Variation at 4 Short Tandem Repeat Loci in 8 Population Groups of India

NAMITA MUKHERJEE,[1] PARTHA P. MAJUMDER,[1] BIDYUT ROY,[1] MONAMI ROY,[1] BADAL DEY,[1] MADAN CHAKRABORTY,[1] AND SANAT BANERJEE[1]

Abstract   We have determined the nature and extent of variation at 4 STR loci (*CSF1PO, TPOX, TH01, VWA*) in 8 caste and tribal population groups of eastern and northern India. Large differences in allele frequencies among the groups were found. Average heterozygosities in all populations were high ($\approx 80\%$). The overall extent of gene differentiation among the 8 groups was high ($G_{ST} = 0.04$). The nature of genomic affinities based on these 4 STR loci does not completely agree with our earlier finding based on classical genetic markers that geographic proximity of habitat has a greater influence on genetic similarity between populations than sociocultural proximity does.

Short tandem repeat (STR) polymorphisms have proven to be extremely useful in the understanding of genomic diversity and affinities of human populations (Edwards et al. 1992; Hammond et al. 1994; Deka et al. 1995; Perez-Lezaun et al. 1997). Although a large number of studies have been carried out among Indian ethnic groups using classical polymorphisms (e.g., blood groups, serum proteins, red cell enzymes), reports from Indian populations on the nature and extent of diversity at STR loci are few (Deka et al. 1995; Pandian et al. 1998). The objective of this study is to assess the patterns of genomic diversity and affinities within and between Indian ethnic groups belonging to different levels of the sociocultural hierarchy and inhabiting disparate geographic areas using 4 STR loci.

It is possible that populations living in close geographic proximity are more likely to exchange genes, thereby enhancing genetic similarity, despite the fact that these populations may not belong to the same sociocultural level. Therefore, in particular, we sought to examine whether genomic affinities among populations belonging to the same level of the sociocultural hierarchy irrespective of their geographic location of habitat are closer than populations

Table 1. Study Populations, Sampling Locations, and Anthropological Information

Population	Sampling Location	Anthropological Information
Agharia	Sundergarh District, Orissa	Hindu Middle caste; Indo-European language primarily agriculturists.
Bagdi	Hooghly District, West Bengal	Hindu low caste; Indo-European language; primary occupations are cultivation and fishi usually admit members of any caste higher than themselves in social ranking.
Brahmin (Uttar Pradesh)	Garhwal, Uttar Pradesh	Hindu upper caste; Indo-European language; traditionally priests but now various occupations.
Brahmin (West Bengal)	Various locations, West Bengal	Hindu upper caste; Indo-European language; traditionally priests but now various occupations.
Chamar	Garhwal, Uttar Pradesh	Hindu low caste; Indo-European language; leather workers.
Gaud	Sundergarh District, Orissa	Hindu middle caste; Indo-European language, primarily agriculturists.
Santal	Medinipur, West Bengal	Tribal group; Austro-Asiatic language; numerically large; wide geographic distribution.
Tanti	Sundergarh District, Orissa	Hindu low caste; Indo-European language; traditionally weavers.

living in close geographic proximity irrespective of their sociocultural affiliation.

The 4 STR loci used for this study were *CSF1PO, TPOX, TH01,* and *VWA,* all of which have specific tetranucleotide repeat motifs. The repeat unit of *CSF1PO* (chromosomal location 5q33.3–q34) is AGAT; for *TPOX* (2p23–2pter) and *TH01* (11p15.5) the repeat unit is AATG, and for *VWA* (12p) it is AGAT. All these loci are polymorphic in most global populations with respect to the repeat number of their respective motifs.

Eight ethnic groups inhabiting different geographic areas and belonging to various levels of the sociocultural hierarchy have been studied. The names of the populations, study locations, and pertinent anthropological information are given in Table 1.

From each unrelated voluntary participant 5–10 ml of blood was collected by venipuncture into tubes containing EDTA. Genomic DNA was isolated from collected blood samples using a standard protocol (Miller et al. 1988). Polymerase chain reaction (PCR) amplification was carried out by using approximately 40 ng of genomic DNA as a template in a final volume of 10 μl containing 10× STR buffer (1 μl), *Taq* polymerase (0.25 U), and primer mix (1.0 μl). Primers for PCR amplification of the STR loci, locus-specific allele ladders, and all other reagents (buffers, etc.) were purchased from Promega Corporation. Protocols suggested by the manufacturer were used at all stages of DNA analysis, including multiplexing. Electrophoresis

of the PCR product was carried out on 6% polyacrylamide-urea gel, which was stained with silver nitrate for band visualization. Bands were matched against locus-specific allele ladders.

Allele frequencies were calculated using the gene-counting method. A gene diversity analysis (Nei 1973) was performed, genetic distances were calculated using the D_A distance measure (Nei et al. 1983), and a neighbor-joining tree was constructed (Saitou and Nei 1987).

The allele frequencies at the 4 STR loci for each population are given in Table 2. At the *CSF1PO* locus, the allele range (number of repeat units) is 7–15 with peaks at 11 and 12 repeats. It is also seen that lower repeat numbers are present only in the lower castes (Chamar and Bagdi). Allele *15* is found only in the tribal group of Santals.

At the *TPOX* locus all populations, except Bagdi and Agharia, show a maximum frequency of allele *11;* among Bagdis and Agharias the peak is at allele *8*. Among the tribal Santals the frequencies of 8 repeats and 11 repeats are equal.

At the *TH01* locus peaks between populations are variable, but in general, alleles *6* and *9* occur with high frequencies. It is noteworthy that the tribal Santals show a very high frequency (65%) of allele *9* at this locus.

At the *VWA* locus repeat numbers 16–18 are common in all populations. A low repeat number allele, allele *13*, was found only among the low-caste Bagdis.

A comparative picture of allele frequencies among populations is provided in Figure 1. At the *TH01* and *VWA* loci, for which data are available from many global populations, the observed ranges of repeat numbers agree with previous studies [Gusmao et al. (1997) and references therein].

We are aware of 2 other Indian populations—Kachari (Assam) and Brahmin (Uttar Pradesh)—that have been studied (Deka et al. 1995) with respect to the *CSF1PO* and *TH01* loci. Although allele frequency patterns in these 2 populations are largely similar to those reported in the present study, there is a notable difference. At the *TH01* locus we found 7.41% of Uttar Pradesh Brahmin chromosomes to carry 10 repeats; this allele was not reported by Deka et al. (1995). Furthermore, at the *TH01* locus, although the 6-repeat allele is frequent in most Indian populations, this allele is absent or has a low frequency among Europeans, Africans, and Chinese (Gusmao et al. 1997).

Based on allele frequencies at these 4 STR loci, average heterozygosities were computed; these are also presented in Table 2. All populations show high levels of heterozygosity ($\approx$80%). Gene diversity analysis was also performed to examine the extent of gene differentiation among the populations. The results are presented in Table 3. The coefficient of gene differentiation among the populations is variable across loci. The *CSF1PO* and *TH01* loci show a 3-fold higher level of gene differentiation than the *TPOX* or *VWA* locus. A reduced level of gene differentiation may result from effects of nat-

Table 2. Allele Frequencies (± SE) at 4 STR Loci and Average Heterozygosities in 8 Ethnic Groups of India

Locus	Allele (Repeat Number)	Agharia (n = 46)	Bagdi (n = 62)	Uttar Pradesh Brahmin (n = 54)	West Bengal Brahmin (n = 46)	Chamar (n = 42)	Gaud (n = 26)	Santal (n = 40)	Tanti (n = 32)
CSF1PO	7	0.000	0.000	0.000	0.000	0.024 ± 0.024	0.000	0.000	0.000
	8	0.000	0.016 ± 0.016	0.000	0.000	0.000	0.000	0.000	0.000
	9	0.000	0.032 ± 0.022	0.074 ± 0.036	0.000	0.024 ± 0.024	0.000	0.025 ± 0.025	0.000
	10	0.413 ± 0.073	0.209 ± 0.052	0.278 ± 0.061	0.239 ± 0.063	0.214 ± 0.063	0.038 ± 0.038	0.025 ± 0.025	0.267 ± 0.081
	11	0.391 ± 0.072	0.339 ± 0.060	0.333 ± 0.064	0.304 ± 0.068	0.381 ± 0.075	0.231 ± 0.083	0.325 ± 0.074	0.267 ± 0.081
	12	0.174 ± 0.056	0.339 ± 0.060	0.222 ± 0.057	0.370 ± 0.071	0.214 ± 0.063	0.692 ± 0.091	0.375 ± 0.077	0.333 ± 0.086
	13	0.000	0.048 ± 0.027	0.074 ± 0.036	0.087 ± 0.042	0.119 ± 0.050	0.038 ± 0.038	0.200 ± 0.063	0.133 ± 0.062
	14	0.022 ± 0.022	0.016 ± 0.016	0.018 ± 0.018	0.000	0.024 ± 0.024	0.000	0.025 ± 0.025	0.000
	15	0.000	0.000	0.000	0.000	0.000	0.000	0.025 ± 0.025	0.000
TPOX	7	0.000	0.065 ± 0.031	0.000	0.000	0.000	0.000	0.000	0.000
	8	0.375 ± 0.070	0.323 ± 0.059	0.327 ± 0.065	0.217 ± 0.061	0.214 ± 0.063	0.231 ± 0.083	0.368 ± 0.078	0.300 ± 0.084
	9	0.125 ± 0.048	0.226 ± 0.053	0.077 ± 0.037	0.261 ± 0.065	0.238 ± 0.066	0.269 ± 0.087	0.132 ± 0.055	0.167 ± 0.068
	10	0.208 ± 0.059	0.113 ± 0.040	0.154 ± 0.050	0.174 ± 0.056	0.191 ± 0.061	0.115 ± 0.063	0.132 ± 0.055	0.100 ± 0.055
	11	0.292 ± 0.066	0.258 ± 0.056	0.404 ± 0.068	0.283 ± 0.066	0.357 ± 0.074	0.385 ± 0.095	0.368 ± 0.078	0.433 ± 0.090
	12	0.000	0.016 ± 0.016	0.038 ± 0.027	0.065 ± 0.036	0.000	0.000	0.000	0.000
TH01	6	0.292 ± 0.066	0.290 ± 0.058	0.296 ± 0.062	0.304 ± 0.068	0.262 ± 0.068	0.286 ± 0.085	0.075 ± 0.042	0.133 ± 0.062
	7	0.146 ± 0.051	0.129 ± 0.043	0.093 ± 0.039	0.283 ± 0.066	0.071 ± 0.040	0.286 ± 0.085	0.125 ± 0.052	0.167 ± 0.068
	8	0.250 ± 0.063	0.129 ± 0.043	0.074 ± 0.036	0.109 ± 0.046	0.047 ± 0.033	0.107 ± 0.058	0.125 ± 0.052	0.133 ± 0.062
	9	0.146 ± 0.051	0.339 ± 0.060	0.296 ± 0.062	0.152 ± 0.053	0.452 ± 0.077	0.214 ± 0.078	0.650 ± 0.075	0.467 ± 0.091
	9.3	0.146 ± 0.051	0.048 ± 0.027	0.167 ± 0.051	0.000	0.143 ± 0.054	0.036 ± 0.035	0.000	0.067 ± 0.046
	10	0.021 ± 0.021	0.064 ± 0.031	0.074 ± 0.036	0.152 ± 0.053	0.024 ± 0.024	0.071 ± 0.049	0.025 ± 0.025	0.033 ± 0.033
VWA	13	0.000	0.016 ± 0.016	0.000	0.000	0.000	0.000	0.000	0.000
	14	0.167 ± 0.054	0.161 ± 0.047	0.074 ± 0.036	0.043 ± 0.030	0.119 ± 0.050	0.000	0.214 ± 0.063	0.094 ± 0.052
	15	0.125 ± 0.048	0.065 ± 0.031	0.037 ± 0.026	0.130 ± 0.050	0.119 ± 0.050	0.115 ± 0.063	0.095 ± 0.045	0.000
	16	0.208 ± 0.059	0.145 ± 0.045	0.296 ± 0.062	0.217 ± 0.061	0.214 ± 0.063	0.231 ± 0.083	0.167 ± 0.058	0.250 ± 0.077
	17	0.104 ± 0.044	0.226 ± 0.053	0.296 ± 0.062	0.326 ± 0.069	0.262 ± 0.068	0.308 ± 0.091	0.214 ± 0.063	0.219 ± 0.073
	18	0.208 ± 0.057	0.258 ± 0.056	0.204 ± 0.055	0.239 ± 0.063	0.167 ± 0.058	0.231 ± 0.083	0.143 ± 0.054	0.344 ± 0.084
	19	0.146 ± 0.051	0.097 ± 0.038	0.018 ± 0.018	0.022 ± 0.022	0.071 ± 0.040	0.115 ± 0.063	0.119 ± 0.050	0.094 ± 0.052
	20	0.042 ± 0.029	0.032 ± 0.022	0.074 ± 0.036	0.022 ± 0.022	0.047 ± 0.033	0.000	0.048 ± 0.033	0.000
H ± SD		0.779 ± 0.043	0.792 ± 0.021	0.778 ± 0.018	0.788 ± 0.017	0.788 ± 0.027	0.737 ± 0.079	0.732 ± 0.063	0.770 ± 0.015

n, number of chromosomes.
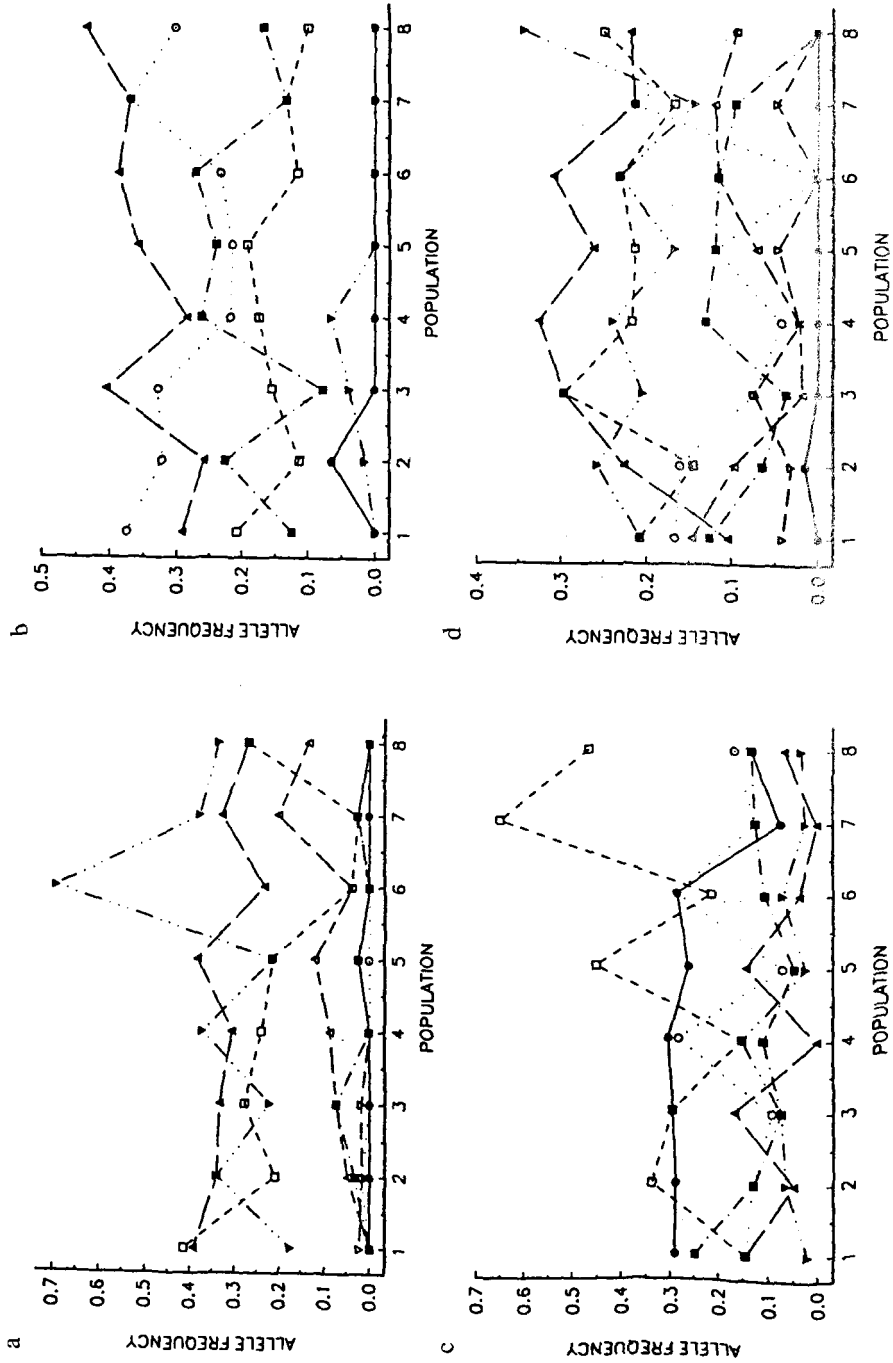
- **Figure 1.** Heterogeneity of allele frequencies at 4 STR loci: (a) *CSF1PO*, (b) *TPOX*, (c) *TH01*, and (d) *VWA*, in 8 population groups of India. (1) Agharia, (2) Bagdi, (3) Brahmin (Uttar Pradesh), (4) Brahmin (West Bengal), (5) Chamar, (6) Gaud, (7) Santal, and (8) Tanti. Symbols used: For loci *CSF1PO* and *TPOX*: allele 7, filled circle; allele 8, open circle; allele 9, filled rectangle; allele 10; open rectangle; allele 11, filled triangle; allele 12, filled inverted triangle; allele 13, open triangle; allele 14, open inverted triangle. For locus *TH01*: allele 6, filled circle; allele 7, open circle; allele 8, filled rectangle; allele 9, open rectangle; allele 9.3, filled triangle; allele 10, filled inverted triangle. For locus *VWA*: allele 13, filled circle; allele 14, open circle; allele 15, filled rectangle; allele 16, open rectangle; allele 17, filled triangle; allele 18, filled inverted triangle; allele 19, open triangle; allele 20, open inverted triangle.

ural selection. However, we are not aware of any selective pressures that may operate on the *TPOX* and *VWA* STR loci.

The level of gene differentiation estimated using data for all 4 STR loci is G_{ST} = 0.041. Thus, when all 4 loci are considered together, most of the observed gene diversity is between individuals within populations; the extent of diversity between populations relative to the total diversity is about 4%. The level of gene differentiation based on these STR loci is lower than those generally observed with respect to protein coding loci in subdivided Indian populations.

The neighbor-joining network depicting affinities among the populations is presented in Figure 2. It can be seen from this figure that, by and large, genetic similarity between the populations correlates with geographic proximity of habitat but not with sociocultural proximity. However, the congruence between genetic affinity is not good with either geographic proximity of habitat or sociocultural proximity. This finding is in contrast with the general observation from Indian populations based on classical genetic markers that the influence of geographic proximity on genetic similarity is stronger than the influence of sociocultural proximity (Majumder 1998).

Table 3. Results of Gene Diversity Analysis Using Data on Allele Frequencies at 4 STR Loci in 8 Ethnic Groups of India

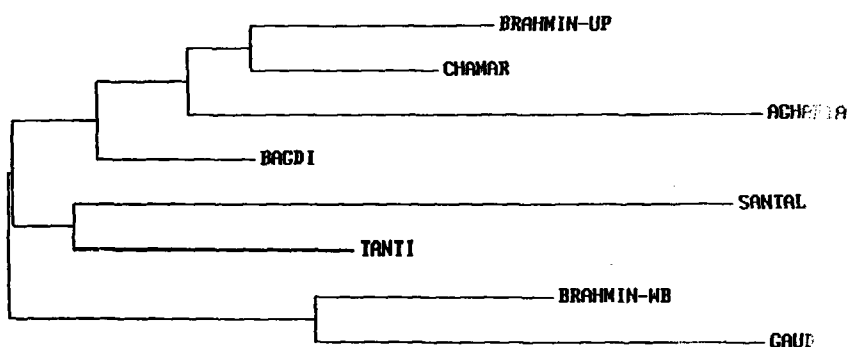| Locus	H_T	H_S	G_{ST}
CSF1PO	0.729	0.685	0.060
TPOX	0.735	0.721	0.019
TH01	0.776	0.728	0.062
VWA	0.816	0.797	0.022
All loci	0.764	0.733	0.041

Figure 2. Neighbor-joining network depicting genomic affinities among 8 population grou,· of India. UP, Uttar Pradesh; WB, West Bengal.

Literature Cited

Deka, R., M.D. Shriver, L.M. Yu et al. 1995. Intra- and interpopulation diversity at short tandem repeat loci in diverse populations of the world. *Electrophoresis* 16:1659–1664.

Edwards, A., H.A. Hammond, L. Jin et al. 1992. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics* 12:241–253.

Gusmao, L., M.J. Prata, A. Amorim et al. 1997. Characterization of four short tandem repeat loci (*TH01, VWA31/A, CD4*, and *TP53*) in northern Portugal. *Hum. Biol.* 69:31–40.

Hammond, H.A., L. Jin, Y. Zhong et al. 1994. Evaluation of 13 short tandem repeat loci for use in personal identification applications. *Am. J. Hum. Genet.* 55:175–189.

Majumder, P.P. 1998. People of India: Biological diversity and affinities. *Evol. Anthropol.* 6:100–110.

Miller, S.A., D.D. Dykes, and H.F. Polesky. 1988. A simple salting-out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* 16:1215.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. USA* 70:3321–3323.

Nei, M., F. Tajima, and Y. Tateno. 1983. Accuracy of estimated phylogenetic trees from mo-lecular data. II. Gene frequency data. *J. Molec. Evol.* 19:153–170.

Pandian, S.K., S. Kumar, K. Dharmalingam et al. 1998. Multilocus DNA profiling using the probe 33.6 in the Tamil Nadu population of South India. *Am. J. Hum. Biol.* 10:87–93.

Perez-Lazaun, A., F. Calafell, E. Mateu et al. 1997. Allele frequencies for 20 microsatellites in a worldwide population survey. *Hum. Hered.* 47:189–196.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molec. Biol. Evol.* 4:406–425.