Examples of STR Population Databases For CODIS and for Casework

Bruce Budowle and Tamyra R. Moretti
FSRTC, FBI Academy, Quantico, VA 22135

The main objective for a national DNA databank is to assist investigators in the identification of perpetrators of violent crimes. For purposes of applying DNA technology to human identity testing and to make effective use of a national DNA databank, defined polymorphic genetic markers are required, and all laboratories that contribute to the database should use the same genetic loci. The short tandem repeat (STR) loci are prime candidates for typing DNA derived from forensic biological evidence and for serving as the core loci for a national DNA databank (1).

Beginning in 1996, the FBI Laboratory sponsored a community-wide science project to validate the performance of STR typing assays, to improve the quality of STR assays, to select the core STR loci to be used in the Combined DNA Index System (CODIS), and to perform population studies using the STR loci. Representatives from 21 laboratories in the United States and Canada participated in the project. The laboratories represented were the FBI, Royal Canadian Mounted Police, Armed Forces Institute of Pathology, National Institute of Standards and Technology, Alabama Department of Forensic Sciences, Arizona Department of Public Safety, California Department of Justice, Detroit Police Department, Florida Department of Law Enforcement, Illinois State Police, Metro-Dade Police Department/ Miami Children's Research Institute, Michigan State Police, Minnesota Bureau of Criminal Apprehension, North Carolina State Bureau of Investigation, Office of the Chief Medical Examiner in New York City, Orange County Sheriff's-Coroner Laboratory, Oregon State Police, Palm Beach County Sheriff's Office, Suffolk County Crime Laboratory, University of North Texas Health Science Center, and the Virginia Division of Forensic Science.

The project has been completed. As a result of this project, high quality kits that enable multiplex amplification of STR loci are commercially available, and subsequently typing procedures have been validated. Moreover, 13 core loci have been selected for use in CODIS. The loci are: CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, and D21S11 (1,2). The final phase of the project, which is the collection of population data, also has been completed.

Over 50 population data sets, totaling more than 10,000 individuals, comprised of African Americans, Caucasians from both the United States and Europe, Hispanics, Native Americans, and Asians have been collected and typed. These data demonstrate that the STR loci are useful for providing estimates of the frequency of a DNA profile in forensic identity testing and that a multiple locus profile is extremely rare. Furthermore, the STR loci were shown to reflect the ethnohistory of human populations, as the population data of other forensically-applied genetic markers have demonstrated. In fact, the Coefficient of Gene Differentiation based on STR loci within each population group of African Americans, Caucasians, Hispanics, and Asians is less than the NRC II Report's recommended value of 0.01 (2,3). The population data are being compiled for entry into a CODIS population data file.

To provide the reader with a general appreciation of the variation that might occur between or among major population groups, this paper provides examples of allele distributions for the 13 STR loci from each major population group relevant to the United States. Only one sample population from each major population group is displayed, because the distributions of allele frequencies for sample populations within a major population group are generally similar (except for Native Americans, where variation between Native American groups may be expected to be greater). The African American sample is from California, the Caucasian sample is from Alabama, the Hispanic sample is from Florida, the Navajo sample is from Arizona, and the Vietnamese sample was collected in California.

The distributions (in percent) of observed alleles for the 13 STR loci for five sample populations are shown in Table 1. The observed and expected homozygosities, exact test for departures from Hardy-Weinberg expectations (HWE), discrimination probability (PD), and probability of exclusion (PE) are also provided. All loci are highly polymorphic in all sample populations, with the loci D3S1358 and TH01 in Navajos having the lowest observed heterozygosities (39.6% and 48.4%, respectively), and the loci FGA in Caucasians (91.3%) and D18S51 in Hispanics (90.0%) displaying the highest observed heterozygosities. Across all sample populations, the most discriminating loci tended to be FGA, D21S11, and D18S51. The TPOX locus was the least discrimin-

ating locus in Vietnamese, Caucasians and Hispanics, and the D3S1358, D5S818, and D13S317 loci were the least discriminating loci in African Americans. These observations are similar to those reported by Budowle and Moretti (2). There was little evidence for departures from HWE in any of the populations. Based on the exact test, the loci that departed significantly from HWE are: FGA ($p=0.012$, African Americans; and $p<10^{-3}$, Navajos); TH01 ($p=0.016$, Navajos); D7S820 ($p=0.029$, Vietnamese); D8S1179 ($p=0.012$, Caucasians; and $p=0.009$, Hispanics); and D21S11 ($p=0.049$, Vietnamese; and $p=0.006$, Caucasians). After employing the Bonferroni correction for the number of loci analyzed (i.e., 13 loci per database), these observations (except, possibly that of locus FGA in Navajos) are not likely to be significant.

An inter-class correlation test analysis was performed to reveal possible correlations between alleles at any of the pair-wise comparisons of the 13 loci. For each database, there is a total of 78 pair-wise comparisons performed. The number of significant departures is at or below expected levels (5%, or 4 observations) in all groups, except Hispanics. In Hispanics, six significant departures were observed (7.7% of the pair-wise tests). However, none of the empirical levels of significance for these six observations are below the adjusted Bonferroni level. Furthermore, five of these correlations for the Hispanic sample are negative; a positive value might be attributed to the effects of substructure. Based on these observations, the data do not support any significant departure from independence between pairs of loci in any sample population. With little evidence of association between loci, the assumption of independence is valid, and a multiple-locus profile frequency can be estimated using the product rule.

Differences in allele frequencies between the loci of the major population databases displayed in Table 1 were observed at the loci and might be expected. Despite these differences in allele frequencies, a 13-locus profile frequency would be rare in all five population groups. The Navajo are the least polymorphic of the five population groups, and these data may provide insight about the upper bound of estimates for isolated populations. The most common profile frequency derived from the 13 core STR loci is less than 1 in 950 million in the Navajo, and typically the estimates are substantially more rare. The estimates of the most common profile frequency for the other four population groups are at least two-to-three orders of magnitude smaller than that for Navajos.

In conclusion, a subset of the collected STR population data is presented. The data are similar to other relevant STR population studies and can be used for estimating the rarity of a multiple locus STR profile. Substantial population data, collected through the concerted effort of the forensic community, provide a solid base for the validity of STR typing and are available for assessing the significance of a multiple locus profile. Cooperative efforts such as the STR project are invaluable for enhancing the capabilities of forensic scientists to help resolve violent crimes.

ACKNOWLEGDEMENTS

We would like to express our thanks to those people and laboratories who contributed to and/or supported this study.

This is publication number 99-04 of the Laboratory Division of the Federal Bureau of Investigation. Names of commercial manufacturers are provided for identification only, and inclusion does not imply endorsement by the Federal Bureau of Investigation.

REFERENCES

1. Budowle B, Moretti TR, Niezgoda SJ, Brown, BL. CODIS and PCR-based short tandem repeat loci: Law enforcement tools. In: Second European Symposium on Human Identification 1998, Promega Corporation, Madison, Wisconsin (in press).
2. Budowle B, Moretti TR. Population data on the thirteen CODIS Core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamaians, Jamaicans, and Trinidanians. J. Forens. Sci. (Submitted).
3. National Research Council II Report. The Evaluation of Forensic Evidence. National Academy Press, Washington, D.C., 1996.

Table 1. Observed allele distributions (as %) for 13 STR loci in five population groups (Note: these data are preliminary and may be edited slightly before being entered into CODIS)

D3S1358	Vietnamese (N=213)	African American (N=200)	Caucasian (N=150)	Hispanic (N=210)	Navajo (N=182)
<12	0.000	0.250	0.333	0.238	0.000
12	0.000	0.500	0.000	0.000	0.000
13	0.469	1.250	0.667	0.952	1.374
14	3.286	8.500	15.333	8.095	3.297
15	30.047	27.500	23.000	35.238	73.901
16	34.038	36.250	25.667	25.238	13.736
17	24.883	20.500	22.667	15.714	6.593
18	5.869	5.000	10.000	14.286	0.824
19	1.408	0.250	2.333	0.238	0.275
Homozygosity (Obs.)	27.7%	25.0%	22.0%	25.2%	60.4%
Homozygosity (Exp.)	27.1%	25.7%	20.2%	23.8%	57.0%
p value	0.846	0.818	0.574	0.620	0.342
exact test	0.953	0.648	0.982	0.954	0.229
PD	0.878	0.889	0.927	0.905	0.635
PE	0.482	0.511	0.595	0.543	0.246

vWA	Vietnamese (N=215)	African American (N=200)	Caucasian (N=150)	Hispanic (N=259)	Navajo (N=182)
13	0.233	1.250	0.000	0.386	0.000
14	26.977	6.250	10.667	7.336	2.473
15	1.860	19.250	9.333	10.232	1.923
16	14.884	25.250	21.333	26.641	42.857
17	22.791	24.500	28.000	29.923	31.319
18	23.953	13.250	20.667	18.533	15.659
19	7.907	7.250	8.333	5.405	5.495
20	1.395	1.500	1.667	1.544	0.275
21	0.000	1.000	0.000	0.000	0.000
22	0.000	0.500	0.000	0.000	0.000
Homozygosity (Obs.)	20.5%	20.5%	16.7%	18.5%	30.8%
Homozygosity (Exp.)	20.9%	18.6%	19.1%	21.2%	30.8%
p value	0.870	0.490	0.444	0.287	0.984
exact test	0.849	0.757	0.515	0.665	0.268
PD	0.922	0.936	0.926	0.920	0.849
PE	0.584	0.629	0.618	0.586	0.441

FGA	Vietnamese (N=212)	African American (N=200)	Caucasian (N=150)	Hispanic (N=210)	Navajo (N=182)
<18	0.943	0.250	0.000	0.000	0.000
18	1.887	0.500	1.333	0.952	1.374
18.2	0.000	1.500	0.000	0.000	0.000
19	8.962	6.750	4.333	8.333	18.681
19.2	0.000	0.750	0.000	0.000	0.000
20	5.896	4.500	13.667	12.619	9.066
20.2	0.236	0.250	0.000	0.000	0.000
21	15.802	9.500	17.000	14.524	12.637
21.2	1.415	0.000	0.000	0.238	0.000
22	19.340	17.500	19.000	14.762	8.516
22.2	0.943	0.250	1.000	0.476	0.000
23	12.972	19.750	14.667	14.048	5.220
23.2	1.887	0.000	0.333	0.476	0.000
24	15.094	16.250	15.667	14.762	11.264
24.2	1.651	0.000	0.000	0.000	0.000
25	6.604	11.250	11.000	10.714	16.758
25.2	0.943	0.000	0.000	0.000	0.000
26	3.538	4.250	2.000	4.762	11.264
26.2	0.472	0.000	0.000	0.000	0.000
27	1.415	3.500	0.000	2.619	4.670
28	0.000	1.750	0.000	0.476	0.549
30	0.000	0.000	0.000	0.238	0.000
>30	0.000	1.500	0.000	0.000	0.000
Homozygosity (Obs.)	12.3%	14.5%	8.7%	12.4%	19.8%
Homozygosity (Exp.)	11.9%	12.6%	14.2%	12.0%	12.3%
p value	0.859	0.417	0.054	0.858	0.002
exact test	0.279	0.012	0.514	0.337	0.000
PD	0.970	0.965	0.954	0.969	0.965
PE	0.758	0.744	0.708	0.752	0.747

TH01	Vietnamese (N=210)	African American (N=200)	Caucasian (N=150)	Hispanic (N=240)	Navajo (N=182)
5	0.000	0.500	0.000	0.000	0.000
6	10.714	11.750	19.667	21.250	16.758
7	28.095	42.500	16.667	25.208	61.264
8	5.000	18.750	13.000	10.417	5.495
8.3	0.000	0.000	0.333	0.000	0.000
9	44.048	13.500	18.667	18.542	0.824
9.3	5.238	12.250	30.667	23.542	15.659
10	6.667	0.750	1.000	1.042	0.000
11	0.238	0.000	0.000	0.000	0.000
Homozygosity (Obs.)	32.9%	26.0%	20.7%	20.0%	51.6%
Homozygosity (Exp.)	29.2%	26.1%	21.0%	20.8%	42.9%
p value	0.250	0.973	0.927	0.766	0.018
exact test	0.501	0.313	0.977	0.390	0.016
PD	0.870	0.894	0.922	0.919	0.757
PE	0.478	0.522	0.582	0.582	0.339

TPOX	Vietnamese (N=210)	African American (N=200)	Caucasian (N=150)	Hispanic (N=240)	Navajo (N=182)
6	0.000	6.250	0.333	0.417	0.000
7	0.476	2.500	0.000	0.208	0.000
8	56.190	33.250	54.333	50.625	34.890
9	9.524	21.000	13.000	8.333	0.824
10	6.190	6.000	4.000	6.250	1.648
11	25.476	27.000	24.000	27.708	37.637
12	2.143	4.000	4.333	6.458	25.000
Homozygosity (Obs.)	38.1%	26.0%	32.0%	39.2%	33.0%
Homozygosity (Exp.)	39.3%	23.5%	37.1%	34.7%	32.4%
p value	0.730	0.412	0.195	0.144	0.879
exact test	0.386	0.430	0.505	0.090	0.302
PD	0.788	0.908	0.806	0.829	0.828
PE	0.364	0.548	0.386	0.411	0.392

CSF1PO	Vietnamese (N=210)	African American (N=200)	Caucasian (N=150)	Hispanic (N=240)	Navajo (N=182)
6	0.000	0.000	0.000	0.000	0.000
7	0.238	5.500	0.000	0.208	0.000
8	0.000	7.000	0.000	0.417	1.648
9	3.333	3.500	2.667	1.250	4.945
10	22.143	31.250	27.000	25.417	26.099
11	27.857	22.500	29.333	29.583	34.890
12	39.048	24.250	32.000	35.625	29.396
12.1	0.000	0.000	0.000	0.000	0.824
13	5.952	5.000	6.667	6.875	1.923
14	0.952	1.000	1.667	0.417	0.000
15	0.476	0.000	0.667	0.208	0.275
Homozygosity (Obs.)	24.3%	21.0%	26.0%	32.1%	28.6%
Homozygosity (Exp.)	28.2%	21.7%	26.4%	28.2%	27.7%
p value	0.206	0.813	0.903	0.187	0.803
exact test	0.771	0.999	0.826	0.675	0.470
PD	0.863	0.921	0.877	0.874	0.871
PE	0.471	0.579	0.490	0.462	0.471

D5S818	Vietnamese (N=212)	African American (N=200)	Caucasian (N=150)	Hispanic (N=259)	Navajo (N=182)
7	2.830	0.500	0.000	2.317	16.209
8	0.236	5.250	0.000	1.158	3.571
9	6.132	1.750	3.000	4.826	0.275
10	23.349	6.000	3.667	4.826	6.044
11	30.425	28.250	41.667	38.224	58.242
12	23.113	35.750	36.667	32.625	10.714
13	13.443	20.750	13.667	15.444	4.945
14	0.236	1.000	1.000	0.579	0.000
15	0.236	0.500	0.333	0.000	0.000
>15	0.000	0.250	0.000	0.000	0.000
Homozygosity (Obs.)	19.8%	26.0%	33.3%	30.5%	37.9%
Homozygosity (Exp.)	22.1%	25.6%	32.7%	28.0%	38.3%
p value	0.416	0.887	0.865	0.377	0.922
exact test	0.096	0.614	0.673	0.134	0.837
PD	0.905	0.890	0.834	0.876	0.822
PE	0.565	0.515	0.412	0.483	0.404

D7S820	Vietnamese (N=212)	African American (N=200)	Caucasian (N=150)	Hispanic (N=259)	Navajo (N=182)
6	0.000	0.000	0.333	0.193	0.000
7	0.236	0.500	2.667	0.965	0.275
8	15.094	23.250	15.667	15.444	12.363
9	7.075	11.250	16.000	11.969	0.000
10	15.094	34.000	29.667	26.255	14.286
11	39.387	21.750	15.000	22.780	40.934
12	20.047	7.750	16.333	18.340	28.297
13	2.594	1.000	3.667	3.668	3.846
14	0.472	0.500	0.667	0.386	0.000
Homozygosity (Obs.)	25.9%	26.5%	16.0%	19.7%	23.6%
Homozygosity (Exp.)	24.5%	23.4%	18.7%	19.3%	28.3%
p value	0.621	0.298	0.401	0.858	0.163
exact test	0.029	0.910	0.675	0.539	0.029
PD	0.900	0.908	0.933	0.931	0.861
PE	0.542	0.547	0.627	0.614	0.476

D18S51	Vietnamese (N=215)	African American (N=200)	Caucasian (N=150)	Hispanic (N=210)	Navajo (N=185)
<11	0.233	0.500	1.333	0.952	0.000
11	0.465	0.750	0.333	1.429	0.270
12	4.186	6.000	13.000	13.333	9.459
13	11.860	4.750	14.667	12.381	32.162
13.2	0.000	0.500	0.000	0.000	0.000
14	18.605	6.250	18.000	13.571	13.784
14.2	0.000	1.000	0.000	0.000	0.000
15	26.744	19.000	15.667	18.571	7.838
16	14.651	17.250	11.000	15.000	18.378
17	10.698	15.250	13.333	9.762	11.622
18	3.256	9.500	7.333	5.476	1.892
19	3.256	9.500	2.667	4.286	2.432
20	2.093	4.250	2.000	3.095	0.811
21	1.395	3.500	0.667	0.952	1.081
22	1.628	1.750	0.000	0.952	0.000
>22	0.931	0.250	0.000	0.238	0.270
Homozygosity (Obs.)	19.1%	11.0%	16.7%	10.0%	20.5%
Homozygosity (Exp.)	15.6%	11.8%	12.9%	12.2%	18.4%
p value	0.161	0.717	0.169	0.325	0.447
exact test	0.288	0.363	0.424	0.957	0.350
PD	0.956	0.970	0.964	0.969	0.942
PE	0.690	0.759	0.732	0.749	0.643

D8S1179	Vietnamese (N=214)	African American (N=200)	Caucasian (N=150)	Hispanic (N=210)	Navajo (N=185)
<9	0.000	0.250	1.667	1.429	0.000
9	0.000	0.250	0.333	0.952	0.000
10	14.953	0.750	12.333	9.286	14.865
11	13.318	6.250	9.667	5.000	5.135
12	10.981	10.750	8.667	10.476	10.811
13	16.822	22.500	31.333	36.429	37.027
14	13.318	31.000	22.000	21.667	22.162
15	16.822	19.750	10.333	11.905	8.378
16	11.449	6.000	3.000	2.619	1.622
17	1.636	2.500	0.667	0.238	0.000
18	0.701	0.000	0.000	0.000	0.000
Homozygosity (Obs.)	17.8%	22.0%	19.3%	27.6%	25.4%
Homozygosity (Exp.)	13.8%	20.3%	18.8%	21.5%	22.8%
p value	0.092	0.562	0.863	0.031	0.395
exact test	0.052	0.567	0.012	0.009	0.682
PD	0.960	0.929	0.929	0.924	0.917
PE	0.715	0.602	0.633	0.595	0.569

D16S539	Vietnamese (N=210)	African American (N=200)	Caucasian (N=150)	Hispanic (N=240)	Navajo (N=185)
5	0.000	0.250	0.000	0.000	0.000
8	0.714	4.000	0.667	2.292	0.000
9	22.143	18.250	11.333	14.583	16.757
10	16.429	10.500	5.000	9.583	16.486
11	26.190	31.000	31.333	28.125	14.595
12	23.571	20.750	31.000	25.417	35.946
13	9.524	13.000	19.000	16.250	14.324
14	1.429	2.250	1.667	3.542	1.892
15	0.000	0.000	0.000	0.208	0.000
Homozygosity(Obs.)	17.1%	20.0%	28.7%	20.0%	20.5%
Homozygosity(Exp.)	20.8%	20.1%	24.4%	20.1%	22.5%
p value	0.196	0.986	0.218	0.979	0.533
exact test	0.770	0.068	0.596	0.696	0.660
PD	0.916	0.923	0.901	0.929	0.914
PE	0.584	0.605	0.529	0.603	0.567

D21S11	Vietnamese (N=215)	African American (N=200)	Caucasian (N=150)	Hispanic (N=210)	Navajo (N=185)
<24.2	0.000	0.000	0.333	0.000	0.000
26	0.000	0.250	0.333	0.238	0.000
26.2	0.000	0.250	0.000	0.000	0.000
27	0.465	3.750	4.000	1.667	0.000
28	5.116	28.750	17.667	12.381	5.405
28.2	0.000	0.000	0.000	0.000	0.000
29	23.721	20.000	18.667	23.333	17.838
29.2	0.000	0.250	0.000	0.238	0.000
30	26.977	14.750	27.333	24.762	50.811
30.2	1.395	2.250	2.333	3.333	0.000
31	8.837	7.000	8.000	7.619	4.865
31.2	8.140	3.750	9.333	9.524	6.216
32	2.791	1.000	1.667	0.952	0.000
32.2	15.814	8.000	7.000	10.952	11.892
33	0.930	0.250	0.000	0.000	0.000
33.1	0.000	0.500	0.000	0.000	0.000
33.2	5.349	3.500	3.000	3.571	2.703
34	0.000	1.000	0.000	0.000	0.000
34.1	0.000	0.250	0.000	0.000	0.000
34.2	0.465	0.250	0.333	0.952	0.270
35	0.000	3.500	0.000	0.000	0.000
36	0.000	0.750	0.000	0.238	0.000
>36	0.000	0.000	0.000	0.238	0.000
Homozygosity(Obs.)	20.5%	15.0%	18.7%	13.3%	32.4%
Homozygosity(Exp.)	17.3%	16.0%	16.1%	15.9%	31.2%
P Value	0.222	0.708	0.399	0.312	0.721
Exact Test	0.049	0.930	0.006	0.451	0.853
PD	0.946	0.955	0.944	0.949	0.875
PE	0.659	0.688	0.679	0.684	0.478

D13S317	Vietnamese (N=211)	African American (N=200)	Caucasian (N=150)	Hispanic (N=259)	Navajo (N=182)
7	0.000	0.000	0.000	0.000	0.000
8	36.256	4.500	11.667	11.197	0.824
9	11.374	3.250	8.000	11.583	24.725
10	11.611	2.250	6.667	7.915	15.110
11	20.142	27.000	32.000	31.660	22.253
12	16.825	40.500	26.667	22.201	19.780
13	2.133	15.250	11.000	10.232	15.934
14	0.948	7.250	3.667	5.212	1.374
15	0.711	0.000	0.333	0.000	0.000
Homozygosity(Obs.)	26.5%	29.5%	20.7%	23.9%	19.2%
Homozygosity(Exp.)	22.6%	26.7%	20.9%	19.3%	19.6%
P Value	0.166	0.374	0.949	0.061	0.899
Exact Test	0.940	0.493	0.389	0.064	0.799
PD	0.917	0.891	0.927	0.931	0.929
PE	0.569	0.508	0.596	0.624	0.603