Congruence of genomic and ethnolinguistic affinities among five tribal populations of Madhya Pradesh (India)

NAMITA MUKHERJEE[1], MITASHREE MITRA[2],
MADAN CHAKRABORTY[1] and PARTHA P. MAJUMDER[1*]

[1]*Anthropology and Human Genetics Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700 035, India*
[2]*School of Studies in Anthropology, Pandit Ravishankar Shukla University, Raipur 492 010, India*

Abstract

The central Indian state of Madhya Pradesh is home to a large number of tribal populations of diverse linguistic and ethnic backgrounds. With a view to examining how well genomic affinities among tribal populations of this state correspond with their ethnic and linguistic affinities, we analysed DNA samples of individuals drawn from five tribes with diverse, but reasonably well-documented, ethnohistorical and linguistic backgrounds. Each DNA sample was scored at 16 biallelic DNA marker loci. On the basis of these data, genomic affinities among these populations were estimated. We have found an extremely good correspondence between the genomic and ethnolinguistic affinities.

[Mukherjee N., Mitra M., Chakraborty M. and Majumder P. P. 2000 Congruence of genomic and ethnolinguistic affinities among five tribal populations of Madhya Pradesh (India). *J. Genet.* **79**, 41–46]

Introduction

Ethnic populations of India are culturally, morphologically, linguistically and genetically very diverse (Majumder 1998). Multiple waves of migration into India during prehistoric and historic times and the subsequent cultural differentiation resulting in strict rules governing mating practices are two of the major causes of the genomic diversity observed among contemporary ethnic groups of India. The tribal populations of India are accepted by anthropologists to be the autochthones. The total number of tribal groups is estimated to be about 450 (Singh 1992).

The central Indian state of Madhya Pradesh (MP) is inhabited by a large number of tribal groups, who are at different stages of modernization. Their occupations today range from hunting and gathering to white-collar jobs. Linguistically, the tribal populations of India speak dialects that belong to one of three language groups: Austro-Asiatic, Dravidian and Tibeto-Burman (Sino-Tibetan). Some large tribal groups (e.g. Bhil) speak a dialect that is classified by many, but not all, as Indo-Aryan (Indo-European). The dialects of the tribal groups of MP represent the major language families present in India. It has been argued (Parpola 1975) that tribals belonging to different language families represent different genetic lineages. Therefore, it is of interest to study the genomic relationships among the tribal groups of this state who speak dialects belonging to different language families.

Materials and methods

Study populations and their ethnohistories: The earliest tribe of MP is the Austro-Asiatic-speaking Baiga tribe, while the Dravidian-speaking Gonds are geographically the most widespread and numerically very large. The Gonds seem to have highly influenced the Baigas. Intermarriages between Gonds and Baigas appear to have been prevalent in historical times, although this practice has now been abandoned (Fuchs 1968). Gonds are said to have migrated from the southern regions of India and some anthropologists consider them as pre-Dravidian (Venkatachar 1935).

*For correspondence. E-mail: ppm@isical.ac.in.

Keywords. DNA polymorphism; ethnic group; language family.

We have studied five ethnic tribal populations from MP: (a) Muria and Halba, sampled in Bastar district, and (b) Kamar, Chinda Bhunjia and Chaukhutia Bhunjia sampled in Raipur district. The Murias are a numerically large ($\approx 100,000$) subtribe of the Gonds, inhabit only the Bastar district, and, as Elwin (1947) has stated, 'are the only civilized group among the other Gond subtribes of the same region'. They are highly endogamous but do not practise close inbreeding. The Kamars are a numerically small tribe ($\approx 13,000$) who primarily inhabit Raipur district and lead a very primitive lifestyle. They are known to be an offshoot of the Gonds (Russel and Hiralal 1916). Inbred marriages (such as a man marrying his father's sister's daughter) are permitted in this tribe. They speak Dravidian mixed with Halbi, a local Indo-European dialect. The Chinda Bhunjia and the Chaukhutia Bhunjia are numerically small (≈ 9500) subtribes of the larger Bhunjia tribe. The Chinda Bhunjias are considered to be an offshoot of the Austro-Asiatic-speaking Baiga, while the Chaukhutias are supposed to have arisen from admixture between the Gonds (maternal) and the Halbas (paternal) (Russel and Hiralal 1916). Both these tribes speak dialects that may be classified in the Indo-European language family (Shukla 1985), which appears to be due to acculturation with the Indo-European-speaking Halba, who number $\approx 60,000$ in Bastar district; the most modernized tribe of MP is Halba. They speak an Indo-European language. They have probably arisen from acculturation between some higher caste groups of Orissa (a neighbouring state) and some tribal people of MP. The Halbas deny any common ancestry with the Gonds.

From the above accounts, it appears that the Murias and Kamars are ethnolinguistically the most close. The Chinda and Chaukhutia are subtribes of Bhunjia and are ethnically close. The Halbas, who speak an Indo-European dialect, are expected to be distinct from these two ethnolinguistic clusters of populations. The objective of this study was to test these ethnolinguistic expectations using genomic data.

Blood samples: Samples of 5–10 ml in EDTA were collected with consent from 220 unrelated individuals belonging to the five tribal groups mentioned above. These samples were transported in ice to the laboratory of the Anthropology and Human Genetics Unit, Indian Statistical Institute, Calcutta, where they were analysed.

Laboratory analysis: High-molecular-weight DNA was isolated from the blood samples by the salting out procedure (Miller *et al.* 1988). Each DNA sample was analysed for polymorphisms at 16 loci, of which nine were insertion/deletion polymorphisms (Indels) and the remaining seven were RFLPs. Primer sequences used for PCR amplification, their corresponding annealing temperatures, and restriction endonuclease digestion protocols are provided in table 1. The reaction mixture for all the amplification reactions contained 50–100 ng DNA , 25 ng of each primer, 200 μM dNTP mix and 1.3 Units of *Taq* DNA polymerase in a total of 10 μl volume. PCR buffer made up of 10 mM Tris.Cl (pH 8.4), 50 mM KCl and 1.5 mM $MgCl_2$ (for two loci T2 and CYP1A, 1.0 mM $MgCl_2$) was used. For *Alu* mtNUC locus PCR cycling temperature protocol was 30 cycles $\times$ (94^0C for 15 s, 63^0C for 1 min, 72^0C for 1 min). For ESR locus the cycling protocol was 30 cycles $\times$ (94^0C for 30 s, 63^0C for 1 min, 72^0C for 1.5 min). For all other loci the cycling protocol was the same except for the annealing temperature, i.e. 30 cycles $\times$ (94^0C for 1 min, x^0C (see table 1 for the value of x) for 1 min, 72^0C for 1 min).

Statistical analysis: Allele frequencies at each of these biallelic loci were estimated for each population by the maximum-likelihood method. Chi-squared tests of significance between the observed genotype frequencies and those expected under Hardy–Weinberg equilibrium were performed. Observed heterozygosities were estimated. The extent of genetic differentiation, G_{ST}, was estimated for individual loci (Nei 1973) and also for the pooled data. Genetic distances between populations were estimated using the D_A distance measure (Nei *et al.* 1983). An unrooted neighbour-joining tree (Saitou and Nei 1987) was constructed to identify affinities among the tribal populations.

Results

Allele frequencies and heterozygosities: Sample sizes and the + allele (insertion allele for the Indel loci and presence of the restriction site for the RFLP loci) frequencies are given in table 2; for the *Alu* CD4 locus, the $-$ allele frequency is presented because the deletion allele is the human-specific allele. All the loci except *Alu* CD4 show high degrees of polymorphism. All populations at most loci show statistically nonsignificant differences of observed genotype frequencies and those expected under Hardy–Weinberg equilibrium (table 2). The heterozygosities at each locus and the average heterozygosities over all the loci for each of the study populations are given in table 3. All the five populations show high levels of diversity at most of the loci. The heterozygosities at the *Alu* CD4 locus are, however, low in all the populations. Diversity at loci *Alu* FX3B and NAT are also comparatively low. However, the average heterozygosities of the populations show considerable variation; 0.388 among Kamars to 0.457 among Chaukhutia Bhujias.

Genomic diversity between populations: Results of gene diversity analysis for individual loci and for all the loci taken together are presented in table 4. The total genomic diversity among the subpopulations is quite high, except that for the *Alu* CD4 locus. It is seen that most of this diversity is due to the diversity between individuals within the same population. This is reflected in the low estimated values of the coefficient of gene differentiation, G_{ST}, which for the pooled data set is 0.025. For four loci, *Alu* FX3B, ESR, LPL and ALB, the G_{ST} values are rather low (< 0.01).

Table 1. Primer sequences, annealing temperatures, and restriction digestion protocols for the loci studied.

Locus	Primer sequence	Annealing temp. (°C)	Restricticton digestion protocol	Reference
Alu mtNUC	5′-ACA AAG TCC AGG TTT CTA ACA G-3′ 5′-AGT CTT GCT TAT TAC AAT GAT GG -3′	63	Not applicable	Zischler *et al.* 1995
Alu ACE	5′-CTG GAG ACC ACT CCC ATC CTT TCT-3′ 5′-GAT GTG GCC ATC ACA TTC GTC AGA T- 3′	58	Not applicable	Stoneking *et al.* 1997
Alu APO	5′-AAG TGC TGT AGG CCA TTT AGA TTA G-3′ 5′-AGT CTT CGA TGA CAG CGT ATA CAG A-3′	50	Not applicable	Stoneking *et al.* 1997
Alu CD4	5′-AGG CCT TGT AGG GTT GGT CTG ATA-3′ 5′-TGC AGC TGC TGA GTG AAA GAA CTG-3′	58	Not applicable	Edwards and Gibbs 1992
Alu D1	5′-TGC TGA TGC CCA GGG TTA GTA AA-3′ 5′-TTT CTG CTA TGC TCT TCC CTC TC-3′	66	Not applicable	Stoneking *et al.* 1997
Alu FX3B	5′-TCA ACT CCA TGA GAT TTT CAG AAG T -3′ 5′-CTG GAA AAA ATG TAT TCA GGT GAG T-3′	56	Not applicable	Stoneking *et al.* 1997
Alu TPA25	5′-GTA AGA GTT CCG TAA CAG GAC AGC T-3′ 5′-CCC CAC CCT AGG AGA ACT TCT CTT T-3′	58	Not applicable	Stoneking *et al.* 1997
Alu PV92	5′-AAC TGG GAA AAT TTG AAG AGA AAG T-3′ 5′-TGA GTT CTC AAC TCC TGT GTG TTA G-3′	54	Not applicable	Stoneking *et al.* 1997
Alu PLAT	5′-GTG AAA AGC AAG GTC TAC CAG-3′ 5′-GAC ACC GAG TTC ATC TTG AC-3′	60	Not applicable	Tishkoff *et al.* 1996
ESR	5′-CTG CCA CCC TAT CTG TAT C-3′ 5′-CTC TGC CAC CCT GGC GTC-3′	63	5 units of *Pvu*II in appropriate buffer was added to the tube, incubated at 37°C for 2 h	Anderson *et al.* 1994
NAT	5′-GAC ATT GAA GCA TAT TTT GAA A-3′ 5′-GAT GAA AGT ATT TGA TGT TTA-3′	56	5 units of *Kpn*I in appropriate buffer was added to the tube, incubated at 37°C for 2 h	Cascorbi *et al.* 1996
CYP1A	5′-CTG ACT GGC TTC AGC AAG TT-3′ 5′-TAG GAG TCT TGT CTC ATG CCT-3′	56	5 units of *Msp*I in appropriate buffer was added to the tube, incubated at 37°C for 2 h	Hayashi *et al.* 1991
PSCR	5′-GGG TTC TAA AGG GAA GAA A-3′ 5′-CCT AAC AGA GGT CAC AAG G-3′	60	5 units of *Taq*I in appropriate buffer was added to the tube, incubated at 65°C for 2 h	Stinissen and Broeckhoven 1991
T2	5′-CTG CAG CTT TTT CTC TAG GG-3′ 5′-CGT CTG CTA CAA GTT CTG GCT T-3′	65	5 units of *Msp*I in appropriate buffer was added to the tube, incubated at 37°C for 2 h	Lynn Jorde (personal communication)
LPL	5′-AGG CTT CAC TCA TCC GTG CCT CC-3′ 5′-TTA TGC TGC TTT AGA CTC TTG TC-3′	62	5 units of *Pvu*II in appropriate buffer was added to the tube, incubated at 37°C for 2 h	Stepanov and Lemza 1993
ALB	5′-GTA GGT GGA CTT GGA GAA GG-3′ 5′-GAT ATA CTT GGC AAG GTC C-3′	63	5 units of *Hae*III in appropriate buffer was added to the tube, incubated at 37°C for 2 h	Lynn Jorde (personal communication)

Genomic affinities among populations: Pairwise genetic distances between the study populations were calculated from the allele frequencies using the D_A distance measure (Nei *et al.* 1983). An unrooted neighbour-joining tree was constructed from this distance matrix, which is depicted in figure 1. It is seen from this figure that the five study populations group themselves in three clusters: {Muria, Kamar}, {Chinda Bhunjia, Chaukhutia Bhunjia} and {Halba}.

Discussion

Human-specific insertion/deletion polymorphisms and the RFLP markers that have been used in this study are known to be selectively neutral in nature. Therefore, observed variations in the allele frequencies among populations are primarily due to random genetic drift or admixture. Since the study populations have generally remained endogamous,

Table 2. Allele frequencies and Hardy–Weinberg chi-square values at 16 biallelic loci in five populations of Madhya Pradesh.

	Population														
	Muria			Halba			Kamar			Chinda Bhunjia			Chaukhutia Bhunhia		
Locus	n	$p(+)$	χ^2	n	$p(+)$	χ^2	n	$p(+)$	χ^2	n	$p(+)$	χ^2	n	$p(+)$	χ^2
Alu mtNUC	49	0.388	0.145	48	0.625	0.592	54	0.463	0.609	25	0.420	0.113	39	0.513	0.027
Alu APO	49	0.714	3.864	47	0.691	5.803	57	0.649	3.008	27	0.481	10.780	37	0.622	0.043
Alu ACE	49	0.531	7.570	48	0.646	0.000	57	0.640	0.876	27	0.796	0.020	37	0.649	0.097
Alu CD4	49	0.010	0.006	48	0.094	0.511	57	0.018	0.019	26	0.038	0.043	39	0.090	0.377
Alu D1	49	0.347	7.210	48	0.427	3.096	57	0.342	9.840	25	0.620	8.280	38	0.289	9.080
Alu FX3B	49	0.786	2.212	48	0.698	0.817	57	0.746	1.392	26	0.731	9.660	38	0.750	0.105
Alu TPA25	49	0.622	1.435	48	0.625	0.592	57	0.614	1.966	27	0.556	1.080	38	0.750	0.125
Alu PV92	49	0.520	2.375	48	0.563	1.131	56	0.554	4.312	27	0.407	1.396	34	0.412	0.294
Alu PLAT	48	0.625	1.113	47	0.574	0.812	57	0.535	0.029	27	0.556	1.080	35	0.729	0.245
ESR	49	0.520	0.980	48	0.563	1.131	57	0.500	0.468	26	0.596	1.021	36	0.611	1.189
NAT	49	0.847	0.882	48	0.750	0.592	57	0.877	1.965	25	0.700	0.510	37	0.581	1.036
CYP1A	47	0.532	2.507	46	0.435	0.728	56	0.598	0.000	25	0.600	0.694	31	0.532	1.650
PSCR	49	0.286	0.490	48	0.365	3.887	57	0.228	0.528	25	0.180	1.205	38	0.263	1.864
T2	49	0.480	0.980	48	0.521	1.043	55	0.655	2.342	25	0.380	0.268	38	0.289	2.975
LPL	49	0.551	0.694	48	0.583	0.157	57	0.640	0.431	27	0.630	0.060	39	0.526	5.880
ALB	49	0.510	1.007	48	0.594	0.304	54	0.454	1.070	26	0.500	0.154	35	0.514	2.330

Table 3. Observed heterozygosities at 16 biallelic loci and pooled heterozygosity in five population groups of Madhya Pradesh.

	Population				
Locus	Muria	Halba	Kamar	Chinda Bhunjia	Chaukhutia Bhunjia
Alu mtNUC	0.449	0.417	0.444	0.520	0.513
Alu APO	0.245	0.277	0.351	0.815	0.486
Alu ACE	0.694	0.458	0.404	0.333	0.432
Alu CD4	0.020	0.188	0.035	0.077	0.179
Alu D1	0.286	0.354	0.263	0.200	0.211
Alu FX3B	0.265	0.479	0.439	0.154	0.447
Alu TPA25	0.551	0.417	0.386	0.593	0.395
Alu PV92	0.429	0.417	0.357	0.593	0.529
Alu PLAT	0.542	0.553	0.509	0.593	0.429
ESR	0.429	0.417	0.544	0.577	0.389
NAT	0.224	0.333	0.175	0.360	0.405
CYP1A	0.383	0.435	0.482	0.400	0.613
PSCR	0.449	0.313	0.386	0.360	0.474
T2	0.429	0.583	0.545	0.520	0.526
LPL	0.531	0.458	0.474	0.444	0.692
ALB	0.571	0.521	0.426	0.538	0.629
Pooled heterozygosity	0.406	0.414	0.388	0.445	0.457

similarities of allele frequency profiles of the populations are a reflection of their common ancestry.

All the five populations in this study are from the central Indian state of Madhya Pradesh. The study populations have been selected from their primary regions of habitat. Muria and Halba have been sampled in Bastar and the other three populations have been sampled in Raipur.

There is significantly greater inter-individual variation within each study population than between the populations. The extent of population differentiation is rather low ($G_{ST} = 0.025$), probably indicating ancestral commonalities of the populations, which are not deep-rooted. The genomic affinities among the study populations indicate that the Muria and Kamar are close to each other. From ethnohistorical accounts it is known that both these groups are descendants of the Dravidian-speaking Gonds. The two subtribes of the Bhunjias, Chinda and Chaukhutia, are also genetically close to each other. The Halbas are genetically distinct from these two clusters of populations. The Halbas are an Indo-European-speaking tribe of MP and do not share any common ancestry with the Gonds. Therefore, it is clear that genomic affinities among these populations of Madhya Pradesh correspond closely with their ethnohistorical and linguistic affinities.

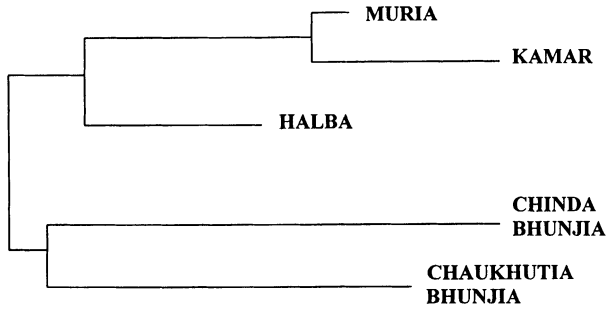

Figure 1. Unrooted neighbour-joining tree depicting genomic relationships among five population groups of Madhya Pradesh.

Table 4. Results of gene diversity analysis for individual loci and for all loci jointly considered.

Locus	H_T	H_S	G_{ST}
Alu mtNUC	0.499	0.486	0.028
Alu APO	0.465	0.452	0.029
Alu ACE	0.454	0.439	0.031
Alu CD4	0.095	0.092	0.027
Alu D1	0.482	0.455	0.056
Alu FX3B	0.383	0.381	0.004
Alu TPA25	0.464	0.456	0.017
Alu PV92	0.500	0.491	0.019
Alu PLAT	0.478	0.469	0.020
ESR	0.493	0.490	0.007
NAT	0.374	0.351	0.061
CYP1A	0.497	0.490	0.015
PSCR	0.389	0.381	0.020
T2	0.498	0.466	0.062
LPL	0.485	0.481	0.008
ALB	0.500	0.496	0.008
All loci	0.441	0.430	0.025

Acknowledgements

We thank the members of the laboratory of the Anthropology and Human Genetics Unit of the Indian Statistical Institute, Calcutta, for help and advice at various stages of this work. This study was partially supported by a grant from the Department of Biotechnology, Government of India, under the 'Genomic Diversity of Indian Populations Project' (GENDIPP).

References

Anderson T. I., Heimdal K. R., Skrede M., Tveit K., Berg K. and Borresen A. L. 1994 Oestrogen receptor (ESR) polymorphism in breast cancer susceptibility. *Hum. Genet.* **94**, 665–670.

Cascorbi I., Brockmoller J., Mrozikiewicz P. M., Baner S., Loddenkemper R. and Roots I. 1996 Homozygous rapid arylamine N-acetyltransferase (NAT2) genotype as a susceptibility factor for lung cancer. *Cancer Res*. **56**, 3961–3966.

Edwards M. C. and Gibbs R. A. 1992 A human dimorphism resulting from loss of an *Alu*. *Genomics* **14**, 590–597.

Elwin V. 1947 *The Muria and their Ghotul*. Oxford University Press, New Delhi.

Fuchs S. 1968 *The Gond and Bhumia of eastern Mandla*. New Literature Publishing House, Bombay.

Hayashi S., Watanabe K., Nakachi K. and Kawajiri K. 1991 Genetic linkage of lung cancer-associated *Msp*I polymorphism with amino acid replacement in the heme binding region of the human cytochrome P450IAI gene. *J. Biochem.* **110**, 407–411.

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

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

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

Nei M., Tajima. F. and Tateno Y. 1983 Accuracy of estimated phylogenetic trees from molecular data II. Gene frequency data. *J. Mol. Evol*. **19**, 153–170.

Parpola A. 1975 On the protohistory of the Indian languages in the light of archeological, linguistic and religious evidence: an attempt at integration. In *South Asian archeology* (ed. J. E. Van Lohuizen-De Leeuw), pp. 73–84. Brill Academic, New York.

Russel R. V. and Hiralal R. B. 1916 *Castes and tribes of Central Provinces of India*. Macmillan, London.

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

Shukla H. L. 1985 *Language, ethnicity and history.* B. R. Publishing, Delhi.

Singh K. S. 1992 *People of India*, vol. 1. Oxford University Press, Delhi.

Stepanov P. and Lemza S. V. 1993 *Pvu*II restriction FLP of lipoprotein lipase in Russians. *Hum. Hered*. **43**, 388–390.

Stinissen P. and Broeckhoven C. V. 1991 PCR detection of the frequent *Taq*I RFLP at the locus D21S13E. *Nucl. Acids Res*. **19**, 2516.

Stoneking M., Fontius J. J., Clifford S. L., Soodyall H., Arcott S. S., Saha N. *et al.* 1997 *Alu* insertion polymorphisms and human evolution: Evidence for a large population size in Africa. *Genome Res*. **7**, 1061–1071.

Tishkoff, S. A., Ruano G., Kidd J. R. and Kidd K. K. 1996 Distribution and frequency of a polymorphic *Alu* insertion at the plasminogen activator locus in humans. *Hum. Genet.* **97**, 759–764.

Venkatachar C. S. 1935 Notes on migration of castes and tribes into central India and their distribution. In *Census of India, 1931* (ed. J. H. Hutton), pp. 60–64. Government of India Press, New Delhi.

Zischler H., Geisert H., von Haeseler A. and Paabo S. 1995 A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern fossils. *Nature* **378**, 489–492.

Received 7 August 2000