

FOR THE RECORD

David D. Einum,¹ Ph.D. and Marco A. Scarpetta,¹ Ph.D.

Genetic Analysis of Large Data Sets of North American Black, Caucasian, and Hispanic Populations at 13 CODIS STR Loci

POPULATIONS: United States, Black, Caucasian, Hispanic

KEYWORDS: forensic science, United States, Black, Caucasian, Hispanic, population genetics, allele frequency, DNA typing, D3S1358, vWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317, D7S820, D16S539, TH01, TPOX and CSF1PO

Over 300,000 blood or buccal specimens were obtained from North American individuals in order to generate genetic data to help determine biological relationships. Genomic DNA was isolated from each specimen using phenol-chloroform extraction or other standard methodologies (1) and quantitated using PICO green (Molecular Probes). Approximately 1 ng of each DNA sample was subjected to PCR amplification with both the Applied Biosystems (ABI) AmpFISTR Profiler Plus and Cofiler systems. These test batteries provide genotype information for the 13 distinct Combined DNA Index System (CODIS) short tandem repeat (STR) loci (D3S1358, vWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317, D7S820, D16S539, TH01, TPOX and CSF1PO) used widely in forensic and paternity applications. The fluorescently-labeled amplification products were analyzed by denaturing polyacrylamide gel electrophoresis on an ABI Prism 377 DNA sequencer. The gels were analyzed by GeneScanner software and allele calls were made using the Genotyper program (both from ABI).

The resultant genetic data were electronically archived and three separate database queries were performed: one for 8000 genotypes (16,000 alleles/locus) of Blacks, another for 8000 genotypes of Caucasian individuals and the last for 1000 genotypes of Hispanics (2000 alleles/locus). Relatively fewer genotypes of Hispanics were included in this analysis due in part to the under-representation of Hispanic individuals submitting to genetic testing for cases of disputed paternity compared to Caucasians and Blacks. Furthermore, only Hispanics genotyped prior to calendar year 2001 were included in the database output because many Hispanic individuals typed after that date originated from a population potentially com-

prised of significant substructure. Additionally, a search protocol was designed for each of the three databases such that only those individuals entered as either a mother or an alleged father involved in a disputed paternity case would be included in the output file because such individuals are not likely biologically related. Finally, it was stipulated that a given person would only be included once in each output file regardless of whether multiple specimens from that individual were tested and the data uploaded, thus representing more than a single iteration of that particular genotype within the database.

Output data from each population were analyzed by Genetic Data Analysis (GDA) software (2). Allele frequencies were calculated at each locus by counting the numbers of each genotype present in the sample set. Unbiased estimates of expected heterozygosity frequencies were also calculated along with observed departures from Hardy-Weinberg equilibrium expectations as computed by the Fisher exact test based on 2000 shufflings (3). Probability of exclusion and power of discrimination statistics are also provided for each STR locus within each population.

The observed allele frequencies for each locus in each population are shown in Tables 1–3. FGA and D18S51 are the most informative systems in each of the 3 populations. The least discriminating loci are D13S317 and TH01 in North American Blacks and D5S818 and TPOX in Caucasians and Hispanics. TPOX ($p = 0.042$) in Caucasians and vWA ($p = 0.019$) in Blacks are the only loci to depart from Hardy-Weinberg expectations based on the exact test but do not comprise significant departures after correction for sampling (4). The combined powers of discrimination of the 13 CODIS loci exceed 0.99999999999999 in each of the three populations.

The complete data set can be accessed by any interested party at <http://www/technology/publications.asp>

¹ Orchid BioSciences, Inc., 2947 Eyde Parkway, Suite 110, East Lansing, MI 48823.

TABLE 1—Allele frequencies at the 13 CODIS STR loci for the North American Black population.

Allele	D3S1358 <i>N</i> * = 7,602	vWA 7,854	FGA 7,419	D8S1179 7,796	D21S11 7,652	D18S51 7,463	D5S818 7,845	D13S317 7,833	D7S820 7,612	D16S539 4,548	TH01 4,530	TPOX 4,532	CSF1PO 4,549
3	0.0001
5	0.0007	0.0034
6	0.0011	...	0.1283	0.0696	0.0007
7	0.0033	0.0001	0.0077	0.0002	0.4046	0.0225	0.0610
8	0.0024	...	0.0001	0.0555	0.0260	0.2040	0.0284	0.2214	0.3207	0.0650
9	0.0049	...	0.0013	0.0190	0.0218	0.1140	0.2097	0.1428	0.2308	0.0370
9.3	0.0894
10	0.0232	...	0.0019	0.0598	0.0273	0.3390	0.1323	0.0092	0.0966	0.2716
10.2	0.0016
11	0.0003	0.0074	...	0.0445	...	0.0050	0.2370	0.2940	0.2040	0.2902	0.0002	0.2351	0.2337
12	0.0045	0.0012	...	0.1190	...	0.0625	0.3530	0.4290	0.1080	0.1918	...	0.0235	0.2743
13	0.0077	0.0169	...	0.2030	...	0.0476	0.2510	0.1520	0.0202	0.1271	...	0.0004	0.0481
13.2	0.0050
14	0.0905	0.0632	...	0.3350	...	0.0658	0.0175	0.0486	0.0021	0.0184	0.0073
14.2	0.0037
15	0.2920	0.2130	...	0.1940	...	0.1660	0.0029	0.0010	...	0.0007	0.0007
15.2	0.0010	0.0001
16	0.3300	0.2600	...	0.0606	...	0.1810	0.0010	0.0002
17	0.2070	0.2030	0.0015	0.0119	...	0.1640	0.0002
17.2	0.0008
18	0.0630	0.1390	0.0075	0.0012	...	0.1160
18.2	0.0124
19	0.0048	0.0697	0.0665	0.0001	...	0.0913
19.2	0.0028	0.0004
20	...	0.0223	0.0620	0.0537
20.2	0.0014
21	...	0.0043	0.1120	0.0222
21.2	0.0015
22	...	0.0007	0.1800	0.0086
22.2	0.0020
23	0.1780	0.0021
23.2	0.0010
24	0.1680	0.0003
24.2	0.0001
25	0.0989	...	0.0003	0.0002
25.2	0.0002	...	0.0001
26	0.0506	...	0.0012
26.2	0.0001
27	0.0319	...	0.0560
28	0.0125	...	0.2430
28.2	0.0001
29	0.0050	...	0.1890
29.2	0.0001	...	0.0006
30	0.0018	...	0.1870
30.2	0.0014	...	0.0192
31	0.0047	...	0.0809
31.2	0.0520
32	0.0172
32.2	0.0694
33	0.0046
33.2	0.0308
34	0.0059
34.2	0.0031
35	0.0314
35.2	0.0003
36	0.0061
36.2
37	0.0014
38	0.0005
<i>p</i> [†]	0.188	0.019	0.148	0.512	0.871	0.451	0.180	0.508	0.665	0.981	0.137	0.544	0.588
He (%)	75.1	81.7	87.3	78.8	85.0	87.7	74.9	70.2	77.7	80.0	74.2	77.3	78.5
Ho (%)	74.6	81.9	87.3	78.8	84.6	88.1	74.4	70.3	77.4	80.5	73.7	77.0	78.9
PD	0.932	0.942	0.971	0.926	0.943	0.973	0.897	0.864	0.917	0.931	0.897	0.914	0.921
PE	0.503	0.635	0.741	0.577	0.687	0.757	0.499	0.433	0.552	0.608	0.488	0.545	0.579

* Number of individuals included in the analysis that fit the qualitative specifications described in the text; † Fisher exact test based on 2,000 shufflings; He: expected heterozygosity; Ho: observed heterozygosity; PD: Power of Discrimination; PE: Probability of Exclusion; ... allele not detected or not applicable.

Interclass correlations yielding $p < 0.05$ for pairwise comparisons: D8S1179/D13S317, vWA/TH01, D7S820/D5S818 and FGA/TH01.

TABLE 2—Allele frequencies at the 13 CODIS STR loci for the North American Caucasian population.

Allele	D3S1358 N* = 7,636	vWA 7,837	FGA 7,674	D8S1179 7,731	D21S11 7,730	D18S51 7,628	D5S818 7,829	D13S317 7,814	D7S820 7,685	D16S539 3,723	TH01 3,699	TPOX 3,701	CSF1PO 3,722
3
5	0.0020
6	0.0001	...	0.0001	...	0.2310	0.0011	0.0003
7	0.0020	0.0003	0.0208	0.0001	0.1879	0.0012	0.0008
8	0.0169	0.0033	0.1200	0.1560	0.0157	0.1077	0.5376	0.0039
9	0.0127	...	0.0008	0.0355	0.0754	0.1650	0.1125	0.1480	0.1038	0.0218
9.3	0.3150
10	0.0915	...	0.0091	0.0584	0.0618	0.2680	0.0608	0.0072	0.0561	0.2636
10.2
11	0.0009	0.0001	...	0.0724	...	0.0112	0.3690	0.3110	0.2040	0.2988	0.0001	0.2588	0.2937
12	0.0007	0.0004	...	0.1470	...	0.1450	0.3620	0.2830	0.1450	0.3002	...	0.0403	0.3291
13	0.0031	0.0010	...	0.3260	...	0.1260	0.1590	0.1040	0.0333	0.1825	...	0.0001	0.0689
13.2	0.0001
14	0.1240	0.0965	...	0.1980	...	0.1640	0.0093	0.0443	0.0071	0.0265	0.0149
14.2
15	0.2690	0.0990	...	0.1060	...	0.1440	0.0010	0.0014	0.0002	0.0017	0.0020
15.2
16	0.2430	0.2180	...	0.0267	...	0.1280	0.0001
17	0.2000	0.2700	0.0008	0.0036	...	0.1160
17.2
18	0.1460	0.2140	0.0162	0.0002	...	0.0773
18.2
19	0.0125	0.0847	0.0636	0.0440
19.2	0.0001
20	0.0001	0.0155	0.1470	0.0173
20.2	0.0015
21	...	0.0011	0.1770	0.0098
21.2	0.0032
22	...	0.0001	0.1770	0.0046
22.2	0.0100
23	0.1410	0.0014
23.2	0.0044
24	0.1340	0.0005
24.2	0.0011	...	0.0008
25	0.0850	...	0.0001
25.2	0.0001	...	0.0008
26	0.0300	...	0.0021	0.0001
26.2	0.0001	...	0.0002
27	0.0059	...	0.0332
28	0.0016	...	0.1580
28.2	0.0002
29	0.0003	...	0.2120
29.2	0.0010
30	0.0001	...	0.2580
30.2	0.0353
31	0.0001	...	0.0737
31.2	0.0896
32	0.0142
32.2	0.0854
33	0.0021
33.2	0.0290
34	0.0001
34.2	0.0035
35	0.0002
35.2	0.0004
36	0.0001
36.2	0.0001
37
38
p [†]	0.282	0.384	0.310	0.097	0.425	0.193	0.988	0.523	0.738	0.647	0.702	0.042	0.641
He (%)	79.1	80.8	86.5	80.7	83.9	87.7	70.3	78.7	81.2	77.0	77.9	62.9	73.1
Ho (%)	78.1	81.2	86.4	80.3	84.0	87.2	71.1	79.5	80.5	76.7	78.5	63.1	73.0
PD	0.923	0.903	0.967	0.939	0.956	0.972	0.860	0.925	0.938	0.912	0.916	0.812	0.879
PE	0.564	0.622	0.723	0.605	0.675	0.738	0.445	0.590	0.608	0.539	0.571	0.330	0.476

* Number of genotypes included in the analysis that fit the qualitative specifications outlined in the text; [†] Fisher exact test based on 2,000 shufflings; He: expected heterozygosity; Ho: observed heterozygosity; PD: Power of Discrimination; PE: Probability of Exclusion; ... allele not detected or not applicable.
 Interclass correlations yielding $p < 0.05$ for pairwise comparisons: FGA/D3S1358 and D3S1358/D21S11.

TABLE 3—Allele frequencies at the 13 CODIS STR loci for the North American Hispanic population.

Allele	D3S1358 N* = 690	vWA 703	FGA 939	D8S1179 679	D21S11 683	D18S51 647	D5S818 705	D13S317 703	D7S820 939	D16S539 993	TH01 993	TPOX 993	CSF1PO 994
3
5
6	0.3610	0.0065	...
7	0.0482	...	0.0138	...	0.3077	0.0020	0.0055
8	0.0066	0.0142	0.0789	0.1283	0.0136	0.0549	0.4562	0.0065
9	0.0096	0.0475	0.1679	0.0863	0.1123	0.0951	0.0645	0.0211
9.3	0.1752
10	0.0862	...	0.0077	0.0489	0.0789	0.2662	0.2205	0.0060	0.0322	0.2420
10.2	0.0008
11	0.0022	0.0021	...	0.0611	...	0.0116	0.4064	0.2198	0.2796	0.2649	...	0.2729	0.2842
12	0.0014	0.0014	...	0.1215	...	0.1221	0.3050	0.2660	0.1837	0.2638	...	0.1636	0.3597
13	0.0051	0.0050	...	0.2946	...	0.1105	0.1234	0.1273	0.0373	0.1067	...	0.0020	0.0729
13.2	0.0008
14	0.0746	0.0626	...	0.2533	...	0.1700	0.0064	0.0597	0.0048	0.0171	0.0070
14.2
15	0.3659	0.1024	0.0005	0.1318	...	0.1406	...	0.0014	...	0.0010	0.0010
15.2
16	0.2862	0.3151	...	0.0309	...	0.1182
17	0.1565	0.2639	0.0016	0.0037	...	0.1708
17.2
18	0.0993	0.1664	0.0096	0.0007	...	0.0665
18.2
19	0.0080	0.0669	0.0815	0.0309
19.2	0.0011
20	0.0007	0.0142	0.1054	0.0193
20.2
21	0.1214	0.0116
21.2
22	0.1432	0.0131
22.2	0.0021
23	0.1470	0.0031
23.2	0.0037
24	0.1539	0.0015
24.2	0.0005
25	0.1315	0.0008
25.2
26	0.0650	...	0.0015
26.2
27	0.0229	...	0.0198
28	0.0064	...	0.1069
28.2
29	0.0011	...	0.2138
29.2	0.0037
30	0.0011	...	0.2562
30.2	0.0300
31	0.0005	...	0.0827
31.2	0.1171
32	0.0095
32.2	0.1061
33	0.0015
33.2	0.0425
34	0.0029
34.2	0.0007
35	0.0037
35.2
36	0.0007
36.2
37	0.0007
38
p^{\dagger}	0.863	0.265	0.894	0.228	0.246	0.317	0.487	0.661	0.752	0.229	0.335	0.481	0.186
He (%)	74.5	78.5	88.0	80.5	84.3	87.5	72.0	82.1	79.2	78.8	73.3	68.6	72.6
Ho (%)	73.9	78.5	88.2	80.9	81.1	87.3	70.4	78.9	78.3	77.1	71.0	68.2	70.6
PD	0.894	0.923	0.973	0.936	0.957	0.971	0.879	0.944	0.926	0.922	0.884	0.852	0.876
PE	0.491	0.571	0.758	0.616	0.620	0.741	0.435	0.580	0.568	0.546	0.444	0.401	0.438

* Number of genotypes included in the analysis that fit the qualitative specifications outlined in the text; \dagger Fisher exact test based on 2,000 shufflings; He: expected heterozygosity; Ho: observed heterozygosity; PD: Power of Discrimination; PE: Probability of Exclusion; ... allele not detected or not applicable.

Interclass correlations yielding $p < 0.05$ for pairwise comparisons: D3S1358/D21S11, D3S1358/D13S317, vWA/D21S11, D21S11/D13S317 and D16S539/TH01.

Acknowledgments

The authors wish to generously thank our many Orchid BioSciences colleagues whose hard work was crucial to performing the analysis reported herein. This study was funded in whole by Orchid BioSciences, Inc.

References

1. Sambrook J, Fritsch EF, Maniatis T. *Molecular cloning: A laboratory manual*. 2nd ed. New York: Cold Spring Harbor Laboratory Press, 1989.
2. Lewis PO, Zaykin D. Genetic data analysis, computer program for the analysis of allelic data, version 1.1. Free software distributed by the au-

thors over the internet from the GDA home page at <http://lewis.eeb.uconn.edu/lewishome/software.html>, 2001.

3. Edwards A, Hammond HA, Jin L, Caskey CT, Chakraborty R. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics* 1992 Feb;12(2):241–53. [\[PubMed\]](#)
4. Weir BS. *Genetic data analysis II*. 2nd rev and exp ed. Sunderland, MA: Sinauer Associates, Inc., 1996.

Additional information and reprint requests:

David D. Einum, Ph.D.
Associate Laboratory Director
Orchid BioSciences, Inc.
2947 Eyde Parkway, Suite 110
East Lansing, MI 48823