www.yhrd.org/Research/Loci/DYS391

Why Consider DYS391?

DYS391 is located on the long arm of the Y-chromosome over 7 Mb away from amelogenin. Thus, it is likely to be detected in the event of an amelogenin Y deletion that could make a male sample falsely appear as a female (X,-).

DYS391 is not very polymorphic. From a data set of 97,575 haplotypes available on the Y-Chromosome Haplotype Reference Database [8], over half of them possess allele 10. However, only two null alleles have been reported and 0.01% duplication events (11 total) have been seen in over 700 different population groups from around the world. Thus, it is a stable locus with a relatively narrow allele range.

DYS391 has a mutation rate of 0.26%, which is comparable to most autosomal STRs commonly in use. There have been 38 mutations observed so far in the 14,621 meioses reported in the literature and compiled on YHRD.

Current 13 CODIS STR loci are in black font

STR Locus	Alleles Observed	Genotypes Observed	P_I value N = 938
SE33	53	292	0.0069
Penta E*	20	114	0.0177
D2S1338	13	68	0.0219
D1S1656	15	92	0.0220
D18S51	21	91	0.0256
D12S391	23	110	0.0257
FGA	26	93	0.0299
D6S1043*	23	88	0.0348
Penta D*	16	71	0.0356
D21S11	25	81	0.0410
D19S433	16	76	0.0561
D8S1179	11	45	0.0582
vWA	11	38	0.0622
D7S820	11	32	0.0734
TH01	8	24	0.0784
D16S539	9	28	0.0784
D13S317	8	29	0.0812
D10S1248	12	39	0.0837
D2S441	14	41	0.0855
D3S1358	11	30	0.0873
D22S1045	11	42	0.0933
CSF1PO	9	30	0.1071
D5S818	9	34	0.1192
TPOX	9	28	0.1283

*Penta D, Penta E, and D6S1043 currently have been examined with <938 samples

Results and Discussion

STR locus allele frequencies have been generated from NIST U.S. samples as previously described during concordance testing [9]. In total, over 1450 samples have been examined with more than a dozen different STR kits. Information from various subsets of this data are shown on this poster.

The table on the left ranks the 24 autosomal STR loci by their probability of identity, which is a measure of locus diversity determined by squaring the sum of the observed genotype frequencies with a lower number indicating a higher power of discrimination.

The table on the top right calculates the random match probabilities (RMP) values for various combinations of loci. Impact of new loci on kinship calculations is illustrated with **the figure on the bottom right**. Adding more loci improves kinship associations in parent-offspring LRs.

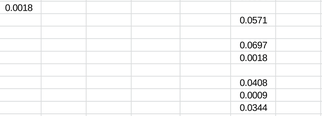
The table below displays the observed allele frequencies from a combined set of 552 U.S. samples consisting of 249 African Americans, 162 Caucasians, 139 Hispanics, and 2 Asians. The most common allele for each locus is listed in bold font.

STR Marker Combinations	RMP*	1 in …
13 CODIS STRs	**6.0E-16**	**1.7E+15**
15 STRs (+D2S1338, D19S433)	7.3E-19	1.4E+18
18 STRs (+D2S441, D10S1248, D22S1045)	4.9E-22	2.0E+21
20 STRs (+D1S1656, D12S391)	2.8E-25	3.6E+24
23 STRs (+SE33, Penta D, Penta E)	**1.2E-30**	**8.4E+29**

14 orders of magnitude improvement

*Average RMP calculated by combining Probability of Identity values for each locus

Parent-Offspring Likelihood Ratio (LR) Distributions



LR distributions produced from 1000 simulations using DNA View software and U.S. Caucasian allele frequencies

Observed Allele Frequencies for 552 U.S. Population Samples

Newly proposed autosomal STR loci are outlined in red (see ref. [5])

Allele	TPOX	CSF1PO	D5S818	D7S820	D13S317	FGA	vWA	D3S1358	D8S1179	D18S51	D21S11	TH01	D16S539	D2S1338	D19S433	D12S391	D1S1656	D2S441	D10S1248	D22S1045	SE33	Penta D	Penta E	D6S1043
2.2																						0.0598		
3.2																						0.0045		
5												0.0018										0.0181	0.0752	
6	0.0471			0.0009								0.1748										0.0072	0.0009	
6.3																					0.0009			
7	0.0100	0.0281	0.0118	0.0136								**0.3098**									0.0009	0.0190	0.1259	
8	**0.4420**	0.0299	0.0263	0.1821	0.0797				0.0072			0.1377	0.0308							0.0018	0.0045	0.0616	0.0779	
8.1				0.0009																				
9	0.1413	0.0245	0.0444	0.1341	0.0779				0.0054	0.0018		0.1422	0.1594		0.0009					0.0018		**0.2065**	0.0317	0.0009
9.1																								
9.3				0.0009								0.2255												
10	0.0661	0.2409	0.0625	**0.3007**	0.0489				0.0743	0.0063		0.0072	0.0969		0.0045		0.0100	0.1839	0.0018	0.0217		0.1051	0.0661	0.0054
10.2																					0.0009			
10.3		0.0009		0.0009																				
11	0.2409	0.2763	0.2980	0.2101	0.2971			0.0009	0.0607	0.0091		0.0009	**0.3025**		0.0317		0.0516	**0.3496**	0.0163	0.1132		0.1594	0.0788	0.2165
11.3																		0.0543						
12	0.0516	**0.3306**	**0.3668**	0.1286	**0.3207**		0.0009		0.1513	0.0960			0.2464		0.0924		0.0915	0.0987	0.0833	0.0317	0.0045	0.1558	**0.1658**	**0.2301**
12.2															0.0208						0.0027			
12.3																		0.0036						
13	0.0009	0.0598	0.1757	0.0272	0.1313		0.0045	0.0027	**0.2645**	0.0861			0.1449		0.2491		0.0933	0.0326	0.2663	0.0072	0.0127	0.1368	0.0987	0.1060
13.2										0.0018					0.0353						0.0027			
13.3																		0.0009						
14		0.0082	0.0118		0.0426		0.0861	0.0924	0.2437	0.1123			0.0190		**0.3134**	0.0009	0.1594	0.2346	**0.2926**	0.0598	0.0389	0.0498	0.0688	0.0643
14.2															0.0507						0.0045			
14.3																	0.0045	0.0009						
15		0.0009	0.0027		0.0018		0.1630	0.2844	0.1413	**0.1739**			0.0009	0.1051	0.0589	**0.1594**	0.0389	0.2065	**0.3306**	0.0471	0.0109	0.0625	0.0389
15.2								0.0009		0.0009				0.0462						0.0027				
15.3																	0.0444						0.0009	
16							0.2391	**0.2926**	0.0471	0.1486			0.0480	0.0217	0.0543	0.1286	0.0009	0.1024	0.2772	0.0525	0.0036	0.0534	0.0082	
16.2					0.0009									0.0208						0.0027				
16.3																0.0870				0.0018				
17							**0.2455**	0.2056	0.0036	0.1341			0.1467	0.0027	0.1304	0.0335	0.0009	0.0236	0.1404	0.0815	0.0018	0.0426	0.0516	
17.1															0.0027									
17.2														0.0027						0.0009				
17.3															0.0136	0.0897				0.0027				
18					0.0118	0.1694	0.1105	0.0009	0.0942				0.0688		**0.1993**	0.0027		0.0018	0.0109	0.0996		0.0236	0.0978	
18.1															0.0009								0.0009	
18.2					0.0054									0.0018										
18.3															0.0163	0.0362							0.0009	
19					0.0625	0.0761	0.0091		0.0670				**0.1449**		0.1413			0.0018	0.0027	**0.1051**		0.0118	0.0879	
19.1															0.0054									
19.2					0.0018															0.0036				
19.3															0.0072	0.0082							0.0018	
20					0.0842	0.0118	0.0009		0.0471				0.1214		0.1178					0.0679		0.0118	0.0525	
20.1															0.0018									
20.2																				0.0072				
20.3															0.0009								0.0018	
21					0.1467	0.0036			0.0091				0.0870		0.1005					0.0290		0.0009	0.0109	
21.2					0.0018				0.0009											0.0163				
21.3																							0.0100	
22					**0.1875**				0.0091				0.0851		0.0534					0.0154			0.0018	
22.2					0.0072										0.0009					0.0217				
22.3					0.0009																		0.0045	
23					0.1495				0.0009				0.1132		0.0516					0.0054			0.0036	
23.2					0.0036															0.0254				
23.3																							0.0009	
24					0.1386				0.0009				0.0870		0.0226					0.0009			0.0027	
24.2					0.0009					0.0009										0.0199				
24.3															0.0009									
25					0.1042								0.0743		0.0136								0.0027	
25.2										0.0009										0.0290				
26					0.0580					0.0009			0.0208		0.0027					0.0009				
26.2																				0.0534				
27					0.0236					0.0480			0.0018		0.0018									
27.2																						0.0571		
28					0.0054					0.1884														
28.2																						0.0697		
28.3																						0.0018		
29					0.0018					0.1993														
29.2										0.0027												0.0408		
30					0.0018					**0.2201**												0.0009		
30.2					0.0009					0.0245												0.0344		
31										0.0824												0.0018		
31.2					0.0009					0.0770												0.0190		
32										0.0072														
32.2										0.0897												0.0054		
33										0.0036												0.0009		
33.1										0.0018														
33.2										0.0290												0.0027		
34										0.0036												0.0027		
34.2										0.0018												0.0009		
35										0.0118														
36										0.0045														
38										0.0009														
39										0.0009														
H_{obs}	0.716	0.745	0.719	0.804	0.746	0.886	0.824	0.770	0.774	0.875	0.844	0.750	0.777	0.882	0.803	0.880	0.880	0.777	0.797	0.772	0.928	0.879	0.871	0.864

Poster available for download from STRBase: http://www.cstl.nist.gov/biotech/strbase/pub_pres/**ButlerISHI2011poster.pdf**