Announcement of population data

Allele frequencies for 27 Y-STR loci with U.S. Caucasian, African American, and Hispanic samples

John M. Butler *, Amy E. Decker, Peter M. Vallone, Margaret C. Kline

Biotechnology Division, National Institute of Standards and Technology, Gaithersburg, MD 20899-8311, USA

Abstract

A total of 263 U.S. Caucasians, 260 African Americans and 140 U.S. Hispanics or a subset of 31 Caucasians, 32 African Americans, and 32 Hispanics were typed for 27 Y-chromosome short tandem repeat (Y-STR) markers: DYS444, DYS446, DYS449, DYS463, DYS485, DYS490, DYS495, DYS504, DYS505, DYS508, DYS520, DYS522, DYS525, DYS532, DYS533, DYS534, DYS540, DYS556, DYS557, DYS570, DYS575, DYS576, DYS594, DYS632, DYS635, DYS641, and DYS643. Allele frequencies for each locus are reported along with nomenclature based on sequence analysis.
© 2005 Elsevier Ireland Ltd. All rights reserved.

Population samples: Anonymous liquid blood samples with self-identified ethnicities were purchased from Interstate Blood Bank, Inc. (Memphis, TN) and Millennium Biotech, Inc. (Ft. Lauderdale, FL). All samples were previously examined with 15 autosomal short tandem repeats and the amelogenin sex-typing marker using the AmpFlSTR Identifiler kit (Applied Biosystems, Foster City, CA) to verify that each sample was unique [1].

N: Two hundred and sixty U.S. Caucasians, 260 African Americans, and 140 U.S. Hispanics were typed for 27 Y-chromosome short tandem repeat (Y-STR) markers. For six of the loci only a subset of each population group was examined that included 31 Caucasians, 32 African Americans, and 32 Hispanics.

DNA extraction: Blood samples were extracted using a modified salting out procedure [2] as described previously [1].

* Corresponding author. Tel.: +1 301 975 4049;
fax: +1 301 975 8505.

E-mail address: john.butler@nist.gov (J.M. Butler).

Quantification: Extracted DNA was quantified using UV spectrophotometry followed by a PicoGreen assay [3] to adjust concentrations to approximately 1 ng/μL (see [1]).

Y-STR markers: Table 1 lists information on the 27 Y-STR markers examined as part of this study (see also http://www.cstl.nist.gov/biotech/strbase/y_strs.htm). The PCR primer sequences used in this study were those present in the Genome Database (GDB; http://www.gdb.org). In an attempt to generate smaller PCR products for some of the Y-STR loci, new primers were designed for DYS485 (−129 bp), DYS490 (−38 bp), and DYS495 (−77 bp) with the size reduction relative to the GDB generated amplicons shown in parentheses. Fluorescent dyes 6-FAM, VIC, or NED were included on the 5′-end of the forward primers to create blue, green, or yellow-labeled PCR products. The reverse primers included an extra seven base tail (shown in lower case letters) to promote non-template addition. The DYS635 (Y-GATA-C4) data were obtained as part of an evaluation of a beta-test version of the Yfiler kit (Applied Biosystems). The exact chromosomal locations were ascertained using

Table 1
Information on the 27 Y-STR loci examined in this study

Y-STR locus	Primer sequence (5′ → 3′)	GenBank accession (repeat position)	GenBank reference Allele	Y-chromosome position (Mb)
DYS444	NED-TCTAAGGGATCCAAAGGCAGAA gtttcttGTGTGAACCATTTGGCATGTTTA	AC007043 (nt75953…76008)	**(TAGA)14**	17.664
DYS446	FAM-TATTTTCAGTCTTGTCCTGTC gtttcttAAATGTATGGCCAACATAGCAAAACCA	AC006152 (nt21421…21490)	**(TCTCT)14**	3.175
DYS449	NED-CCTGGAAGTGGAGTTTGCTGT gtttcttTGGAGTCTCTCAAGCCTGTTCTA	AC051663 (nt93299…93464)	(TTTC)15… (TTTC)14 = **29**	8.261
DYS463	FAM-AATTCTAGGTTTGAGCAAAGACA gtttcttATGAGGTTGTGTGACTTGACTG	AC007275 (nt64620…64739)	(AAAGG)7(AAGGG)15 (AAGGA)2 = **24**	7.687
DYS485	NED-CCTGGGTGACAAGAGTTATACTCT gtttcttGCAGACTTCGCCACTACATAAT	AC009233 (nt86438…86485)	**(TTA)16**	20.488
DYS490	VIC-CTGAGCTGAGATCACGCC gtttcttACGATATGAAAAAGCAGTATGTCCT	AC019058 (nt32127…32162)	**(TTA)12**	3.487
DYS495	FAM-AGCAAACTTTGAAGCCAGAAAG gtttcttCTTGGGCAACAGAGCGAGA	AC004474 (nt49600…49644)	**(AAT)15**	13.449
DYS504	FAM-TCTACACCACTGTGCCAAGC gtttcttGGCAACAGAGCAACCCTCT	AC006157 (nt11761…117832)	**(TCCT)18**	2.946
DYS505	FAM-TCTGGCGAAGTAACCCAAAC gtttcttTCGAGTCAGTTCACCAGAAGG	AC012078 (nt50281…50328)	**(TCCT)12**	3.684
DYS508	VIC-ACAATGGCAATCCCAAATTC gtttcttGAACAAATAAGGTGGGATGGAT	AC006462 (nt29187…29230)	**(TATC)11**	16.232
DYS520	NED-AACAGCCTGCCCAACATAGT gtttcttACCATCATGCCCTGCAATA	AC007275 (nt151532…151622)	(ATAG)10 (ATAC)10 = **20**	7.774
DYS522	VIC-CCTTTGAAATCATTCATAATGC gtttcttTCATAAACAGAGGGTTCTGG	AC007247 (nt70433…70472)	**(GATA)10**	7.459
DYS525	VIC-ATTCACACCATTGCACTCCA gtttcttCCATCTGTTTATCTTCCCATCA	AC010104 (nt92809…92848)	**(TAGA)10**	7.120
DYS532	NED-TTGGTTTTATGCCTTTCACT gtttcttTAGGTGACAGAGCAGGATTC	AC016991 (nt68764…68819)	**(CTTT)14**	8.423
DYS533	VIC-CATCTAACATCTTTGTCATCTACC gtttcttTGATCAGTTCTTAACTCAACCA	AC053516 (nt2747…2794)	**(ATCT)12**	16.831
DYS534	NED-CATCTACCCAACATCCATCTA gtttcttGACAAAGATGTTAGATGAATAGACA	AC053516 (nt2497…2556)	**(CTTT)15**	16.831
DYS540	NED-GACCGTGTACTCTGGCCAAT gtttcttCAGGAGGCTAGCTCAGGAGA	AC010135 (nt147642…147689)	**(TTAT)12**	17.003
DYS556	FAM-TGCTGTCACATCACCAATGA gtttcttTTTGGTTGCTGAAGCATTGA	AC011745 (nt19776…19819)	**(AATA)11**	20.940
DYS557	VIC-TTTTCTGTGCCAAGCCTACA gtttcttTCTAATGCACCTTGAGGGATG	AC007876 (nt86525…86592)	**(TTTC)16**	21.573
DYS570	NED-GAACTGTCTACAATGGCTCACG gtttcttTCAGCATAGTCAAGAAACCAGACA	AC012068 (nt43502…43569)	**(TTTC)17**	6.905
DYS575	FAM-GGTGGTGGACATCCGTAATC gtttcttAGTAATGGGATGCTGGGTCA	AC007247 (nt91065…91104)	**(AAAT)10**	7.479
DYS576	FAM-TTGGGCTGAGGAGTTCAATC gtttcttGGCAGTCTCATTTCCTGGAG	AC010104 (nt69314…69381)	**(AAAG)17**	7.096
DYS594	VIC-GATGTGCCTAATGCCACAGA gtttcttCCCTGGTGTTAATCGTGTCC	AC010137 (nt50161…50210)	**(TAAAA)10**	20.045
DYS632	VIC-GGCCGTTGCAAAATAAACTG gtttcttTCTGGGCAACAGAAGGAGAC	AC006371 (nt153697…153732)	**(CATT)9**	14.347
DYS635 (GATA-C4)	Yfiler kit (Applied Biosystems) used	AC004772 (reverse & complement)	[TCTA]4(TGTA)2 [TCTA]2(TGTA)2 [TCTA]2(TGTA)2 [TCTA]9 = **23**	13.690
DYS641	NED-CTTGAGCCCAGGAAGCATAG gtttcttCCACACGATGCAATTTTGTC	AC018677 (nt3842…3881)	**(TAAA)10**	14.572
DYS643	FAM-AAGCCATGCCTGGTTAAACT gtttcttTGTAACCAAACACCACCCATT	AC007007 (nt25665…25719)	**(CTTTT)11**	15.864

Chromosome positions were determined using hgBLAT (http://genome.ucsc.edu/cgi-bin/hgBlat) and the July 2003 human genome reference sequence.

BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat) and are based on the July 2003 assembly of the human genome. Primers were crosschecked against one another using the AutoDimer program [4] to ensure that only compatible loci were combined with one another in testing.

PCR amplification: For each sample, the Y-STR markers were typed in small multiplexes containing 3, 4, or 5 loci. For example, the 20 loci examined in most detail were run in five sets: (a) DYS495, DYS508, DYS520, and DYS533; (b) DYS505, DYS540, DYS522, and DYS532; (c) DYS485, DYS556, DYS594, and DYS449; (d) DYS576, DYS463, DYS570, and DYS446; (e) DYS643, DYS557, DYS534, and DYS444. Amplifications were performed in reaction volumes of 10 μL using a master mix containing 1X GeneAmp PCR Gold buffer (Applied Biosystems, Foster City, CA), 1.5 mmol/L MgCl$_2$, 250 μmol/L deoxynucleotide triphosphates (dNTPs; Promega Corporation, Madison, WI), 0.16 mg/mL bovine serum albumin (BSA) fraction V (Sigma, St. Louis, MO), and 1.0 unit of AmpliTaq Gold DNA polymerase (Applied Biosystems). The thermal cycling program was carried out on a GeneAmp 9700 (Applied Biosystems) using the following conditions in 9600-emulation mode (i.e., ramp speeds of 1 °C/s): 95 °C for 10 min; 30 cycles of {95 °C for 1 min, 55 °C for 1 min, 72 °C for 1 min}; 60 °C for 45 min; and 25 °C until removed from thermal cycler.

Electrophoresis and typing: A 1.0 μL aliquot of each Y-STR PCR product was diluted in 14 μL Hi-Di^{TM} formamide and 0.4 μL GS500-LIZ internal size standard (Applied Biosystems) and analyzed on the 16-capillary ABI Prism 3100 Genetic Analyzer (Applied Biosystems) using filter set G5 without prior denaturation of samples. Samples were injected electrokinetically for 10 s at 3 kV. Separations were performed in approximately 60 min on a 36 cm array using POP^{TM}-6 (Applied Biosystems). Automated allele sizing was performed with Genotyper 3.7.

DNA sequencing: At least two different alleles were sequenced for each Y-STR locus with the exception of DYS575 and DYS641 to aid calibration of observed size to repeat number and to assist in nomenclature decisions. Unlabeled primers (see Table 1) were used for generating both forward and reverse sequencing reactions. Thermal cycling was performed using conditions described above but with 35 cycles instead of 30. Amplicon purity and concentration were assessed with the Agilent BioAnalyzer 2100 (Agilent Technologies, Palo Alto, CA). Following this assessment, unincorporated primers and dNTPs were removed by adding 2 μL of ExoSAP-IT (USB Corp., Cleveland, OH) to each 5 μL PCR reaction. Reactions were mixed briefly and incubated at 37 °C for 15 min and then 80 °C for 15 min to inactive the enzymes. Cycle sequencing was performed with the ABI PRISM BigDye Terminator (Version 3.1) cycle sequencing kit (Applied Biosystems) using 2 μL 5X Sequencing Buffer (Applied Biosystems), 4 μL BigDye Terminator kit mix (Applied Biosystems), 3.2 μL of 1 μmol/L forward or reverse primer, and Exo-SAP treated

PCR product plus sterile water to bring the total reaction volume to 20 μL. The target quantity for each PCR product in the sequencing reaction was approximately 7 ng based on the Agilent 2100 determined value. Thermal cycling was conducted in a GeneAmp 9700 thermocycler with the following conditions: an initial 1 min denaturation at 96 °C; followed by 25 cycles of 15 s at 94 °C (denaturation), 5 s at 50 °C (annealing), and 2 min at 60 °C (extension). The DNA product was purified by filtration though a single tube cartridge or 96-well plate DTR filtration system (Edge BioSystems, Gaithersburg, MD). Purified sequencing products were separated on an ABI 3100 Genetic Analyzer using POP-6 polymer. Sequences were aligned with the GenBank reference allele (accession numbers listed in Table 1) and edited using Sequencher Plus 4.0.5b11 (Gene-Codes, Ann Arbor, MI).

Y-STR allele nomenclature: The repeat structure used to define the nomenclature for each Y-STR marker is given in Table 1. The repeat motifs for each of the markers were determined using guidance provided by the DNA Commission of International Society for Forensic Genetics (ISFG) [5].

Analysis of data: Allele frequencies were estimated by direct counting. Gene diversities (D) were calculated according to $D = (n/n - 1)(1 - \Sigma p_i^2)$, where n is the sample size and p_i is the allelic frequency [6].

Access to the data: Genotyping results for these Y-STRs are posted at http://www.cstl.nist.gov/biotech/strbase/NISTpop.htm.

Results and discussion: This data set provides a continuation of information made available on these same samples by Schoske et al. [7], where results from 27 amplified regions of the Y-chromosome were reported. In addition, a concordance study with a beta-test version of Applied Biosystems' Yfiler kit found that there was complete agreement with the Schoske et al. [7] data with the exception of a nomenclature difference in DYS448. The nomenclature by Redd et al. [8] for DYS448 is now recommended and part of the certified values for the NIST Human Y-chromosome DNA Profiling Standard Reference Material SRM 2395 (see http://ts.nist.gov/ts/htdocs/230/232/232.htm, then enter 2395). The DYS635 information collected from the Yfiler testing of our U.S. population samples is presented in Table 2 along with the other new Y-STR loci examined as part of this study.

Most of the loci examined in the present study were selected from the most polymorphic simple repeats reported by Kayser et al. [9] in their comprehensive survey of the Y-chromosome. Since the Kayser et al. [9] survey only examined eight different samples albeit from diverse backgrounds, the information collected as part of this study represents a first attempt to aid assessment of these loci in populations of interest through collection of allele frequency information and thus help establish which loci are most polymorphic and desirable to pursue in future population studies and human identity applications. It should be noted that some allele frequency information has been

Table 2

The distribution of allele frequencies observed in U.S. Caucasian (Cau), African American (AA), and Hispanic (His) sample sets listed by Y-STR locus (see Table 1)

Locus	Allele	Combined frequency ($N = 654$)	Cau frequency ($N = 258$)	AA frequency ($N = 257$)	His frequency ($N = 139$)
DYS444	10	0.0107	0.0039	0.0117	0.0216
	11	0.1101	0.0659	0.1206	0.1727
	12	**0.5872**	**0.5930**	**0.6304**	**0.4964**
	13	0.2202	0.2597	0.1712	0.2374
	14	0.0596	0.0698	0.0545	0.0504
	15	0.0107	0.0078	0.0117	0.0144
	16	0.0015	0.0000	0.0000	0.0072

Locus	Allele	Combined frequency ($N = 660$)	Cau frequency ($N = 262$)	AA frequency ($N = 258$)	His frequency ($N = 140$)
DYS446	8	0.0015	0.0000	0.0000	0.0071
	10	0.0061	0.0115	0.0000	0.0071
	11	0.0318	0.0420	0.0310	0.0143
	12	0.1576	0.1450	0.1357	0.2214
	13	**0.5121**	**0.5763**	**0.4767**	**0.4571**
	14	0.1682	0.1450	0.1938	0.1643
	15	0.0545	0.0382	0.0736	0.0500
	15.1	0.0015	0.0000	0.0000	0.0071
	16	0.0288	0.0076	0.0465	0.0357
	17	0.0152	0.0076	0.0155	0.0286
	18	0.0076	0.0076	0.0116	0.0000
	19	0.0091	0.0115	0.0078	0.0071
	20	0.0030	0.0038	0.0039	0.0000
	21	0.0030	0.0038	0.0039	0.0000

Locus	Allele	Combined frequency ($N = 651$)	Cau frequency ($N = 261$)	AA frequency ($N = 253$)	His frequency ($N = 137$)
DYS449	24	0.0015	0.0038	0.0000	0.0000
	25	0.0061	0.0038	0.0000	0.0219
	26	0.0077	0.0115	0.0040	0.0073
	27	0.0445	0.0421	0.0553	0.0292
	28	0.1459	0.1724	0.1186	0.1460
	29	**0.2611**	**0.2989**	0.1818	**0.3358**
	30	0.2043	0.2529	0.1779	0.1606
	31	0.1536	0.1149	**0.2016**	0.1387
	32	0.1045	0.0651	0.1542	0.0876
	33	0.0369	0.0230	0.0435	0.0511
	34	0.0123	0.0038	0.0198	0.0146
	35	0.0108	0.0038	0.0237	0.0000
	36	0.0077	0.0038	0.0119	0.0073
	37	0.0031	0.0000	0.0079	0.0000

Locus	Allele	Combined frequency ($N = 661$)	Cau frequency ($N = 262$)	AA frequency ($N = 258$)	His frequency ($N = 140$)
DYS463	17	0.0015	0.0038	0.0000	0.0000
	18	0.0408	0.0304	0.0194	0.1000
	19	0.0106	0.0038	0.0233	0.0000
	20	0.2284	0.0076	**0.5388**	0.0714
	21	0.1014	0.1331	0.0969	0.0500
	22	0.1120	0.1483	0.0465	0.1643
	23	0.0530	0.1027	0.0116	0.0357
	24	**0.3873**	**0.4905**	0.2287	**0.4857**
	25	0.0560	0.0722	0.0310	0.0714
	26	0.0076	0.0076	0.0000	0.0214
	28	0.0015	0.0000	0.0039	0.0000

Locus	Allele	Combined frequency ($N = 655$)	Cau frequency ($N = 262$)	AA frequency ($N = 254$)	His frequency ($N = 139$)
DYS485	10	0.0061	0.0000	0.0157	0.0000
	12	0.0702	0.1298	0.0276	0.0360

Table 2 (*Continued*)

Locus	Allele	Combined frequency ($N = 655$)	Cau frequency ($N = 262$)	AA frequency ($N = 254$)	His frequency ($N = 139$)
	13	0.0244	0.0344	0.0118	0.0288
	14	0.2824	0.0992	**0.5197**	0.1942
	15	**0.5084**	**0.6336**	0.2992	**0.6547**
	16	0.0809	0.0649	0.1102	0.0576
	17	0.0229	0.0267	0.0157	0.0288
	18	0.0046	0.0115	0.0000	0.0000

Locus	Allele	Combined frequency ($N = 92$)	Cau frequency ($N = 29$)	AA frequency ($N = 31$)	His frequency ($N = 32$)
DYS490	8,13	0.0435	0.0000	0.1290	0.0000
	9,12	0.0978	0.0690	0.1290	0.0938
	9,13	0.0326	0.0000	0.0645	0.0313
	9,14	0.0109	0.0000	0.0000	0.0313
	9,15	0.0109	0.0000	0.0323	0.0000
	9,16	0.0109	0.0000	0.0323	0.0000
	10,13	0.0109	0.0000	0.0323	0.0000
	12,13	0.0109	0.0000	0.0000	0.0313
	12,14	**0.3804**	**0.5172**	0.1290	**0.5000**
	12,15	0.0217	0.0345	0.0000	0.0313
	12,16	0.1304	0.1379	0.0323	0.2188
	12,17	0.0652	0.1724	0.0323	0.0000
	12,18	0.0217	0.0690	0.0000	0.0000
	13,13	0.0217	0.0000	0.0645	0.0000
	13,14	0.0870	0.0000	**0.2258**	0.0313
	13,15	0.0109	0.0000	0.0323	0.0000
	13,16	0.0217	0.0000	0.0323	0.0313
	14,15	0.0109	0.0000	0.0323	0.0000

Locus	Allele	Combined frequency ($N = 659$)	Cau frequency ($N = 262$)	AA frequency ($N = 257$)	His frequency ($N = 140$)
DYS495	12	0.0076	0.0000	0.0039	0.0286
	13	0.0076	0.0000	0.0039	0.0286
	14	0.0455	0.0725	0.0156	0.0500
	15	0.3748	0.1870	**0.6070**	0.3000
	16	**0.4446**	**0.5725**	0.2802	**0.5071**
	17	0.1153	0.1641	0.0856	0.0786
	18	0.0046	0.0038	0.0039	0.0071

Locus	Allele	Combined frequency ($N = 94$)	Cau frequency ($N = 30$)	AA frequency ($N = 32$)	His frequency ($N = 32$)
DYS504	11	0.0213	0.0000	0.0313	0.0313
	12	0.0213	0.0000	0.0625	0.0000
	13	0.2128	0.0000	**0.4688**	0.1563
	14	0.1489	0.1667	0.1250	0.1563
	15	0.0638	0.0667	0.0313	0.0938
	16	0.1702	0.2333	0.0625	0.2188
	17	**0.3085**	**0.4667**	0.2188	**0.2500**
	18	0.0426	0.0667	0.0000	0.0625
	19	0.0106	0.0000	0.0000	0.0313

Locus	Allele	Combined frequency ($N = 654$)	Cau frequency ($N = 262$)	AA frequency ($N = 253$)	His frequency ($N = 139$)
DYS505	9	0.0076	0.0115	0.0000	0.0144
	10	0.0031	0.0038	0.0040	0.0000
	11	0.2232	0.2366	0.1818	0.2734
	12	**0.4541**	**0.5000**	**0.3874**	**0.4892**
	13	0.2783	0.2290	0.3676	0.2086
	14	0.0275	0.0191	0.0435	0.0144
	15	0.0061	0.0000	0.0158	0.0000

Locus	Allele	Combined frequency ($N = 658$)	Cau frequency ($N = 261$)	AA frequency ($N = 257$)	His frequency ($N = 140$)
DYS508	8	0.0015	0.0038	0.0000	0.0000
	9	0.0015	0.0038	0.0000	0.0000

Table 2 (*Continued*)

Locus	Allele	Combined frequency (N = 658)	Cau frequency (N = 261)	AA frequency (N = 257)	His frequency (N = 140)
	10	0.1292	0.2146	0.0506	0.1143
	10.1	0.0015	0.0038	0.0000	0.0000
	11	**0.4833**	**0.6207**	0.3074	**0.5500**
	12	0.2325	0.1264	**0.3385**	0.2357
	13	0.0638	0.0192	0.1206	0.0429
	14	0.0608	0.0077	0.1206	0.0500
	15	0.0258	0.0000	0.0623	0.0071

Locus	Allele	Combined frequency (N = 659)	Cau frequency (N = 262)	AA frequency (N = 257)	His frequency (N = 140)
DYS520	18	0.0061	0.0115	0.0039	0.0000
	19	0.0395	0.0382	0.0156	0.0857
	20	0.0425	0.0611	0.0156	0.0571
	21	**0.4856**	**0.6450**	0.2918	**0.5429**
	22	0.1608	0.1718	0.1479	0.1643
	23	0.2231	0.0687	**0.4280**	0.1357
	24	0.0319	0.0038	0.0700	0.0143
	25	0.0091	0.0000	0.0233	0.0000
	26	0.0015	0.0000	0.0039	0.0000

Locus	Allele	Combined frequency (N = 655)	Cau frequency (N = 259)	AA frequency (N = 258)	His frequency (N = 138)
DYS522	8	0.0015	0.0039	0.0000	0.0000
	9	0.0031	0.0039	0.0000	0.0072
	10	0.3099	0.3707	0.2093	**0.3841**
	11	**0.4519**	**0.4208**	0.5233	0.3768
	12	0.2015	0.1815	0.2248	0.1957
	13	0.0290	0.0154	0.0426	0.0290
	14	0.0015	0.0000	0.0000	0.0072
	17	0.0015	0.0039	0.0000	0.0000

Locus	Allele	Combined frequency (N = 93)	Cau frequency (N = 29)	AA frequency (N = 32)	His frequency (N = 32)
DYS525	7	0.0108	0.0000	0.0313	0.0000
	8	0.0108	0.0000	0.0313	0.0000
	9	0.0538	0.0000	0.0000	0.1563
	10	**0.6774**	**0.7586**	**0.7813**	**0.5000**
	11	0.1505	0.2069	0.0625	0.1875
	12	0.0645	0.0000	0.0938	0.0938
	13	0.0323	0.0345	0.0000	0.0625

Locus	Allele	Combined frequency (N = 655)	Cau frequency (N = 262)	AA frequency (N = 254)	His frequency (N = 139)
DYS532	9	0.0153	0.0267	0.0039	0.0144
	10	0.0443	0.0305	0.0433	0.0719
	11	0.1252	0.1641	0.0945	0.1079
	12	0.2290	0.1489	0.3150	0.2230
	13	**0.3588**	**0.3893**	**0.3425**	**0.3309**
	14	0.1450	0.1603	0.1181	0.1655
	15	0.0626	0.0649	0.0591	0.0647
	16	0.0168	0.0153	0.0197	0.0144
	17	0.0031	0.0000	0.0039	0.0072

Locus	Allele	Combined frequency (N = 660)	Cau frequency (N = 262)	AA frequency (N = 258)	His frequency (N = 140)
DYS533	9	0.0258	0.0305	0.0078	0.0500
	10	0.0333	0.0115	0.0543	0.0357
	11	0.3424	0.2023	**0.5155**	0.2857
	12	**0.4848**	**0.5954**	0.3295	**0.5643**
	13	0.0864	0.1260	0.0698	0.0429
	14	0.0258	0.0305	0.0233	0.0214
	12,13	0.0015	0.0038	0.0000	0.0000

Table 2 (*Continued*)

Locus	Allele	Combined frequency (N = 657)	Cau frequency (N = 261)	AA frequency (N = 256)	His frequency (N = 140)
DYS534	10	0.0015	0.0000	0.0039	0.0000
	11	0.0030	0.0038	0.0039	0.0000
	12	0.0091	0.0038	0.0195	0.0000
	13	0.0487	0.0421	0.0430	0.0714
	14	0.1339	0.1533	0.1445	0.0786
	15	**0.3805**	**0.3525**	**0.4141**	**0.3714**
	15.1	0.0030	0.0077	0.0000	0.0000
	16	0.2542	0.2644	0.2383	0.2643
	17	0.1187	0.1111	0.1016	0.1643
	18	0.0381	0.0498	0.0273	0.0357
	19	0.0061	0.0038	0.0039	0.0143
	20	0.0015	0.0038	0.0000	0.0000
	15,17	0.0015	0.0038	0.0000	0.0000

Locus	Allele	Combined frequency (N = 659)	Cau frequency (N = 263)	AA frequency (N = 257)	His frequency (N = 139)
DYS540	10	0.0137	0.0076	0.0195	0.0144
	11	0.2200	0.1787	0.2062	0.3237
	12	**0.7132**	**0.7567**	**0.7160**	**0.6259**
	13	0.0501	0.0532	0.0545	0.0360
	14	0.0030	0.0038	0.0039	0.0000

Locus	Allele	Combined frequency (N = 655)	Cau frequency (N = 262)	AA frequency (N = 254)	His frequency (N = 139)
DYS556	9	0.0031	0.0038	0.0039	0.0000
	10	0.0076	0.0153	0.0039	0.0000
	10.1	0.0015	0.0038	0.0000	0.0000
	11	**0.5863**	**0.7214**	0.3937	**0.6835**
	12	0.3603	0.2328	**0.5276**	0.2950
	13	0.0412	0.0229	0.0709	0.0216

Locus	Allele	Combined frequency (N = 658)	Cau frequency (N = 262)	AA frequency (N = 257)	His frequency (N = 139)
DYS557	11	0.0015	0.0038	0.0000	0.0000
	12	0.0015	0.0000	0.0000	0.0071
	13	0.0030	0.0000	0.0038	0.0071
	14	0.0441	0.0830	0.0154	0.0214
	15	0.1520	0.2302	0.0654	0.1571
	16	**0.3754**	**0.4566**	0.2462	**0.4429**
	17	0.1717	0.1396	0.1923	0.1857
	18	0.1672	0.0604	**0.2923**	0.1286
	19	0.0593	0.0113	0.1231	0.0286
	20	0.0167	0.0038	0.0308	0.0143
	21	0.0061	0.0000	0.0154	0.0000
	22	0.0015	0.0000	0.0038	0.0000

Locus	Allele	Combined frequency (N = 661)	Cau frequency (N = 262)	AA frequency (N = 259)	His frequency (N = 140)
DYS570	12	0.0015	0.0000	0.0039	0.0000
	13	0.0015	0.0000	0.0039	0.0000
	14	0.0076	0.0000	0.0193	0.0000
	15	0.0197	0.0153	0.0309	0.0071
	16	0.0893	0.0878	0.0618	0.1429
	17	**0.3404**	**0.4084**	**0.2934**	**0.3000**
	18	0.2390	0.2405	0.2548	0.2071
	19	0.1679	0.1221	0.1969	0.2000
	20	0.0832	0.0878	0.1004	0.0429
	21	0.0227	0.0229	0.0193	0.0286
	22	0.0212	0.0115	0.0116	0.0571
	23	0.0061	0.0038	0.0039	0.0143

Locus	Allele	Combined frequency (N = 94)	Cau frequency (N = 30)	AA frequency (N = 32)	His frequency (N = 32)
DYS575	9	0.0106	0.0000	0.0313	0.0000
	10	**0.9894**	**1.0000**	**0.9688**	**1.0000**

Table 2 (*Continued*)

Locus	Allele	Combined frequency (N = 659)	Cau frequency (N = 261)	AA frequency (N = 258)	His frequency (N = 140)
DYS576	13	0.0030	0.0038	0.0039	0.0000
	14	0.0273	0.0115	0.0426	0.0286
	15	0.1047	0.0421	0.1977	0.0500
	16	0.1745	0.1762	0.2016	0.1214
	17	0.2640	0.2835	0.2209	**0.3071**
	18	**0.2822**	**0.3257**	**0.2403**	0.2786
	19	0.1077	0.1188	0.0698	0.1571
	20	0.0288	0.0307	0.0194	0.0429
	21	0.0076	0.0077	0.0039	0.0143

Locus	Allele	Combined frequency (N = 657)	Cau frequency (N = 262)	AA frequency (N = 256)	His frequency (N = 139)
DYS594	9	0.0137	0.0153	0.0156	0.0072
	10	**0.5693**	**0.8015**	0.3047	**0.6187**
	11	0.3714	0.1756	**0.5977**	0.3237
	12	0.0335	0.0076	0.0586	0.0360
	13	0.0091	0.0000	0.0195	0.0072
	14	0.0030	0.0000	0.0039	0.0072

Locus	Allele	Combined frequency (N = 94)	Cau frequency (N = 30)	AA frequency (N = 32)	His frequency (N = 32)
DYS632	9	**0.5426**	0.2333	**0.8438**	**0.5313**
	10	0.4574	**0.7667**	0.1563	0.4688

Locus	Allele	Combined frequency (N = 661)	Cau frequency (N = 263)	AA frequency (N = 259)	His frequency (N = 139)
DYS635	17	0.0121	0.0000	0.0270	0.0072
GATA-C4	19	0.0061	0.0000	0.0154	0.0000
	20	0.0575	0.0494	0.0463	0.0935
	21	0.2753	0.1559	**0.4324**	0.2086
	21.3	0.0015	0.0000	0.0000	0.0072
	22	0.1180	0.0989	0.1274	0.1367
	23	**0.3964**	**0.5095**	0.2471	**0.4604**
	24	0.1089	0.1407	0.0965	0.0719
	25	0.0197	0.0380	0.0077	0.0072
	26	0.0030	0.0076	0.0000	0.0000
	27	0.0015	0.0000	0.0000	0.0072

Locus	Allele	Combined frequency (N = 94)	Cau frequency (N = 30)	AA frequency (N = 32)	His frequency (N = 32)
DYS641	9	0.0106	0.0000	0.0313	0.0000
	10	**0.9362**	**0.9333**	**0.8750**	**1.0000**
	11	0.0532	0.0667	0.0938	0.0000

Locus	Allele	Combined frequency (N = 657)	Cau frequency (N = 263)	AA frequency (N = 257)	His frequency (N = 137)
DYS643	7	0.0046	0.0000	0.0078	0.0073
	8	0.0046	0.0076	0.0039	0.0000
	9	0.0594	0.0798	0.0233	0.0876
	10	**0.4125**	**0.5741**	0.2062	**0.4891**
	11	0.1065	0.1217	0.0817	0.1241
	11.1	0.0015	0.0038	0.0000	0.0000
	12	0.1613	0.1483	0.1518	0.2044
	13	0.1766	0.0608	**0.3580**	0.0584
	14	0.0563	0.0038	0.1245	0.0292
	15	0.0167	0.0000	0.0428	0.0000

The most frequent allele is bolded for each group.

previously described with DYS444 [10], DYS446 [8], DYS449 [8], and DYS463 [8]. In addition, Dai et al. [11] recently published DYS520 allele frequencies from a Chinese population. The allele frequencies for DYS446, DYS449, and DYS463 observed for the Caucasian samples in this study were quite similar to those reported for 148 Caucasians in Redd et al. [8] (data not shown).

Table 2 contains the observed allele frequencies for U.S. Caucasian, African American, and Hispanic samples. Note that 21 of the loci were tested across more than 650 total
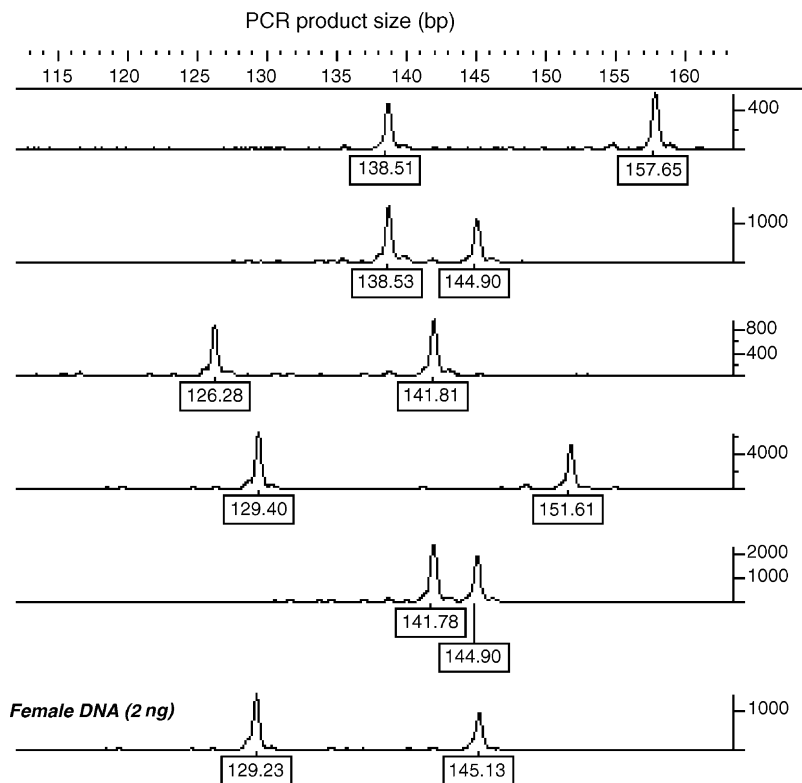
Fig. 1. Various allele patterns observed with the duplicated trinucleotide repeat locus DYS490 in multiple males and a single female sample. All DNA templates were approximately 1–2 ng in amount. Peak labels are PCR product sizes in base pairs (bp) relative to the GS500 LIZ size standard on an ABI 3100.

samples while DYS490, DYS504, DYS525, DYS575, DYS632, and DYS641 were only examined in 92–94 samples due to low variation, poor performance with the current primers, or potential X-chromosome homology. For example, DYS490 produced double peak patterns in most male samples examined as well as a female control (Fig. 1). Due to its female cross-reactivity, forensic laboratories desiring male-specific amplification will not likely adopt DYS490, even though this X-chromosome homology may be reduced or eliminated through careful primer design in the future. None of the other 26 loci examined in this study produced PCR products with the 2 ng female DNA template tested.

The rest of the loci examined in this study were all single-copy markers and thus exhibited only a single peak at each locus. However, two PCR products were observed in a single Caucasian sample at DYS533 and DYS534. Table 2 lists these allele pairs in a separate category rather than incorporating them into the overall allele count for the two individual alleles. Since the repeat structures for these two loci are only 190 bp apart it is likely that this entire region near 16.831 megabases on the Y-chromosome has been duplicated within this sample.

Nomenclature of allele designation is important to enable effective communication of information across the DNA typing community. Previous nomenclatures, which were published prior to this study, were followed in Table 1 for DYS444 [10], DYS446 [8], DYS449 [8], DYS463 [8], and DYS635 [12]. The nomenclature for the simple repeat loci, such as DYS485, DYS533, or DYS643, is straightforward and follows ISFG recommendations with the earliest possible repeat motif along the GenBank top strand being designated [5]. An examination of the supplementary material attached to the Kayser et al. [9] publication describing these loci indicates that we are in agreement for most of the loci.

There are, however, four loci with possible alternative nomenclatures to the ones presented in Table 1. These include DYS520, DYS532, DYS534, and DYS557. DYS520 contains a compound repeat that was designed $(GATA)_m (CATA)_n$ by Dai et al. [11]. We have chosen to call the repeat structure $(ATAG)_m (ATAC)_n$ in order to reflect the earliest possible motif and thus follow the ISFG guidelines. In this case, there should not be a difference in the total number of repeats using either approach. With DYS532 there are five adjacent repeats possessing similar structure to the core CTTT repeat. If the full motif pattern of $(CTTT)_3(CCTT)\mathbf{(CTTT)_{14}}(CTTC)$ was used, then the Gen-Bank reference allele would be designated 19 repeats rather than the 14 repeats based solely on the core CTTT repeat listed in Table 1. DYS534 contains a compound repeat

structure around the core CTTT repeat: $(CTTT)_3$ $(CCTT)(CCTC)\mathbf{(CTTT)_{15}}(CTGT)(CTTT)$. We have chosen to only designate the core 15 CTTT repeats rather than 22 $(3 + 1 + 1 + 15 + 1 + 1)$ compound repeats. Likewise DYS557 possesses extra TTTC repeats in close proximity to the core TTTC repeat that might be included in an alternative approach to designating the repeat structure. The bottom strand of GenBank accession AC004772 is used to denote DYS635 in order to maintain consistency with previous work, referred to then as Y-GATA-C4 [12].

The variant alleles observed in this study have been confirmed through mixing experiments with adjacent alleles for each locus. The following variant alleles have been sequenced: DYS556 allele 10.1 ($[AATA]_2$ AAT$\underline{T}$A $[AATA]_7$), DYS635 allele 21.3 ($[TCTA]_4(TGTA)_2[TC$-TA$]_2(TGTA)_2[TCTA]_2(TGTA)_2$ $[TCTA]_5$ TC-A $[TCTA]_2$), and DYS643 allele 11.1 ($[CTTTT]_8\underline{C}$ $[CTTTT]_3$).

The data presented in this work are preliminary. However, it appears that several simple repeat loci including DYS570 and DYS576 are highly polymorphic and thus may aid resolution of common Y-STR haplotypes, which cannot be separated with the widely used minimal haplotype loci. A future manuscript will address optimal loci in terms of haplotype resolution as well as construction of new large multiplexes to enable simultaneous amplification of the better performing loci.

Table 3 compares the gene diversity values for the individual loci in all the samples examined as well as the various U.S. population groups. Haplotype information across our U.S. samples for these 27 Y-STRs has not been included in the print version of this announcement due to space considerations but is available in electronic format at http://www.cstl.nist.gov/biotech/strbase/NISTpop.htm (data from Schoske et al. [7] on many of the same samples with 22 additional loci is also available on our website). Future reports are planned based on this data that will investigate the value of additional loci to those currently available in commercial STR kits.

This paper follows the guidelines for publication of population data requested by the journal [13].

Table 3
Comparison of gene diversity values for the 27 Y-STR loci across the U.S. samples examined in this study

Locus	N	Diversity	Rank	Caucasian	African American	Hispanic
DYS449	651	0.8318	1	0.8001	0.8541	0.8149
DYS490	92	0.8201	2	0.6970	0.9118	0.7097
DYS504	94	0.8101	3	0.7149	0.7298	0.8528
DYS576	661	0.7972	4	0.7684	0.8098	0.7889
DYS570	661	0.7836	5	0.7470	0.7975	0.8063
DYS532	656	0.7769	6	0.7705	0.7578	0.7976
DYS557	658	0.7739	7	0.7110	0.7992	0.7318
DYS463	661	0.7684	8	0.7056	0.6464	0.7178
DYS534	657	0.7557	9	0.7686	0.7404	0.7579
DYS643	657	0.7554	10	0.6258	0.7847	0.6968
DYS635	661	0.7385	11	0.6852	0.7260	0.7168
DYS508	658	0.6883	12	0.5544	0.7583	0.6290
DYS520	659	0.6851	13	0.5465	0.7066	0.6538
DYS446	660	0.6804	14	0.6245	0.7105	0.7151
DYS505	654	0.6666	15	0.6435	0.6823	0.6466
DYS522	656	0.6593	16	0.6548	0.6325	0.6762
DYS485	655	0.6501	17	0.5678	0.6293	0.5312
DYS495	659	0.6473	18	0.6074	0.5476	0.6471
DYS533	660	0.6387	19	0.5889	0.6197	0.5982
DYS444	654	0.5918	20	0.5738	0.5577	0.6690
DYS594	657	0.5374	21	0.3277	0.5480	0.5146
DYS556	655	0.5255	22	0.4263	0.5638	0.4486
DYS525	93	0.5157	23	0.3941	0.3871	0.6996
DYS632	94	0.5017	24	0.3701	0.2722	0.5141
DYS540	659	0.4409	25	0.3941	0.4432	0.5056
DYS641	94	0.1219	26	0.1287	0.2319	0.0000
DYS575	94	0.0213	27	0.0000	0.0625	0.0000

The loci are listed according to their diversity values with the combined data set. Note that the total number of samples run for each locus varies, as does the ranking in each population group.

References

[1] J.M. Butler, R. Schoske, P.M. Vallone, J.W. Redman, M.C. Kline, Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations, J. Forensic Sci. 48 (2003) 908–911.

[2] S.A. Miller, D.D. Dykes, H.F. Polesky, A simple salting out procedure for extracting DNA from human nucleated cells, Nucleic Acids Res. 16 (1988) 1215.

[3] V.L. Singer, L.J. Jones, S.T. Yue, R.P. Haugland, Characterization of PicoGreen reagent and development of a fluorescence-based solution assay for double-stranded DNA quantitation, Anal. Biochem. 249 (1997) 228–238.

[4] P.M. Vallone, J.M. Butler, AutoDimer: a screening tool for primer–dimer and hairpin structures, Biotechniques 37 (2004) 226–231.

[5] P. Gill, C. Brenner, B. Brinkmann, B. Budowle, A. Carracedo, M.A. Jobling, P. de Knijff, M. Kayser, M. Krawczak, W.R. Mayr, N. Morling, B. Olaisen, V. Pascali, M. Prinz, L. Roewer,

P.M. Schneider, A. Sajantila, C. Tyler-Smith, DNA commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs, Int. J. Leg. Med. 114 (2001) 305–309.

[6] M. Nei, Molecular Evolutionary Genetics, Columbia University Press, New York, 1987, pp. 176–179.

[7] R. Schoske, P.M. Vallone, M.C. Kline, J.W. Redman, J.M. Butler, High-throughput Y-STR typing of U.S. populations with 27 regions of the Y chromosome using two multiplex PCR assays, Forensic Sci. Int. 139 (2004) 107–121.

[8] A.J. Redd, A.B. Agellon, V.A. Kearney, V.A. Contreras, T. Karafet, H. Park, P. de Knijff, J.M. Butler, M.F. Hammer, Forensic value of 14 novel STRs on the human Y chromosome, Forensic Sci. Int. 130 (2002) 97–111.

[9] M. Kayser, R. Kittler, A. Ralf, M. Hedman, A.C. Lee, A. Mohyuddin, S.Q. Mehdi, Z. Rosser, M. Stoneking, M.A. Jobling, A. Sajantila, C. Tyler-Smith, A comprehensive survey of human Y-chromosomal microsatellites, Am. J. Hum. Genet. 74 (2004) 1183–1197.

[10] R. Iida, E. Tsubota, K. Sawazaki, M. Masuyama, T. Matsuki, T. Yasuda, K. Kishi, Characterization and haplotype analysis of the polymorphic Y-STRs DYS443, DYS444, and DYS445 in a Japanese population, Int. J. Leg. Med. 116 (2002) 191–194.

[11] H.L. Dai, X.D. Wang, Y.B. Li, J. Wu, J. Zhang, H.J. Zhang, J.G. Dong, Y.P. Hou, Characterization and haplotype analysis of 10 novel Y-STR loci in Chinese Han population, Forensic Sci. Int. 145 (2004) 47–55.

[12] L. Gusmao, A. Gonzalez-Neira, C. Alves, M. Lareu, S. Costa, A. Amorim, A. Carracedo, Chimpanzee homologous of human Y specific STRs: a comparative study and a proposal for nomenclature, Forensic Sci. Int. 126 (2002) 129–136.

[13] P. Lincoln, A. Carracedo, Publication of population data of human polymorphisms, Forensic Sci. Int. 110 (2000) 3–5.