1817M

Final Progress Report of Award No. 98-LB-VX-0002
"Evaluation of New STR Markers for Forensic Analysis"
Principal Investigator: Ranjan Deka

We had received support to characterize a set of ethnically and geographically diverse human populations with respect to the occurrence and frequency of alleles at STR loci representing a large fraction of the human genome; examine the generality of fit of genotypes to HWE and allelic independence across the studied loci. The objective was to generate worldwide databases of allele frequencies at these loci, obtain indirect estimates of mutation rates from population-locus contrasts and to develop a panel of STR markers suitable for individual identification and parentage testing. The project has just completed, some data are still being analyzed. In the following, we submit a report to the National Institute of Justice on the work that has been accomplished.

We analyzed DNA samples from 19 ethnically distinct populations belonging to five major continental populations. These populations are:

Africans: Sudanese, Nigerian, Benin, South Carolina Black
Caucasian: German, Spanish, United Arab Emirates, Brazilian White
Asian: Chinese, Japanese, Kachari (Northeast India), Thai, Kampuchean
Native American: Dogrib, Bri Bri, Pehuenche, Panama Indian
Pacific Islander: Samoan, New Guinea Highlander

We characterized a set of 54 unlinked tri- and tetranucleotide loci distributed over all of the autosomal chromosomes. These markers were chosen based on their observed heterozygosity between 70 and 85% among the Caucasians, which corresponds to a mutation rate of 1.5×10^{-4} to 10^{-5} for an effective population size of 5000. From this set of markers we selected a set of 32 markers, which are user-friendly for PCR and analysis on silver-stained gels as well as showing agreement to acceptable population genetic properties such as HWE, linkage equilibrium. These markers can be analyzed in 11 maultiplex PCR reactions and resolved on silver stained polyacrylamide gels as shown below:

Multiplex Panels:
1. CSF1R, TH01, PLA2A1
2. F13A1, CYP19, LPL
3. D1S552, D6S1006, D1S1453
4. D11S1392, D10S1239, D2S1649
5. D2S1352, D21S1440, D13S325
6. D4S2394, D3S3045, D3S2409
7. D9S930, D9S925, D4S2366
8. D8S1110, D1S518, D8S1132
9. D5S816, D12S1064, D12S1042
10. D20S481, D20S473, D20S604
11. D21S1446, D21S1435

The most significant achievement of this project, however, was characterization of genetic variation at nine CODIS loci, because of their direct relevance in forensic analysis. Most forensic laboratories in the US are now using these loci. However, databases from ethnically defined populations are still lacking. These are preliminary data and we propose to analyze them in greater detail, for which we have requested further support from the NIJ. We generated genotype data at these loci in a total of 982 individuals from the19 ethnically defined populations named above, representing five major human groups (Table 1).

Table 1. Summary Statistics of within Population Variation at 9 STR Loci in 19 Global Populations

Population (N)	Average (s.e.) over 9 loci		
	Number of alleles	Allele size variance	Expected Heterozygosity
African			
Sudanese (46)	9.1 (1.2)	3.46 (0.77)	0.813 (0.02)
Nigerian (46)	9.4 (1.2)	2.88 (0.61)	0.794 (0.02)
Benin (51)	9.2 (1.1)	2.90 (0.62)	0.792 (0.02)
S.C. Black (48)	8.9 (1.1)	3.12 (0.69)	0.797 (0.03)
Caucasian			
German (49)	8.4 (0.7)	2.63 (0.44)	0.814 (0.02)
Spanish (46)	8.6 (0.8)	2.72 (0.40)	0.807 (0.02)
United Arab Emirates (53)	8.6 (0.9)	2.83 (0.55)	0.811 (0.02)
Brazilian (81)	9.4 (0.8)	2.94 (0.50)	0.817 (0.02)
Asian			
Chinese (103)	8.7 (0.9)	2.61 (0.42)	0.802 (0.02)
Japanese (47)	8.6 (0.7)	3.08 (0.66)	0.799 (0.02)
Kachari (54)	8.9 (0.8)	3.05 (0.56)	0.816 (0.02)
Thai (48)	8.7 (1.2)	2.86 (0.43)	0.810 (0.02)
Kampuchean (39)	7.9 (0.8)	2.56 (0.31)	0.806 (0.01)
Native American			
Dogrib (48)	5.9 (0.6)	2.39 (0.38)	0.744 (0.03)
Panama Indian (44)	6.9 (0.9)	2.49 (0.66)	0.698 (0.04)
Bri Bri (43)	6.6 (0.7)	2.64 (0.73)	0.733 (0.04)
Pehuenche (37)	6.8 (0.7)	3.07 (0.87)	0.732 (0.04)
Pacific Islander			
Samoan (48)	7.7 (0.5)	2.41 (0.32)	0.785 (0.01)
New Guinea Highlander (51)	6.4 (0.5)	2.53 (0.54)	0.755 (0.01)

As shown in Table 1, the 9 loci studied thus far are all highly polymorphic even in the isolated populations (such as the Native Americans and Pacific Islanders). This is true with respect to all three measures of genetic diversity with a locus (number of alleles, allele size variance and heterozygosity). Therefore, it can be argued that the discriminatory power of the battery of markers is expected to be adequate even for casework in which the source of DNA is from an isolated population.

Table 2. Tests of Multi-Locus Independence of Allele Frequencies in 19 Global Populations

Population	S_k^2 (95% CI)	Mean (s.d.) number of shared alleles	
		Observed	Expected
Sudanese	1.22 (0.89 – 2.11)	5.0 (1.6)	5.4 (1.7)
Nigerian	1.37 (0.77 – 1.85)	5.6 (1.7)	5.7 (1.7)
Benin	1.90 (0.88 – 1.97)	5.5 (1.7)	5.8 (1.8)
S.C. Black	1.05 (0.85 – 1.97)	5.4 (1.7)	5.5 (1.7)
German	1.33 (0.77 – 1.81)	5.3 (1.8)	5.4 (1.7)
Spanish	1.09 (0.77 – 1.85)	5.4 (1.8)	5.6 (1.7)
United Arab Emirates	1.45 (0.82 – 1.84)	5.4 (1.7)	5.5 (1.7)
Brazilian	1.24 (0.95 – 1.81)	5.1 (1.7)	5.3 (1.7)
Chinese	1.62 (1.09 – 1.90)	5.4 (1.7)	5.6 (1.8)
Japanese	1.41 (0.82 – 1.93)	5.6 (1.8)	5.7 (1.8)
Kachari	1.20 (0.82 – 1.83)	5.2 (1.8)	5.4 (1.7)
Thai	1.67 (0.80 – 1.88)	5.3 (1.8)	5.5 (1.7)
Kampuchean	1.68 (0.80 – 2.04)	5.5 (1.7)	5.7 (1.8)
Dogrib	2.22 (1.00 – 2.30)	6.6 (1.8)	6.7 (1.8)
Panama Indian	1.92 (1.06 – 2.49)	6.4 (1.6)	6.8 (1.7)
Bri Bri	1.62 (0.97 – 2.31)	6.6 (1.8)	6.6 (1.7)
Pehuenche	2.20 (0.87 – 2.26)	6.7 (1.9)	6.7 (1.7)
Samoan	0.85 (0.82 – 1.92)	6.1 (1.7)	6.1 (1.8)
New Guinea Highlander	1.25 (1.02 – 2.28)	6.4 (1.8)	6.6 (1.8)

Note: n = No. of individuals with 9-locus genotype data available; S_k^2 = Variance of the number of heterozygous loci in 9-locus genotype, computed over all individuals in the population. The number of shared alleles was evaluated by pairwise comparisons of 9-locus genotypes for all possible pairs of individuals within the population.

Table 2 presents a summary of mutual test of independence of the 9 loci. In a separate manuscript (Chakraborty et al., in preparation), we have shown that when each multi-locus genotype occurs only once in a sample, the summed number of heterozygous loci is a sufficient statistic for testing the hypothesis of mutual independence of loci. Therefore, the test statistic, used in Table 2 for testing the above hypothesis, contains all information in a database of multi-locus genotypes that is relevant for answering the issue of mutual independence of alleles across all loci.

Table 3. Estimates of Coefficient of Gene Differentiation (G_{ST}) among Populations for five major Groups of Humans based on 9 STR loci

Population Groups	Based on gene diversity		Based on allele size variance	
	G_{ST} (H) in %	Prob.	G_{ST} (V) in %	Prob.
African	0.18 ± 0.16	0.049	0.80 ± 0.44	0.006
Caucasian	0.22 ± 0.07	0.011	0.21 ± 0.31	0.148
Asian	0.50 ± 0.10	$<10^{-4}$	0.47 ± 0.34	0.021
Native American	3.47 ± 0.38	$<10^{-4}$	4.38 ± 1.14	$<10^{-4}$
Pacific Islander	2.27 ± 0.63	$<10^{-4}$	9.20 ± 3.80	$<10^{-4}$

Tables 3 and 4 provide a summary of gene diversity analyses of the 9 loci we have examined thus far. For geographic populations within each of the five major groups, we evaluated the coefficient of gene diversity (G_{ST}, equivalent to coefficient of coancestry, θ) by two methods in Table 3. Although for studying the evolutionary relationship between populations, the allele-size variance based estimates are preferred, for forensic applications, the estimates shown in the second column (G_{ST} H) are the most relevant. Thus, data shown in Table 3 establishes two points worthy to note. First, for all major groups estimates of $\theta<3\%$ are adequate (as suggested in the recent report of NRC, 1996). Second, the levels of significance (under the heading of Probability of Table 3), obtained by a permutation-based method (detailed in a recent publication of our group, Chakraborty et al. 1999), indicate that even small values of θ can be statistically significant. In other words, even when two databases from two different samples from the same population show statistically significant differences of allele frequencies, such observations do not compromise forensic calculations, since such departures can be taken into account in forensic calculations by invoking values of θ suggested in the NRC (1996) report.

Table 4. Gene Diversity Analysis of 19 Global Populations subdivided as five major Groups and Sub-populations within each Group

Locus	Between groups		Between populations within group			
	G_{gt} (H) in %	G_{gt} (V) in %	G_{sg} (H) in %	Prob.	G_{sg} (V) in %	Prob.
D3S1358	2.21	7.04	1.16	0.0002	0.85	0.0096
VWA	4.11	0.66	1.81	$<10^{-4}$	5.03	$<10^{-4}$
FGA	1.77	4.11	1.33	$<10^{-4}$	6.31	$<10^{-4}$
D8S1179	1.78	3.76	1.05	$<10^{-4}$	0.39	0.1018
D21S11	2.30	3.78	1.65	$<10^{-4}$	2.11	$<10^{-4}$
D18S51	1.64	2.92	1.17	$<10^{-4}$	1.96	$<10^{-4}$
D5S818	3.56	10.66	1.32	$<10^{-4}$	1.82	$<10^{-4}$
D13S317	5.78	10.82	1.91	$<10^{-4}$	1.73	$<10^{-4}$
D7S820	2.28	6.32	0.64	0.0006	2.12	$<10^{-4}$
Average	2.81	5.11	1.34	$<10^{-4}$	2.77	$<10^{-4}$
s.e.	0.47	1.08	0.13		0.80	

Table 4 illustrates another aspect of the gene diversity analysis. Note that the estimates of G_{ST} for geographic populations within each of the major groups are smaller than that among groups. Thus, with the data that we have generated thus far, we have now obtained an empirical support of the notion that databasing for forensic calculations done on the basis of broad definition of population is adequate, a point correctly asserted by Chakraborty and Kidd (1991).

Table 5. Match Probability and Paternity Exclusion Probability with the Combined Testing of 9 STR Loci in Global Populations

Population	Match Probability with			Exclusion Probability in %			
				With M, C data and exclusion based on		With data on C only and exclusion based on	
	$\theta = 0$	$\theta = 0.01$	$\theta = 0.03$	≥ 1 locus	≥ 2 loci	≥ 1 locus	≥ 2 loci
Sudanese	8.2×10^{-12}	7.8×10^{-12}	7.2×10^{-12}	99.990	99.813	99.654	96.787
Nigerian	3.7×10^{-11}	3.6×10^{-11}	3.3×10^{-11}	99.979	99.659	99.417	95.179
Benin	5.2×10^{-11}	5.0×10^{-11}	4.6×10^{-11}	99.974	99.610	99.335	94.700
S.C. Black	2.0×10^{-11}	2.0×10^{-11}	1.8×10^{-11}	99.984	99.733	99.540	95.948
German	1.2×10^{-11}	1.1×10^{-11}	9.9×10^{-12}	99.987	99.787	99.600	96.465
Spanish	2.0×10^{-11}	1.9×10^{-11}	1.7×10^{-11}	99.984	99.736	99.517	95.896
United Arab Emirates	1.4×10^{-11}	1.4×10^{-11}	1.2×10^{-11}	99.986	99.766	99.567	96.224
Brazilian	6.2×10^{-12}	5.9×10^{-12}	5.3×10^{-12}	99.991	99.836	99.686	97.065
Chinese	2.6×10^{-11}	2.5×10^{-11}	2.2×10^{-11}	99.981	99.704	99.458	95.543
Japanese	4.3×10^{-11}	4.1×10^{-11}	3.7×10^{-11}	99.976	99.643	99.359	94.928
Kachari	9.9×10^{-12}	9.4×10^{-12}	8.5×10^{-12}	99.988	99.799	99.623	96.614
Thai	1.3×10^{-11}	1.2×10^{-11}	1.1×10^{-11}	99.987	99.777	99.591	96.366
Kampuchean	3.2×10^{-11}	3.1×10^{-11}	2.8×10^{-11}	99.979	99.673	99.403	95.203
Dogrib	1.3×10^{-9}	1.2×10^{-9}	1.1×10^{-9}	99.884	98.647	98.041	87.993
Panama Indian	5.5×10^{-9}	5.4×10^{-9}	5.3×10^{-9}	99.800	97.866	97.054	83.598
Bri Bri	1.5×10^{-9}	1.4×10^{-9}	1.3×10^{-9}	99.885	98.611	98.065	87.788
Pehuenche	1.5×10^{-9}	1.4×10^{-9}	1.3×10^{-9}	99.887	98.642	98.135	88.243
Samoan	1.6×10^{-10}	1.5×10^{-10}	1.4×10^{-10}	99.955	99.398	98.962	92.713
N.G. Highlander	1.2×10^{-9}	1.1×10^{-9}	1.0×10^{-9}	99.883	98.685	97.963	87.920

Finally, in Table 5 we illustrate the power of the battery of the 9 loci for forensic and parentage testing applications. As seen in this table, in general, the 9 loci do indeed have adequate discriminatory power (for forensic identification of individuals) as well as large enough exclusionary power (for parentage analysis). However, this table also illustrates that additional loci may be needed for further advancement of DNA forensic technologies. This is guided by several observations. First, one should note that with the increased efficiency of DNA typing more complex cases are now increasingly becoming part of DNA forensic analysis. Such cases include statistical assessment of DNA mixtures and parentage analysis involving relatives that are not typed. In such cases, the exclusionary power or the likelihood ratio, obtained from 9-locus genotypes may not reach an appreciable value. Thus, additional loci are needed. Furthermore, as some of the loci may show occasional mutations in parentage analysis, assessment of cases exhibiting one (or even two) locus exclusions are needed. The efficiency of this should also be increased with more loci typed and validated. Third, in part the conservative features of the calculations shown in Table 5 are prompted by rather high values of minimum threshold allele frequency we used, because in our work thus far, for several populations the sample sizes were not adequate for forensic databases. These justify the need of further work that we propose. We contend that additional genotyping of individuals are needed so that for each locus we have genotype data on at least 100 individuals for each population. Further, we also need to document that all of the above assertions are valid

for all of the current set of 13 CODIS loci and any additional loci developed by forensic laboratories

In summary, work done from our current project has now provided empirical support of the statistical calculations suggested in the NRC (1996) report. At the same time, these results also illustrate that with the completion of similar work on all of 14 loci (13 CODIS and the Penta-E locus) in enlarged samples of at least 100 individuals from 20 well-defined populations, we will be able to provide a resource to the forensic community to do case work analysis in almost any population of the world, including the complex cases now emerging in the community. For this purpose, as stated above, we have requested further support from the National Institute of Justice.

Publications:

1. Deka R, Guangyun S, Smelser D, Zhong Y, Kimmel M, Chakraborty R. 1999. Rate and directionality of mutations and effects of allele size constraints at anonymous, gene-associated, and disease-causing trinucleotide loci. Mol. Biol. Evol. 16:1166-1177.
2. Parra E, Shriver MD, Soemantri A, McGarvey ST, Hundrieser J, Saha N, Deka R. 1999. Analysis of five Y-specific microsatellite loci in Asian and Pacific populations. Am. J. Phys. Anthrop. 110:1-16
3. Parra E, Saha N, Soemantri A, McGarvey ST, Hundrieser J, Shriver MD, Deka R. 1999. Genetic variation at 9 autosomal microsatellite loci in Asian and Pacific populations. Hum. Biol. 71:757-779.
4. **Deka R**, Shriver MD, Yu LM, Heidreich EM, McGarvey ST, Bunker CH, Miki T, Yin S-H, Raskin S, Barrantes R, Ferrell RE, Chakraborty R. Genetic variation at 23 microsatellite loci in 16 human populations. J Genet (in press)
5. Chakraborty R, Zhong Y, Stivers DN, **Deka R**, Kimmel M. Estimation and test of significance of gene diversity at microsatellite loci in a substructured population. Ann Hum Genet (submitted).
6. Deka R, Guangyun S, Zhong Y, Chakraborty R. World-wide genetic variation at nine STR (CODIS) loci and forensic implications. J. Forensic Sci. (submitted).