ORIGINAL INVESTIGATION

Susanta Roychoudhury · Sangita Roy · Analabha Basu
Rajat Banerjee · H. Vishwanathan · M. V. Usha Rani
Samir K. Sil · Mitashree Mitra · Partha P. Majumder

Genomic structures and population histories of linguistically distinct tribal groups of India

Abstract There are various conflicting hypotheses regarding the origins of the tribal groups of India, who belong to three major language groups – Austro-Asiatic, Dravidian and Tibeto-Burman. To test some of the major hypotheses we designed a genetic study in which we sampled tribal populations belonging to all the three language groups. We used a set of autosomal DNA markers, mtDNA restriction-site polymorphisms (RSPs) and mtDNA hypervariable segment-1 (HVS-1) sequence polymorphisms in this study. Using the unlinked autosomal markers we found that there is a fair correspondence between linguistic and genomic affinities among the Indian tribal groups. We reconstructed mtDNA RSP haplotypes and found that there is extensive haplotype sharing among all tribal populations. However, there is very little sharing of mtDNA HVS-1 sequences across populations, and none across language groups. Haplogroup M is ubiquitous, and the subcluster U2i of haplogroup U occurs in a high frequency. Our analyses of haplogroup and HVS-1 sequence data provides evidence in support of the hypothesis that the Austro-Asiatic speakers are the most ancient inhabitants of India. Our data also support the earlier finding that some of the western Eurasian haplogroups found in India may have been present in India prior to the entry of Aryan speakers. However, we do not find compelling evidence to support the theory that haplogroup M was brought into India on an "out of Africa" wave of migration through a southern exit route from Ethiopia. On the contrary, our data raise the possibility that this haplogroup arose in India and was later carried to East Africa from India.

Introduction

India has served as a major corridor for the dispersal of modern humans out of Africa (Cann 2001). The date of entry of modern humans into India remains uncertain. However, modern human remains dating back to the late Pleistocene (55,000–25,000 years before present, ybp) have been found (Kennedy et al. 1987) and by the Middle Paleolithic Period (50,000–20,000 ybp) humans appear to have spread to many parts of India (Misra 1992). The contemporary people of India are culturally stratified as tribals and non-tribals. It is generally accepted that the tribal people are the original inhabitants of India (Ray 1973; Thapar 1966). The tribals constitute 8.08% of the total population of India (1991 census of India). There are an estimated 461 tribal communities in India (Singh 1992), who speak about 750 dialects (Kosambi 1991) which can be classified into one of the following three language families: Austro-Asiatic, Dravidian and Tibeto-Burman. There is considerable debate about the evolutionary histories of the Indian tribals. The proto-Australoid tribals, who speak dialects belonging to the Austric linguistic group, are believed to be the basic element in the Indian population (Thapar 1966). Many other anthropologists, historians and linguists (Pattanayak 1998; Rapson 1955; Risley 1915) have also supported the view that the Austro-Asiatic (a subfamily of the Austric language family) speaking tribals to be the original inhabitants of India. Some other scholars (Buxton 1925; Sarkar 1958) have, however, proposed that the Dravidians are the original inhabitants, and that the Austro-Asiatics are later immigrants. Many linguists (Renfrew 1987; Ruhlen 1991) contend that Elamo-

S. Roychoudhury · S. Roy · R. Banerjee
Human Genetics and Genomics Department,
Indian Institute of Chemical Biology, Calcutta, India

A. Basu · P.P. Majumder (✉)
Anthropology and Human Genetics Unit,
Indian Statistical Institute, 203 B.T. Road, Calcutta 700 035, India
e-mail: ppm@isical.ac.in,
Tel.: +91-33-5778085 ext 3209, Fax: +91-33-5773049

H. Vishwanathan · M.V. Usha Rani
Department of Environmental Science, Bharathiar University,
Coimbatore, India

S.K. Sil
Department of Life Science, Tripura University, Agartala, India

M. Mitra
School of Studies in Anthropology,
Pandit Ravishankar Shukla University, Raipur, India

Dravidian languages originated in the Elam province of southwestern Iran, and that the dispersal of the Dravidian languages into India took place with the migration of humans from this region, who brought with them the technologies of agriculture and animal domestication. The Tibeto-Burman speaking tribals, who inhabit primarily the northeastern regions of India, are thought to be immigrants to India from Tibet and Myanmar (Guha 1935).

There is hardly any molecular genetic evidence to shed light on these archaeological and linguistic hypotheses. Further, the origin and dispersal of the aboriginal inhabitants of India remain unclear. However, most studies on Indian ethnic populations have used either samples drawn from caste populations or samples of mixed ethnicity. An early study by Mountain et al. (1995) used a small sample (n=7) of Kadars, a tribal population inhabiting southern India. A recent study by Kivisild et al. (1999), which included some tribal samples, addressed the question, using mtDNA polymorphisms, whether the links between Europeans and Indians are due to recent admixture (Indo-Aryan invasion of India) or are more ancient. Another study, using Y-chromosomal DNA markers by Quintana-Murci et al. (2001) supported a demic diffusion model of early farmers from southwestern Iran – and nomads from western and central Asia – into India, bringing the spread of genes and culture (including language) to southwestern Asia. The DNA samples analysed by Quintana-Murci et al. (2001) did not include any specific ethnic population of India; it is unclear whether any tribal samples from India were included. Similarly, the studies by Barnabas et al. (1996, 2000) were also on samples of mixed ethnicity.

Since the tribals are the autochthones of India we considered it more appropriate to study tribal populations. Cavalli-Sforza et al. (1994) have also emphasised that in India "tribals may represent relic populations of unknown origin but potentially great genetic interest." Although we recognise that there are many different modes of evolution/development, acquisition and spread of language (Renfrew 2000), past genetic studies, albeit within restricted global geographical regions, have shown that linguistic proximity is a good predictor of genetic affinity (Barbujani 1997), and that within India "linguistic differences account for much of the genetic diversity" (Cavalli-Sforza et al. 1994). We designed our study to include tribal groups representing all the three linguistic families – Austro-Asiatic, Dravidian and Tibeto-Burman. We generated data using a battery of mitochondrial (mtDNA) and autosomal DNA biallelic polymorphisms and mtDNA hypervariable-segment 1 (HVS1) sequence polymorphisms, and analysed these data statistically to test the concordance of linguistic and genomic similarities among Indian tribals and to draw other relevant population genetic inferences.

Materials and methods

Blood samples (5–10 ml by venipuncture) were collected with informed consent from individuals belonging to eight tribal groups of India representing all the language groups to which the tribal dialects belong. The tribal groups were as follows:

- Austro-Asiatic speakers: Santal (n=21), Munda (n=7), Lodha (n=32)
- Dravidian speakers: Muria (n=49), Kota (n=45), Kurumba (n=54), Irula (n=50)
- Tibeto-Burman speakers: Tipperah (n=51)

These tribal communities inhabit the eastern (Santal, Munda, Lodha), southern (Kota, Kurumba, Irula), central (Muria) and northeastern (Tipperah) regions of India. The locations of sampling of these populations are given in Fig. 1. All the tribal populations have traditionally been hunter-gatherers but now practise shifting and/or settled cultivation. Most groups continue to remain maritally isolated from the other groups, although some degree of acculturation has been reported among the numerically large tribes, for example, Santal. The population sizes of these groups are highly variable from about 1,500 individuals (Kota) to several million (Santal). The numerically large groups inhabit geographically large areas, but because of primitive modes of communication and movement often form local subgroups.

DNA from each blood sample was isolated using a standard protocol (Miller et al. 1988). Each DNA sample was screened for ten mtDNA restriction site polymorphisms (RSPs) and one insertion/deletion polymorphism (IDP). The RSPs screened were *Hae*III nt 663, *Hpa*I nt 3592, *Alu*I nt 5176, *Alu*I nt 7025, *Dde*I nt 10394, *Alu*I nt 10397, *Hinf*I nt 12308, *Hinc*II nt 13259, *Alu*I nt 13262, *Hae*III nt 16517; and the IDP screened was the COII/tRNA^{Lys} intergenic 9-bp deletion. These sites were chosen such that individuals could be classified into haplogroups that are most relevant for Indian populations. mtDNA RSP analyses were performed using standard primers and protocols (Torroni et al. 1993, 1996). The HVS1 region (nt 16033–nt 16370) was amplified using standard primers (Vigilant et al. 1991). Nucleotide sequences were determined directly from purified PCR products using a ABI-377 automated DNA sequencer and the ABI prism dideoxyterminator system. Each sample was also screened for 12 autosomal biallelic polymorphisms, of which 4 are human-specific IDPs and 8 are RSPs. The names and GDB accession numbers of the RSP loci are: ESR1 (GDB: 185229), NAT (GDB: 187676), PSCR (GDB: 182305), T2 (GDB: 196856), LPL (GDB: 285016), ALB (GDB: 178648), *Hox*B4 (GDB: 120663) *Msp*I, ADH2 (GDB: 119651) *Rsa*I. The names of the IDPs are *Alu*FXIIIB, *Alu*CD4, *Alu*PLAT and mtNUC. Primers and protocols used for screening of the IDPs were as given in Majumder et al. (1999) and Tishkoff et al. (1996), and those for RSPs were as given in Jorde et al. (1995) and K. Kidd (personal communication). Because of paucity of DNA and/or PCR amplification failures, sample sizes are variable across loci. Further, lack of sufficient DNA from the Munda population also precluded the screening of autosomal loci on this population. Since the major focus of this study is on mtDNA polymorphisms, details and data on autosomal polymorphisms are not provided here, for brevity. These can be obtained by writing to the authors.

AMOVA analysis of mtDNA haplotypes was performed using Arlequin (Schneider et al. 2000). DNA sequences were aligned using ClustalW. The Cambridge sequence was used as reference during alignment. Descriptive statistics, nucleotide diversities and mismatch statistics were calculated using the DnaSP (version 3) package (Rozas and Rozas 1999). Expansion times were estimated using the method proposed by Slatkin and Hudson (1991), assuming a mutation rate of 20.5% per site per million years, which is appropriate for the HVS1 region (Bonatto and Salzano 1997). The 95% confidence interval of an estimated expansion time was calculated using 2×SD of the sampling variance of nucleotide diversity. Calculation of Fu's (1997) F_S statistic and its test of significance using coalescent simulation were performed using Arlequin (Schneider et al. 2000). Within and between population components of nucleotide diversity were calculated by the method of Nei and Jin (1989). For phylogenetic analyses using mtDNA sequence data, DNA distances were calculated using the maximum likelihood method assuming a 30:1 transition:transversion ratio, which has been suggested as appropriate for the HVS1 region (Lundstrom et al. 1992). For phylogenetic analysis using autosomal and mtDNA haplotype frequency data, genetic distances were calculated using Nei's (1987) D_A distance. Standard deviation of an es-

Fig. 1 Map of India showing locations of sampling of the tribal populations included in the present study
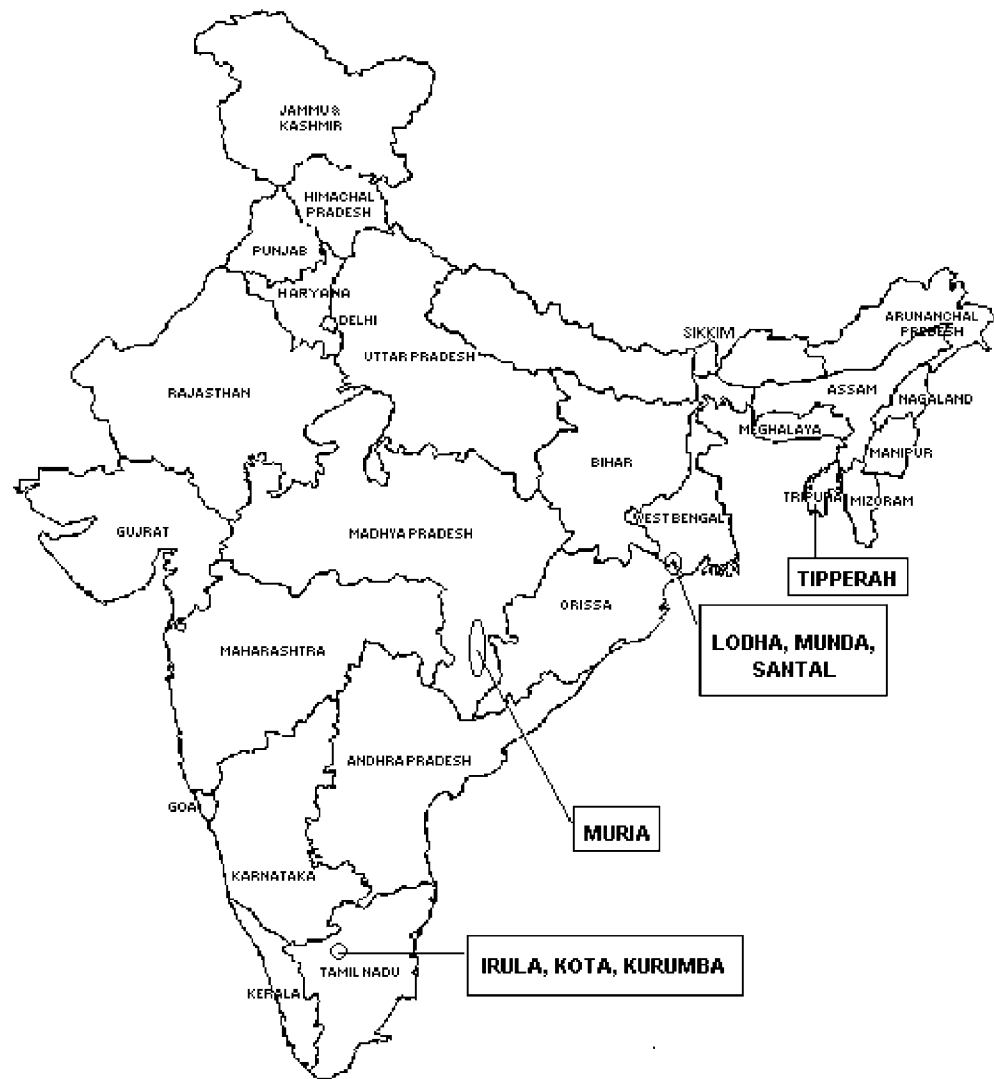
timated value of genetic distance was calculated by resampling. The neighbour-joining method (Saitou and Nei 1987) was used for phylogenetic reconstruction. All phylogenetic calculations were performed using the DISPAN package or the DNADIST (Jukes-Cantor) and NEIGHBOR modules of the PHYLIP version 3.5c package (http://evolution.genetics. washington.edu/phylip.html).

Results

Since mtDNA evolves essentially as a single locus because of lack of recombination, we used data on a set of 12 unlinked autosomal polymorphic markers to examine whether the extent of congruence between linguistic and genomic affinities for the tribal groups included in the present study, except the Munda (for whom the lack of sufficient DNA precluded determination of autosomal genotypes). We calculated pairwise genetic distances between these populations. The estimated distance matrix, with resampling standard deviations, is presented in Table 1. We performed a phylogenetic analysis using this estimated distance matrix; the neighbour-joining tree is presented in Fig. 2. It is seen that tribal groups belonging to

Table 1 Matrix of pairwise genetic distances (standard errors) among tribal populations of India

	Lodha	Santal	Tipperah	Muria	Irula	Kota
Santal	0.0171±0.0069	–	–	–	–	–
Tipperah	0.0518±0.0082	0.0257±0.0010	–	–	–	–
Muria	0.0181±0.0055	0.0097±0.0063	0.0161±0.0061	–	–	–
Irula	0.0109±0.0049	0.0134±0.0070	0.0201±0.0065	0.0050±0.0045	–	–
Kota	0.0343±0.0075	0.0300±0.0092	0.0272±0.0079	0.0147±0.0055	0.0186±0.0061	–
Kurumba	0.0156±0.0055	0.0196±0.0085	0.0289±0.0068	0.0112±0.0053	0.0106±0.0037	0.0185±0.0059

Fig. 2 Neighbour-joining tree constructed on the basis of data on autosomal DNA polymorphisms depicting relationships among tribal populations of India
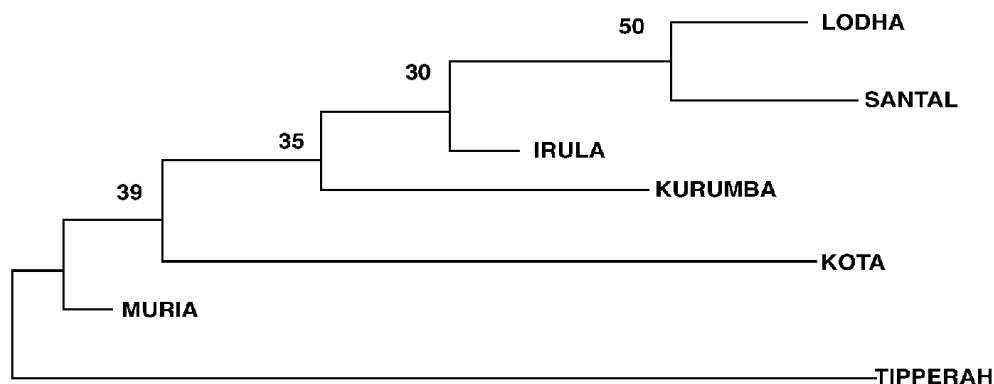

Table 2 mtDNA haplotypes based on ten restriction site and one insertion/deletion polymorphisms in eight tribal populations of India belonging to three language groups. Order of loci: *Hae*III nt 663, *Hpa*I nt 3592, *Alu*I nt 5176, *Alu*I nt 7025, *Dde*I nt 10394, *Alu*I nt 10397, *Hin*fI nt 12308, *Hin*cII nt 13259, *Alu*I nt 13262, *Hae*III nt 16517, 9-bp deletion (*1* presence of restriction site, *0* absence of restriction site)

Haplotype		Austro-Asiatic			Dravidian				Tibeto-Burman	
Sl. no.	Type	Lodha (*n*=32)	Munda (*n*=7)	Santal (*n*=20)	Irula (*n*=30)	Kota (*n*=30)	Kurumba (*n*=30)	Muria (*n*=30)	Tipperah (*n*=45)	Total (*n*=224)
1	00111101010	15 46.87	3 42.85	12 60.0	14 46.66	29 96.66	23 76.66	23 76.66	14 31.11	133 59.37
2	00110001010	– –	2 28.57	3 15.0	7 23.33	1 3.33	4 13.33	4 13.33	11 24.44	32 14.28
3	00111101000	11 34.37	2 28.57	2 10.0	1 3.33	– –	– –	1 3.33	4 8.89	21 9.37
4	00110011010	6 18.75	– –	– –	1 3.33	– –	2 6.66	1 3.33	2 4.44	12 5.35
5	00110011000	– –	– –	2 10.0	6 20.0	– –	– –	1 3.33	1 2.22	10 4.46
6	10110001000	– –	– –	– –	– –	– –	– –	– –	4 8.89	4 1.78
7	00111100110	– –	– –	– –	1 3.33	– –	– –	– –	1 2.22	2 0.89
8	00011101010	– –	– –	– –	– –	– –	– –	– –	2 4.44	2 0.89
9	00010000100	– –	– –	– –	– –	– –	– –	– –	2 4.44	2 0.89
10	00110001000	– –	– –	1 5.0	– –	– –	– –	– –	1 2.22	2 0.89
11	00011101000	– –	– –	– –	– –	– –	– –	– –	2 4.44	2 0.89
12	00110000110	– –	– –	– –	– –	– –	– –	– –	1 2.22	1 0.44
13	10110001010	– –	– –	– –	– –	– –	1 3.33	– –	– –	1 0.44

the same language group form a cluster, although somewhat loose. The Austro-Asiatic tribal groups (Lodha, Santal) cluster together, the Tibeto-Burman speaking Tipperah forms a separate cluster, and the Dravidian tribal groups (Irula, Kurumba, Kota, Muria), although not forming a tight cluster, are positioned intermediate between the Austro-Asiatic and Tibeto-Burman clusters. One reason for the lack of tightness of the clusters and moderate bootstrap support values of the branches (Fig. 2) may be that the number of loci used (12) is not adequate. In spite of these observations our results indicate that the use of language as a classificatory variable in understanding the genomic structure of tribal populations of India appears largely justified. It is, however, worth pointing out that the cluster configurations can also be explained by geographical, instead of linguistic, contiguity of the populations. Unfortunately, the distribution of the tribal groups in India is such that there is considerable confounding of geography and language. The Dravidian speaking tribals are found only in restricted areas of southern India, the Tibeto-Burman speaking tribals are confined to the northeast, and the Austro-Asiatics are restricted to small pockets in the central and eastern regions.

With respect to the 11 biallelic mtDNA polymorphisms, 13 distinct haplotypes were observed among the 224 individuals screened from the eight populations (Table 2).
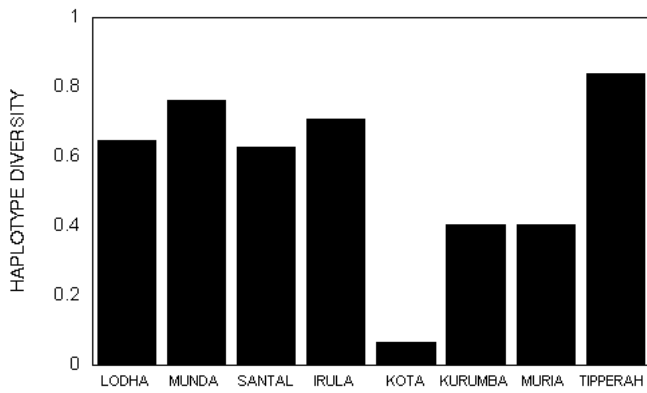
Fig. 3 mtDNA haplotype diversities among tribal populations of India

None of the individuals possessed the COII/tRNA[Lys] intergenic 9-bp deletion. Two of the ten restriction sites (*Hpa*I nt 3592 and *Alu*I nt 7025) were also monomorphic. It was found that only 4 of the 13 distinct haplotypes were present in many populations; the remaining nine haplotypes occurred sporadically. The Tibeto-Burman speaking Tipperahs harboured the maximum number (12 of 13) of haplotypes. In the pooled sample only one haplotype (haplotype 1 in Table 2) was present in about 60% of the individuals. The frequencies of this haplotype differed significantly between the three language groups ($P<0.05$; Austro-Asiatic 50.85%, Dravidian 73.33%, Tibeto-Burman 31.11%). Several haplotypes were present in small frequencies only among the Tibeto-Burmans. The haplotype diversities of the populations are presented in Fig. 3. The Tibeto-Burman speaking Tipperah, in view of the fact that they harbour a large number of haplotypes at small frequencies, exhibit the highest haplotype diversity. The Dravidian speaking Kota exhibit the lowest diversity. In general, the Austro-Asiatic tribals are more diverse than the Dravidian tribals. We carried out AMOVA analysis using haplotype frequencies. It was found that 89.73% of the variation in haplotype frequencies is between individuals within populations, 5.37% is between populations within language groups and only 4.89% is between language groups. The coefficient of gene differentiation, G_{ST}, between the eight populations was 0.126.

Our RSP data permitted the classification of individuals into eight haplogroups: Asian and Amerindian haplogroups A, B, C and D; East Asian haplogroup M; Caucasian haplogroups U and H; and African haplogroup L. The frequencies of the various haplogroups are presented in Table 3. Haplogroups B, H and L were not observed. Thirty-seven (16.5%) individuals could not be classified into any of the eight haplogroups based on the RSP sites examined by us. Haplogroup M was found to be the most frequent – 71.4% of the individuals in the pooled sample belonged to this haplogroup. The frequency (51.11%) of this haplogroup was found to be significantly lower among Tibeto-Burman tribals than among the Austro-Asiatic (76.27%) and the Dravidian (76.66%). Of the remaining haplogroups observed in the study populations, haplogroup U was also found to occur in most populations. The frequency of this haplogroup in the pooled sample was about 10%. Considerable differences in the frequencies of this haplogroup were observed among Austro-Asiatic (13.56%), Dravidian (9.17%) and Tibeto-Burman (6.7%) tribals; these differences were, however, not statistically significant at the 5% level.

The stretch of 338 nucleotides of the HVS1 segment showed deletions of one or two nucleotides, at positions 16182 and 16183, in several individuals. These two nucleotides were not considered in the statistical analyses of the sequence data. Among the 115 individuals there were a total of 104 mutations at 94 polymorphic sites. Details are provided in Table 4. Twelve sequences were shared by at least two individuals. Sharing of sequences was primarily between individuals within the same population, although three of these sequences were shared among individuals belonging to different populations. No sequence was shared between populations belonging to different language groups. Our AMOVA results indicated that although with respect to RSPs, about 5% of the haplotype variation was attributable to differences between the three language groups, observed HVS1 sequence variation could not be attributed to differences between these three groups. As with the variation in RSP haplotypes, about 90% of the HVS1 sequence variation was attributable to differences among individuals within tribal groups, and the remaining 10% of the variation to differences among tribal groups within language groups. The Austro-Asiatic speaking trib-

Table 3 Haplogroup frequencies (percentages) among eight tribal populations of India

[a]Haplogroups B, H and L were not observed in the study samples
[b]Haplogroups C and D are subsets of haplogroup M; therefore individuals belonging to haplogroups C and D are also counted as belonging to haplogroup M

Population	Geographical region of sampling	Haplogroup frequency[a]					
		A	C[b]	D[b]	M	U	Other
Lodha	East	–	–	–	26 (81.3)	6 (18.7)	–
Munda	East	–	–	–	5 (71.4)	–	2 (28.6)
Santal	East	–	–	–	14 (70.0)	2 (10.0)	4 (20.0)
Irula	South	–	1 (3.3)	–	16 (53.3)	7 (23.3)	7 (23.3)
Kota	South	–	–	–	29 (96.7)	–	1 (3.3)
Kurumba	South	1 (3.3)	–	–	23 (76.7)	2 (6.7)	4 (13.3)
Muria	Central	–	–	–	24 (80.0)	2 (6.7)	4 (13.3)
Tipperah	Northeast	4 (8.9)	1 (2.2)	4 (8.9)	23 (51.1)	3 (6.7)	15 (33.3)
Pooled		5 (2.2)	2 (0.9)	4 (1.8)	160 (71.4)	22 (9.8)	37 (16.5)

Table 4 Details of polymorphisms observed in the mtDNA HVS1 region of the 115 Indian tribals along with their common haplogroup designations

Table 4 (continued)

Sample name**	HVS1 sequence polymorphisms*	Haplogroup M	Haplogroup U
IR-01[a]	51,189,234,247	–	+
IR-02[b]	93,266,304,309,325,356	–	–
IR-03	129,362	–	–
IR-04[a]	51,189,234,247	–	+
IR-05	51,144(A),189,234,247,368(A)	–	+
IR-06	172,223,239,298,327,357	+	–
IR-07	124(G),172,278	–	–
IR-08[a]	51,189,234,247	–	+
IR-09	124,179,183,189,223,294	+	–
IR-10	52(G),085,179,183,189,223,294	+	–
IR-11	64,085,179,183,189,223,294	+	–
IR-12	193,223	+	–
IR-13[a]	51,189,234,247	–	+
IR-15[a]	51,189,234,247	–	+
KT-01	223,368	+	–
KT-02	126,223,311	+	–
KT-03	172,223	+	–
KT-04[c]	126,223	+	–
KT-05[c]	126,223	+	–
KT-06[b]	93,266,304,309,325,356	–	–
KT-07[c]	126,223	+	–
KT-08[c]	126,223	+	–
KT-09[c]	126,223	+	–
KT-10	40,126,223,311	+	–
KT-11[d]	39(T),126,223	+	–
KT-12[d]	39(T),126,223	+	–
KT-13	39(T),126,223,330	+	–
KT-14	223,368	+	–
KT-15	39(T),064(A),223	+	–
KT-16[e]	39(T),223,368	+	–
KT-17	39(T),172,223	+	–
KT-18[e]	39(T),223,368	+	–
KT-19	39(T),223	+	–
KT-20	126,223	+	–
KT-21[f]	223	+	–
KT-22	126,223,342	+	–
KT-23[f]	223	+	–
KT-24[f]	223	+	–
KT-25[f]	223	+	–
KR-01[f]	223	+	–
KR-02[f]	223	+	–
KR-03	94(G),223,358(G)	+	–
KR-04	126(G),127(G),129,223	+	–
KR-06	140,183(del),189,223,274,295,(T)319,320	+	–
KR-07[g]	182(del),183(del),189,249	–	+
KR-08	39(C),129	–	–
KR-09	140,183(del),189,223,274,295,319,320	+	–
KR-10[g]	182(del),183(del),189,249	–	+
KR-12	129,250,360	–	–
LO-01	51,93,126,145,223,250,360,362(G)	+	–
LO-02[i]	51,92,168,291,311	–	+
LO-03	51,93(A),183(del),189,223,295	+	–
LO-04	51,183(del),189,223,295	+	–
LO-05[i]	51,92,168,291,311	–	+
LO-06	51,92,168,291,311,358,360,362(G),363(G),364	–	+
LO-07[h]	93,126,145,223	+	–
LO-08[i]	51,92,168,291,311	–	+
LO-09	39(C),085,223,318(T)	+	–
LO-10	147(G),172,223,319	+	–
LO-12[j]	129,218,223	+	–
LO-13	86,108,129,223,278	+	–
LO-14[h]	93,126,145,223	+	–
LO-15	207,309,318(T)	–	+
MU-01	223,272,358(G)	+	–
MU-02	129,266,290,362(G),363	–	–
MU-03[k]	51,189,223,316	+	–
MU-05[k]	51,189,223,316	+	–
MU-06	75,86,189,223,270.274,319,352,362	+	–
MU-07	92,179,223,289,294,319	+	–
MR-01	223,278	+	–
MR-02	129,144(A),223,362	+	–
MR-03	153,266,304	–	–
MR-04	223,362	+	–
MR-05	223,270,319,352	+	–
MR-06	223,284,311	+	–
MR-07	223,304,362	+	–
MR-08	94,260,261,311,319,362	–	–
MR-09	153,266,304	–	–
MR-10	51,209,239,358,360,362(G),363	–	+
MR-11	129,223,311	+	–
MR-12	39(C),45,51,206(C),230,264,311,321(A),353	–	+
SA-01	51,182(del),183(del),189,223,325	+	–
SA-02	48,93,129,223	+	–
SA-03	129,223,294	+	–
SA-04	51,168,215	–	+
SA-05	223,318(T)	+	–
SA-06[j]	129,218,223	+	–
SA-07	86,223,270,274,319,352	+	–
SA-08	51,167,189,223	+	–
SA-09	189,223,300	+	–
SA-10	39(T),129,145,185,239,325	–	–
SA-11	45,92,145,185,239,325	–	–
SA-12	51,93(A),154,206(C),230,311	–	+
SA-13	75,260,261,319,362	–	–
SA-14	92,145,185,239,325	–	–
TR-01	104,223,234,362	+	–
TR-02	93,164,172,182(del),183(del),189,223,266	–	–
TR-03	92,187,353	–	–
TR-04	51,206(C)231(G),234,304,311,362	–	+
TR-05	223,278,284,319	+	–
TR-06	223,298,311,319,357	+	–
TR-07	51,168,172,327,359	–	+
TR-08	64(A),223,320	+	–
TR-09	85,86,223,290,293(C),319	–	–
TR-10	192,288,304,309	–	–
TR-11	111(G),162,172,304	–	–
TR-12	183(del),189,223,241,290,319	–	–
TR-13	183(del),189,304,311	–	–
TR-14	189,197,284,304,362	–	–
TR-15	223,234,300,311	+	–
TR-16[l]	189,284,304,362	–	–
TR-17[l]	189,284,304,362	–	–
TR-18	51,131,318	–	+
TR-19	223,362	+	–
TR-20	39(T),223,309,320	+	–

* Nucleotide positions (minus 16,000) at which mutations – transitions and transversions (*parentheses* nucleotide change) – were noted relative to the Cambridge Reference Sequence

** *Superscripts* shared sequences

Table 5 Descriptive statistics and estimated expansion times of Indian tribals belonging to various language groups and haplo-groups

	Linguistic group			Haplogroup M
	Austro-Asiatic	Dravidian	Tibeto-Burman	
No. of sequences	34	61	20	71
No. of polymorphic sites	54	59	42	59
No. of mutations	59	66	42	63
Nucleotide diversity (π)±2 SD	0.023±0.002	0.016±0.002	0.021±0.001	0.013±0.001
Mean no. of mismatches (k)	7.747	5.432	7.170	4.632
Raggedness (r)	0.0157	0.0058	0.0281	0.0130
F_S (P)	–19.176 (0.000)	–25.417 (0.000)	–12.203 (0.000)	–25.678 (0.000)
Expansion time in years before present (95% confidence intervals)	56,098 (51,220–60,975)	39,024 (34,146–43,902)	51,220 (48,780–53,659)	31,707 (29,268–34,146)

als exhibited the maximum amount (0.023; Table 5) of nucleotide diversity in the HVS1 region, and the Dravidians the lowest (0.016). There was no clear clustering of HVS1 sequences by population or even language group (Fig. 4), which is not unexpected from the results of the AMOVA analysis.

Quintana-Murci et al. (1999) discovered that HVS1 motif defined by four transitions at nt 16129, 16189, 16249 and 16311 characterised haplogroup M (clade M1) in East Africa. This motif has not been found in the tribal populations of India, although all these four transitions are individually and in various pairwise combinations found in our data at varying frequencies (Table 4). The frequencies of these four transitions in the pooled set of samples were 23.5% (nt 16189), 13.0% (nt 16311), 11.3% (nt 16129) and 1.7% (nt 16249). Kivisild et al. (1999) found that there are several subclusters of haplogroup U, of which they found six to be present in their sample of Indians. We found only two of these subclusters to be present among the tribals in India. These are subclusters U2i and U1, with frequencies of 77.3% and 9.1%, respectively. Interestingly, we found that all the six Irulas who belonged to haplogroup U also possessed transitions at nucleotide positions 16051, 16189, 16234 and 16247. This association was not found in any other tribal belonging to haplogroup U.

From these sequence data we calculated nucleotide diversities and statistics (mean number of mismatches and raggedness) pertaining to mismatch distributions. The Austro-Asiatics exhibited the maximum genomic diversity (Table 5), both nucleotide diversity and mean number of mismatches. We estimated the parameters of a population expansion model (Harpending et al. 1993) and examined the fits of the observed and expected mismatch distributions for the three linguistic groups of tribals. The observed and expected distributions are presented in Fig. 5a–c. From the unimodalities of the observed mismatch distributions, smoothness of the distributions (as revealed by the very small values of the raggedness statistic; Harpending et al. 1993; Table 5) and their reasonably good fits with the expected distributions, it is clear that there were significant expansions of these linguistic

groups of tribals. To detect traces of population expansions we also used a second approach. We computed Fu's (1997) F_S statistic, which is particularly sensitive to population growth. Significantly large negative values indicate population expansion (Fu 1997), which is what is observed (Table 5) in our data set for each of the three language groups. We estimated the expansion times, which are also presented in Table 5. The estimated expansion time of the Austro-Asiatics (approx. 56,000 ybp) is much higher than that of the Dravidians (approx. 39,000 ybp). This difference in expansion times is significantly different, as indicated by the disjoint 95% confidence intervals of these estimates. Our tentative estimate, in view of the limited sample size, of the expansion time of the Tibeto-Burmans is approx. 52,000 ybp.

In the set of 71 individuals (22 Austro-Asiatic, 42 Dravidian, 7 Tibeto-Burman) who belonged to haplogroup M the HVS1 nucleotide diversity was 0.013 (Table 5). Among them also, the Austro-Asiatic speakers showed the maximum nucleotide diversity, followed by the Tibeto-Burmans and Dravidians, respectively. The mismatch distribution among individuals of haplogroup M (Fig. 5d) and the relevant statistics (Table 5) also clearly show that haplogroup M underwent a significant population expansion. The estimated expansion time for this haplogroup from our data is approx. 32,000 ybp.

Discussion

Although anthropologists, archaeologists and historians accept that the tribal populations are the original inhabitants of India, most studies on Indian populations using DNA markers have not included the tribals. It has been argued (Pattanayak 1998; Risley 1915; Thapar 1966) that the Austro-Asiatic speaking tribals are the original inhabitants of India. Some other scholars have, however, argued that tribal groups speaking Dravidian and Austro-Asiatic languages have evolved from an older original substrate of proto-Australoids (Keith 1936), while the Tibeto-Burman speaking tribals are later immigrants from Tibet and Myanmar (Guha 1935). Parpola (1975) has con-
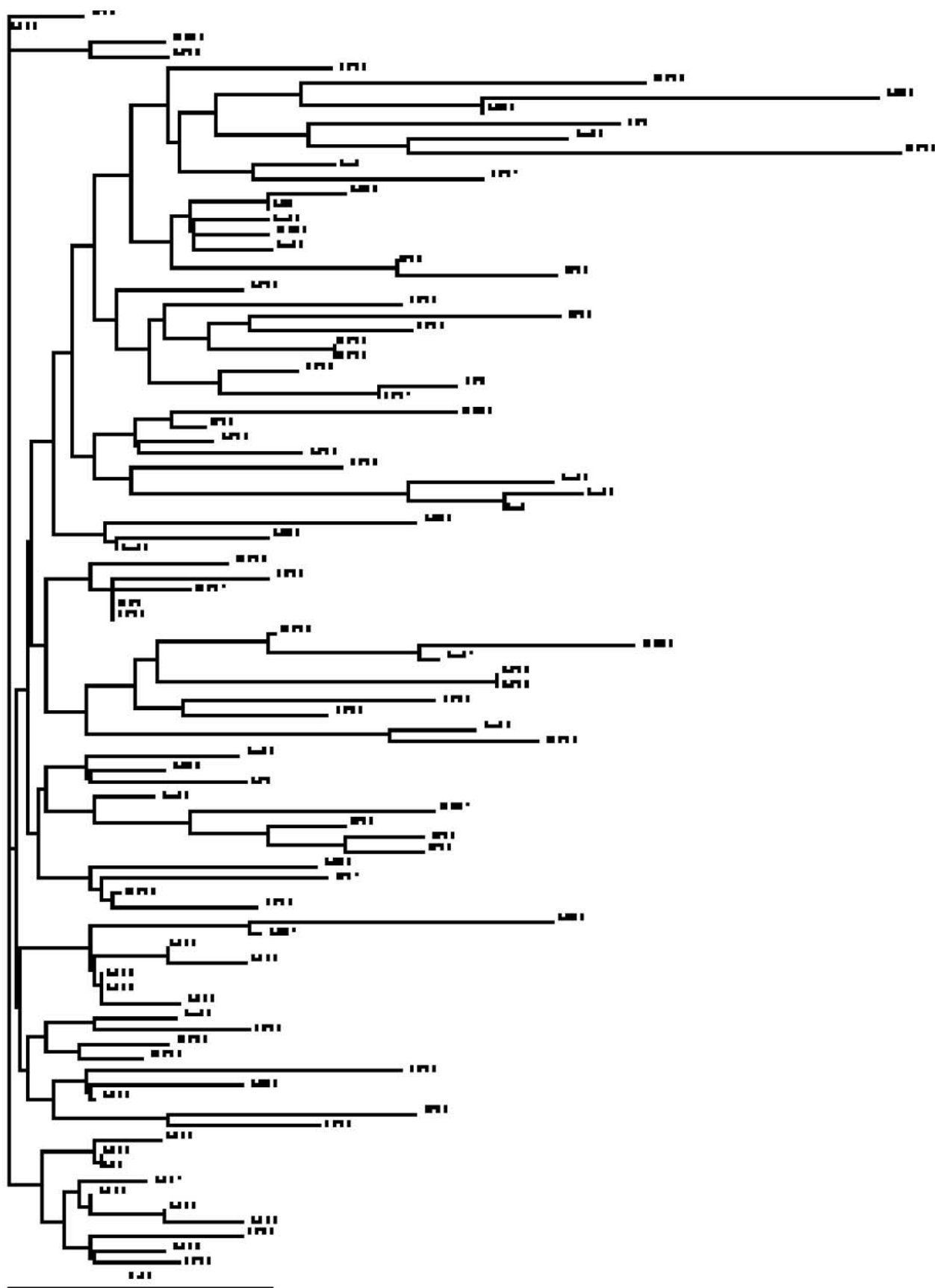
Fig. 4 Neighbour-joining tree depicting relationships among distinct mtDNA HVS-1 sequences observed in tribal populations of India
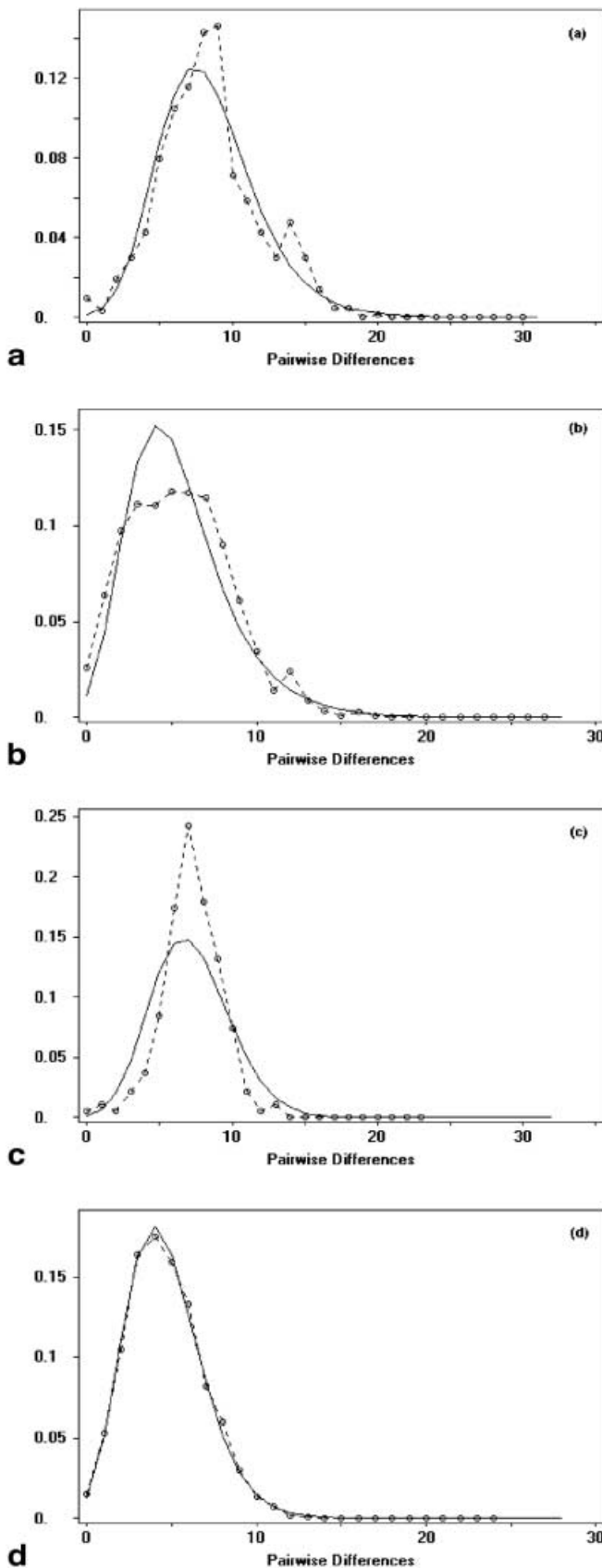
tended that the different language families in India may represent different lineages, which is consistent with the finding by Cavalli-Sforza et al. (1994) that, within India, linguistic differences account for much of the genetic diversity. The present study was undertaken to shed light on these various conflicting hypotheses regarding the peopling of India.

Our study was based primarily on mtDNA polymorphisms. However, since the mtDNA evolves essentially as a single locus, we first examined the extent of evidence in support of a linguistic grouping of Indian tribals. Using data on a set of 12 autosomal biallelic DNA polymorphisms generated by us, we found (Fig. 2) that populations belonging to the same language group are also, by and large, genetically closer. The result is consistent with earlier results based on biochemical genetic markers (Majumder 1998). Thus there appeared to be justification in grouping tribal populations of India by language, although we recognise that a language group may not necessarily represent an evolutionary unit because the modalities of language change may be vastly different from those of genetic change (Renfrew 2000). Further, the geographical distribution of the tribal groups in India is such that there is considerable confounding of geography and language. Therefore the observed genetic clustering of linguistically similar tribal groups may also be due in part to their geographical contiguity of habitat.

The 9-bp deletion on mtDNA that occurs frequently between nt 8272 and nt 8279 in many Polynesian and Native American populations was not found in our study populations. Earlier studies (Barnabas et al. 2000; Clark et al. 2000; Watkins et al. 1999) have also indicated that this 9-bp deletion is extremely infrequent in India. As observed by us earlier (Roychoudhury et al. 2000), there is extensive sharing of haplotypes across ethnic groups; one haplotype (00111101010) was modal across all the populations. Most populations harbour a small number of haplotypes. Several haplotypes are present in small frequencies only among the Tibeto-Burmans; this may be indicative of their recent admixture with other, possibly Southeast Asian, populations. From our estimates of mtDNA haplotype diversities (Fig. 3) it is also clear that drift effects have accentuated the process of genetic differentiation of the Indian tribals. The fact that Austro-Asiatic speaking tribals show greater mtDNA haplotype diversity may be indicative of their greater antiquity.

The presence of mtDNA restriction sites *Dde*I nt 10394 and *Alu*I nt 10397 defines the haplogroup M. This haplogroup was originally identified in Southeast Asian populations (Ballinger et al. 1992; Chen et al. 1995) but were later shown to be an ancient marker in India, predating the migration of Indo-European speakers into India (Passarino et al. 1996a, 1996b). Consistent with our ear-

lier study (Roychoudhury et al. 2000), all tribal populations are found to possess this haplogroup in high frequencies (51.1%–96.7%). The frequencies among Austro-Asiatic and Dravidian tribals are virtually the same (76%), while it is significantly lower among the Tibeto-Burmans (51.1%). Haplogroup M is also found in fairly high frequencies (18%) in the East African population of Ethiopia (Quintana-Murci et al. 1999). Quintana-Murci et al. (1999) also found that the HVS1 motif defined by four transitions at nt 16129, 16189, 16249 and 16311 that characterises haplogroup M (clade M1) in East Africa is not found in India. They estimated using restriction fragment length polymorphism data that the age of the East African haplogroup M is 48,000±15,000 years, and that that of the Indian haplogroup M is 56,000±7,000 years, whereas lower age estmates were obtained using HVS1 data (36,000±11,000 years for East Africans and 53,000±7,000 years for Indians). These facts, they contend, provide support to Lahr and Foley's (1994) hypothesis of a southern exit route of modern humans out of Africa; east Africa to India. However, our present data show (Table 4) that all these four transitions are individually and in various pairwise combinations found among Indian tribals at varying frequencies. Thus, the possibility of haplogroup M having originated in India and carried towards eastern Africa cannot be completely ruled out. Although it has not yet been found in India, the M1 clade may be present in small frequency in India and may have risen to a high frequency in eastern Africa by genetic drift. Alternatively, the haplogroup M may be an assemblage of lineages, some of which may have evolved independently. We have also shown that the subgroup of individuals belonging to haplogroup M expanded (Fig. 5 (d)) and that the estimated expansion time is 32,000±3,000 ybp (Table 5), which nearly coincides with the estimated age of this haplogroup in eastern Africa based on

HVS-1 data (Quintana-Murci et al. 1999). If indeed haplogroup M expanded in eastern Africa, one would have expected that this haplogroup would be more widely dispersed in that region. On the contrary, our past (Roychoudhury et al. 2000) and the present data show that haplogroup M is ubiquitous throughout India. Further, there is no significant cline of the haplogroup M frequency from southern India to northern India (Roychoudhury et al. 2000 and present data), which is what would have been expected if indeed this haplogroup was carried into India from Ethiopia, even if as a very early dispersal event of modern humans. It is therefore plausible that haplogroup M arose in India and was carried to Ethiopia, possibly via the sea route, when there was an expansion of this haplogroup in India about 32,000 ybp. Indeed, the data of Passarino et al. (1996a, 1996b) indicate that this haplogroup is very infrequent in sub-Saharan Africa and the Fertile Crescent region, which was the main exit corridor of modern humans, including to India.

It is remarkable that our estimate (77.3%) of the proportion of tribals belonging to Indian-specific subcluster U2i of haplogroup U coincided with that (77.9%) estimated earlier by Kivisild et al. (1999) based on samples

primarily from caste populations. Because the antiquities of the tribal populations are far greater than the time of entry (3000–4000 ybp) of Indo-Aryan speakers in India, our data support the conclusion reached by Kivisild et al. (1999) that haplogroup U was introduced in India by an ancestral population that preceded the arrival of Indo-Aryan speakers into India. However, while Kivisild et al. (1999) found several western Eurasian mtDNA lineages belonging to haplogroup H and subclusters U1, K, U4, U5 with frequencies between 1% and 5% in their samples from India, we found presence of only the subcluster U1 in our tribal samples with a frequency of 9%. The subcluster U7, found at a frequency of about 13% in the samples of Kivisild et al. (1999) but not found in our tribal samples, may also be western Eurasian. Since the samples included by Kivisild et al. (1999) were obtained primarily from Indo-Aryan speaking caste populations, it is possible that these western Eurasian specific haplogroups and subclusters, except U1, which are not found among the tribals in India, were introduced into India with the entry of Aryan speakers from western and central Asia. This is contrary to the suggestion of Kivisild et al. (1999) that all of the western Eurasian subclusters of haplogroup U were present in India before the entry of the Aryan speakers.

Based on both mtDNA haplotype and HVS1 sequence data, we find that the Austro-Asiatic tribals show a higher diversity than Dravidian tribals. This is consistent with Renfrew's (1992) observation that the present distribution of the Austric language group is due to the initial dispersal process out of Africa, whereas later agricultural dispersal can account for the Elamo-Dravidian or Sino-Tibetan (to which family Tibeto-Burman languages belong) distributions. Our observation is also consistent with the view of many scholars (Pattanayak 1998; Risley 1915; Rapson 1955; Thapar 1966) who have hypothesised that the Austro-Asiatics in India may have been the original inhabitants. Indeed, if the Austro-Asiatic speaking tribals are the most ancient group of humans in India, they are expected to show the highest genetic diversity. Of course, it is possible that this group has descended from a group of founders after the Dravidian or the Tibeto-Burman speakers, but the founding group had a larger effective population size.

We found that there is no sharing of mtDNA HVS1 sequences across language groups. This is also consistent with the hypothesis (Parpola 1975) that populations belonging to different language groups in India represent different genetic lineages. Why do we then not observe distinct clusters of HVS1 sequences by language groups, and why is there so much sharing of mtDNA RSP haplotypes? While this can be variously answered, the simplest explanations are that the mtDNA RSP haplotypes are strongly shared across tribal groups because the haplotypes are based on a small number of polymorphic sites, or that the haplotypes are derivatives of some ancient African haplotypes, or both. On the other hand, because of the much higher mutation rate of the HVS-1 region, ancient signatures may have been lost even before language differentiation and the entry of the people belonging to

various language groups into India. Admixture effects may have accentuated the process. We, however, note that it has been reported that the Indian mtDNA HVS-1 lineages all derived from the African mtDNA lineage cluster L3a (Kivisild et al. 1999; Watson et al. 1997).

Our data indicate Austro-Asiatic speakers underwent population expansion about 17,000 years prior to the Elamo-Dravidian speakers and about 5,000 years prior to the Tibeto-Burman speakers. The confidence intervals of the expansion times of Austro-Asiatic and Tibeto-Burman speakers are non-overlapping with those of the Dravidian speakers, while those of Austro-Asiatic and Tibeto-Burman speakers do overlap, indicating that the antiquity of expansion of the Austro-Asiatics is significantly greater than that of the Dravidians, but not of the Tibeto-Burmans. These data do not provide any evidence that the expansions took place within India. However, among hunter-gatherers, in particular, population expansions led to enormous pressures on natural resources, which result in population movements. Therefore it is probable that the Austro-Asiatic speakers who expanded earlier also migrated earlier. Thus our data provide support to the theories that different language groups in India represent distinct founding groups, and that the Austro-Asiatic speakers are likely to have been the most ancient inhabitants of India.

References

Ballinger SW, Schurr TG, Torroni A, Gan YY, Hodge JA, Hassan K, Chen K-H, Wallace DC (1992) Southeast Asian mitochondrial DNA analysis reveals continuity of ancient mongoloid migrations. Genetics 130:139–152

Barbujani G (1997) DNA variation and language affinities. Am J Hum Genet 61:1011–1014

Barnabas S, Apte RV, Suresh CG (1996) Ancestry and interrelationships of the Indians and their relationship with other world populations: a study based on mitochondrial DNA polymorphisms. Ann Hum Genet 60:409–422

Barnabas S, Joshi B, Suresh CG (2000) Indian-Asian relationship: mtDNA reveals more. Naturwissenschaften 87:180–183

Bonatto SL, Salzano FM (1997) A single and early origin for the peopling of the Americas supported by mitochondrial DNA sequence data. Proc Natl Acad Sci USA 94:1866–1871

Buxton LHD (1925) The peoples of Asia. London

Cann RL (2001) Genetic clues to dispersal of human populations: retracing the past from the present. Science 291:1742–1748

Cavalli-Sforza LL, Menozzi P, Piazza, A (1994) History and geography of human genes. Princeton University Press, Princeton

Chen Y-S, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC (1995) Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. Am J Hum Genet 57:133–149

Clark VJ, Sivendren S, Saha N, Bentley GR, Aunger R, Sirajuddin SM, Stoneking M (2000) The 9-bp deletion between the mitochondrial Lysine tRNA and COII genes in tribal populations of India. Hum Biol 72:273–285

Fu Y-X (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147:915–925

Guha BS (1935) The racial affinities of the people of India. In Census of India, 1931. III. Ethnographical. Government of India Press, Simla

Harpending HC, Sherry ST, Rogers AR, Stoneking M (1993) The genetic structure of ancient human populations. Curr Anthropol 34:483–496

Jorde LB, Bamshad MJ, Watkins WS, Zenger R, Fraley AE, Krakowiak PA, Carpenter KD, Soodyall H, Jenkins T, Rogers AR (1995) Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am J Hum Genet 57:523–538

Keith A (1936) Review of B.S. Guha's racial affinities of the peoples of India. Man 29:37

Kennedy KAR, Deraniyagala SU, Roertgen WJ, Chiment J, Sisotell T (1987) Upper Pleistocene fossil hominids from Sri Lanka. Am J Phys Anthropol 72:441–461

Kivisild T, Bamshad MJ, Kaldma K, Metspalu M, Metspalu E, Reidla M, Laos S, Parik J, Watkins WS, Dixon ME, Papiha SS, Mastana SS, Mir MR, Ferak V, Villems R (1999) Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. Curr Biol 9:1331–1334

Kosambi DD (1991) The culture and civilisation of ancient India in historical outline. Vikas, New Delhi

Lahr MM, Foley RA (1998) Towards a theory of modern human origins: Geography, demography, and diversity in recent human evolution. Yearbook of Physical Anthropology 41:127–176

Lundstrom RS, Tavare S, Ward RH (1992) Estimating substitution rates from molecular data using the coalescent. Proc Natl Acad Sci USA 89:5961–5965

Majumder PP (1998) People of India: Biological diversity and affinities. Evol Anthropol 6:100–110

Majumder PP, Roy B, Banerjee S, Chakraborty M, Dey B, Mukherjee N, Roy M, Guha Thakurta P, Sil SK (1999) Human-specific insertion/deletion polymorphisms in Indian populations and their possible evolutionary implications. Eur J Hum Genet 7:435–446

Miller SA, Dykes DD, Polesky HF (1988) A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 16:1215

Misra VN (1992) Stone Age in India: an ecological perspective. Man Environ 14:17–64

Mountain JL, Hebert JM, Bhattacharyya S, Underhill PA, Ottolenghi C, Gadgil M, Cavalli-Sforza LL (1995) Demographic history of India and mtDNA sequence diversity. Am J Hum Genet 56:979–992

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nei M, Jin L (1989) Variances of the average number of nucleotide substitutions within and between populations. Mol Biol Evol 6:290–300

Parpola A (1975) On the protohistory of the Indian languages in the light of archaeological, linguistic and religious evidence: an attempt at integration. In: van Lohuizen-De Leeuw JE (ed) South Asian Archaeology. Brill, New York, pp 73–84

Passarino G, Semino O, Modiano G, Bernini LF, Santachiara-Benerecetti AS (1996a) mtDNA provides the first known marker distinguishing proto-Indian from the other Caucasoids; it probably predates the diversification between Indians and Orientals. Ann Hum Biol 23:121–126

Passarino G, Semino O, Bernini LF, Santachiara-Benerecetti AS (1996b) Pre-Caucasoid and Caucasoid genetic features of the Indian population, revealed by mtDNA polymorphisms. Am J Hum Genet 59:927–934

Pattanayak DP (1998) The language heritage of India. In: Balasubramanian D, Rao NA (eds) The Indian human heritage. Universities Press, Hyderabad, pp 95–99

Quintana-Murci L, Semino O, Bandelt H-J, Passarino G, McElreavey K, Santachiara-Benerecetti AS (1999) Genetic evidence of an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. Nat Genet 23:437–441

Quintana-Murci L, Krausz C, Zerjal T, Sayar SH, Hammer MF, Mehdi SQ, Ayub Q, Qamar R, Mohyuddin A, Radhakrishna U, Jobling MA, Tyler-Smith C, McElreavey K (2001) Y-chromosoem lineages trace diffusion of people and languages in southwestern Asia. Am J Hum Genet 68:537–542

Rapson EJ (1955) Peoples and languages. In Rapson EJ (ed) Cambridge history of India, vol 1. Chand, Delhi

Ray N (1973) Nationalism in India. Aligarh Muslim University, Aligarh

Renfrew C (1987) Archaeology and language: the puzzle of Indo-European origins. Jonathan Cape, London

Renfrew C (1992) Archaeology, genetics and linguistic diversity. Man 27:445–487

Renfrew C (2000) At the edge of knowability: towards a prehistory of languages. Cambridge Archaeol J 10:7–34

Risley HH (1915) The people of India. Thacker Spink, Calcutta

Roychoudhury S, Roy S, Dey B, Chakraborty M, Roy M, Roy B, Ramesh A, Prabhakaran N, Usha Rani MV, Vishwanathan H, Mitra M, Sil SK, Majumder PP (2000) Fundamental genomic unity of ethnic India is revealed by analysis of mitochondrial DNA. Curr Sci 79:1182–1192

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174–175

Ruhlen M (1991) A Guide to the World's Languages. Stanford University Press, Stanford

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sarkar SS (1958) Race and race movements in India. In: Chatterjee SK (ed) The cultural heritage of India, vol. Ramakrishna Mission Institute of Culture, Calcutta pp 17–32

Schneider S, Kueffer J-M, Roessli D, Excoffier L (2000) Arlequin: a software for population genetic data analysis. University of Geneva, Geneva

Singh KS (1992) People of India: an introduction. Anthropological Survey of India, Calcutta

Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129:555–562

Thapar R (1966) A history of India, vol 1. Penguin, Middlesex

Tishkoff SA, Ruano G, Kidd JR, Kidd KK (1996) Distribution and frequency of a polymorphic Alu insertion at the plasminogen activator locus in humans. Hum Genet 97:759–764

Torroni A, Schurr TB, Cabell MF, Brown MD, Neel JV, Larsen M, Smith DG (1993) Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet 53:563–590

Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, Obinu D, Savontaus M-L, Wallace DC (1996) Classification of European mtDNAs from an analysis of three European populations. Genetics 144:1835–1850

Vigilant LA, Wilson AC, Harpending H (1991) African populations and the evolution of human mitochondrial DNA. Science 253:1503–1507

Watkins WS, Bamshad M, Dixon ME, Bhaskara Rao B, Naidu JM, Reddy PG, Prasad BV, Das PK, Reddy PC, Gai PB, Bhanu A, Kusuma YS, Lum JK, Fischer P, Jorde LB (1999) Multiple origins of the mtDNA 9-bp deletion in populations of South India. Am J Phys Anthropol 109:147–158

Watson E, Forster P, Richards M, Bandelt H-J (1997) Mitochondrial footprints of human expansions in Africa. Am J Hum Genet 61:691–704