Keywords. Demography; migration; mitochondrial DNA; polymorphism; Y

India is a land of enormous genetic, cultural and linguistic diversity. With the exception of Africa, India harbours more genetic diversity than other comparable global regions (Majumder 1998). The enormous diversity in social and cultural beliefs and practices has been well documented and emphasized (Karve 1961; Beteille 1998). The population of India is culturally stratified, broadly into tribals and non-tribals. It is generally accepted that the tribal people, who constitute 8·

isolation and/or social regulations governing matings. Therefore, any general statement on an ethnic group must rule out the existence of such subgroups, or must be confined to the subgroups that have been studied. Identifying subgroups from a genetic standpoint is, by itself, a major scientific endeavour. In the present paper, we have used names of ethnic groups without investigating the existence of genetic substructuring within the groups. While this may be acceptable for providing insights in relation to some macro-level questions such as the ones entertained in the present study, certainly fine-tuning of our inferences may need to be done in the future. Second, because our studies and questions have evolved over a period of time, our choice of ethnic groups has sometimes been opportunistic, and sometimes more focussed in relation to our hypotheses. Therefore, not all hypotheses have been tested using DNA samples drawn from the same set of ethnic groups. We do not consider this to be a limitation, as not all hypotheses need to be or can be tested using a uniform set of ethnic groups. Third, although from each group we have sampled individuals who were unrelated at least to the first-cousin level, because we have sampled them from restricted geographical areas, it is possible that our sample may not represent the entire spectrum of genomic diversity of the group, especially if there are subgroups within the group. Fourth, our sample sizes from some of the ethnic groups are small. Although our inferences have been based on statistical tests that take variable sample sizes into account, we acknowledge that larger sample sizes would have added robustness to our inferences. However, it may be pointed out that it is well-known in population genetics that in comparison with inferences based on data of a large number of individuals, inferences based on data of a smaller number of individuals but a larger number of loci are equally robust. We have, therefore, studied a larger number of loci to compensate for our restricted sample sizes.

3. Fundamental genomic unity of India

Since the seminal study of Cann *et al* (1987), mitochondrial DNA (mtDNA) data have proven to be extremely useful in the study of human evolution, including prehistoric migrations and demographic events such as sudden population expansion or extreme bottlenecks (Sherry *et al* 1994). In other words, mtDNA enables to probe the distant past.

We have studied 644 mtDNA samples collected from 23 ethnic populations; 10 populations from the eastern states of West Bengal (5 populations), Orissa (4 populations) and Tripura (1 population), 1 population from the central state of Madhya Pradesh, 4 populations from the northern state of Uttar Pradesh, and 8 populations from the southern state of Tamil Nadu. These populations were chosen to include both tribal and caste populations at different levels of social hierarchy. One group of Muslims from Uttar Pradesh has also been included. The tribal populations belong to three different linguistic groups (Austro-Asiatic, Dravidian and Tibeto-Burman), and the caste populations are either Indo-Aryan speakers (northern Indian castes) or Dravidian speakers (southern Indian castes). The Muslims studied are all Indo-Aryan language speakers.

We have screened 8 mtDNA loci. The 9-bp COII/tRNA^{Lys} intergenic length mutation revealed that all populations were monomorphic; no sampled individual showed the presence of the 9 bp deletion. The remaining seven RFLP loci were polymorphic in the pooled data set (see footnote of table 1 for details). Seven-locus haplotypes were constructed and their frequencies estimated in each population. A total of 19 different haplotypes were observed in the pooled data set. However, in none of the populations were all 19 haplotypes observed. The maximum number of haplotypes (13) was observed among Rajputs; the Kotas harboured only 2 haplotypes. Frequencies of haplotypes in each study population, as also in the pooled sample, are presented in table 1. The frequency distribution of haplotypes in the pooled data set is nearly unimodal; only one haplotype (0111011) accounted for about 50% of all mtDNA molecules. It can, therefore, be inferred that this is the most ancient haplotype in Indian populations. It is also seen from table 1 that in 20 of the 23 study populations, this modal haplotype is the most frequent. The three populations in which this haplotype is not the most frequent are all inhabitants of Uttar Pradesh in northern India –

demographic expansions, geographic dispersal and social groupings. The lack of correspondence of clusters based on mtDNA haplotype frequencies with either geographical location of habitat, language or social proximity is consistent with such a model for the peopling of India. Further, because of the extensive haplotype sharing among ethnic groups, the extent of observed variation in haplotype frequencies attributable to differences between groups is small; most observed haplotype variation is between individuals within groups (Roychoudhury *et al* 2000).

4. Where to?

We have also compared the distributions of haplotypes found in the populations included in the present study with those of other populations of southeast Asia. For this purpose, we collated and compacted the haplotype data presented in Ballinger *et al* (1992). Since Ballinger *et al* (1992) did not study the RFLP site at nt 12308 studied by us, this locus had to be excluded for purposes of comparison. The results, presented in figure 1, show that there is a considerable sharing of haplotypes between Indian and southeast Asian populations. The distributions of haplotype frequencies are also similar. There is, however, one notable difference. The southeast Asian populations harbour a set of haplotypes, albeit with low or medium frequencies, on the 9 bp deletion background, which are completely absent in the present study populations. Most of these haplotypes are also on a DdeI(10394)-AluI(10397) –/– background. Ballinger *et al* (1992) have hypothesized that the 9 bp deletion arose in central China and radiated out from this region as migrants moved to populate parts of southeast Asia. If India was also populated by migrants radiating out from central China, one would have expected that a significant proportion of the migrants would carry the (–/– 9-bp-*del*) haplotype; hence this haplotype should be present in Indian populations in polymorphic frequencies. However, this haplotype has not been observed in any of the populations investigated in the present study, nor was it detected in an earlier study (Rickards 1995). On the other hand, a significant proportion of the southeast Asian populations possess the 9-bp ‘non-deletion’ allele on DdeI(10394)-AluI(10397) +/+ or +/– backgrounds. In fact, the two classes of haplotypes observed in southeast Asian populations (see figure 1), one of which is completely absent in Indian populations, leads us to believe that southeast Asian populations were derived from two sources; one from India and the other possibly from central or southern China. It may be noted that the 9 bp deletion is present in high frequencies among Tharus (Passarino *et al* 1993) and Japanese (Cann *et al* 1987; Horai and Matsunaga 1986) populations that are postulated to have arisen from human migrations originating from southern China. It is known (Beteille 1998; Diamond 1997) that there were two waves of human migration from mainland Asia through southeast Asia to New Guinea and Java. One of these was an early wave that occurred about 40,000 ybp. The other wave, the so-called Austronesian migration from south China, took place about 4000–3500 ybp. Although we cannot be certain, we postulate that the early wave of migration was from India and carried the +/+ haplotype into southeast Asia. The second wave of migration from south China may have carried the (–/– 9-bp-*del*) haplotype into this region. An early wave of migration from India, actually from Africa through

Table 1. Absolute and percentage frequencies of 7-locus mtDNA haplotypes in 23 ethnic populations of India.

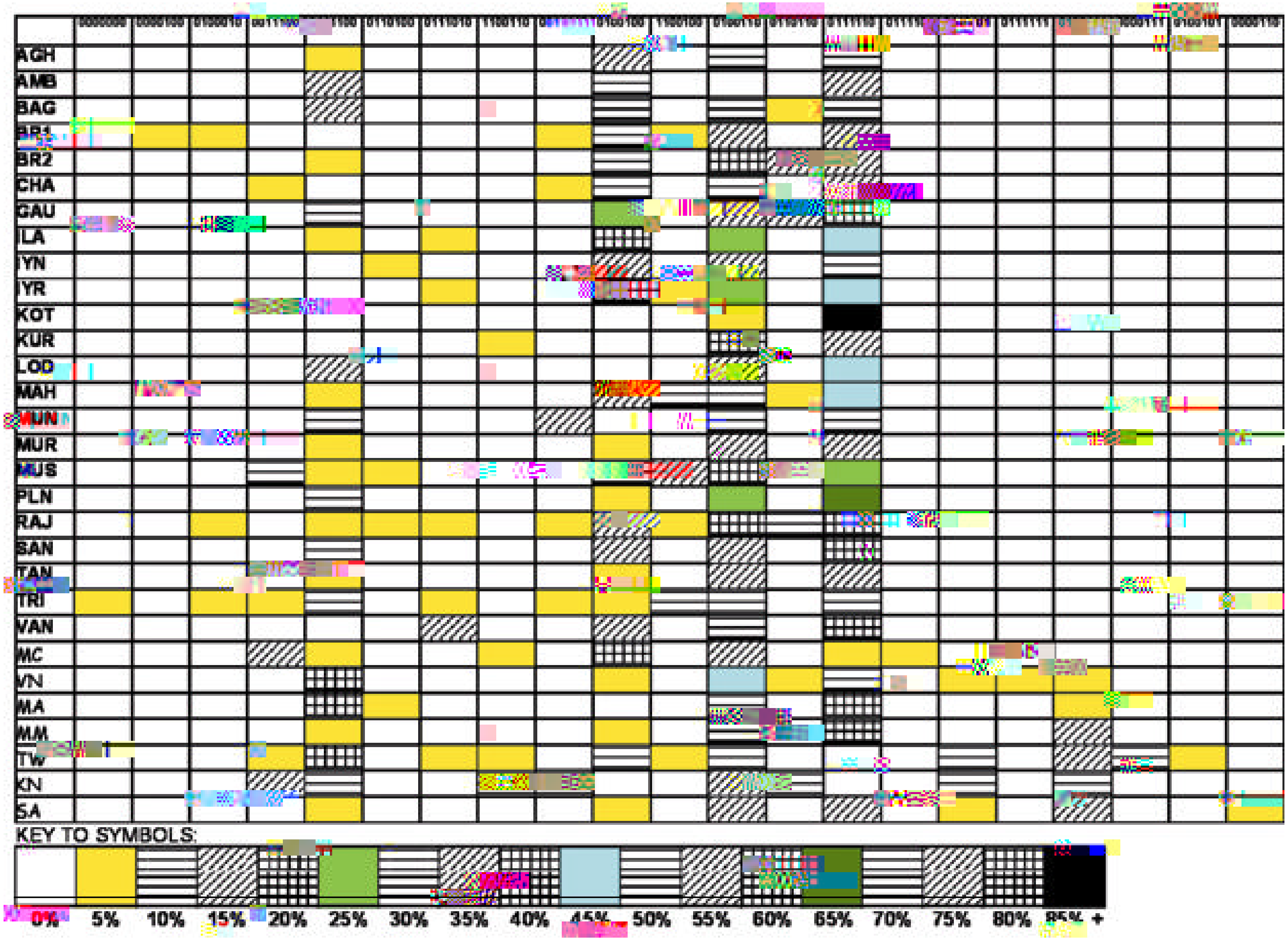

Figure 1. mtDNA haplotype diversities in 23 ethnic populations of India.

expansion model (Harpending *et al* 1993) and examined the fit of the observed and expected mismatch distributions for the three linguistic groups of tribals. The observed and expected distributions are presented in figure 2(a–c). From the unimodal nature of the observed mismatch distributions, their smoothness [as revealed by the very small values of the raggedness statistic (Harpending *et al* 1993); table 2] and the reasonably good fits with 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 2) in our data set for each of the three language groups. We estimated the expansion times, which are also presented in table 2, using the methodology proposed by Slatkin and

Table 2. Descriptive statistics and estimated expansion times of Indian tribals belonging to various language groups and haplogroups.

Linguistic group	No. of sequences	No. of poly-morphic sites	No. of muta-tions	Nucleotide diversity (*p*) ± 2 SD	Mean No. of mis-matches (*k*)	Ragged-ness (*r*)	F_S (*P*-value)	Expansion time in years before present (95% confidence interval)
Austro-Asiatic	34	54	59	0·023 ± 0·002	7·747	0·0157	– 19·176 (0·000)	56098 (51220–60975)
Dravidian	61	59	66	0·016 ± 0·002	5·432	0·0058	– 25·417 (0·000)	39024 (34146–43902)
Tibeto-Burman	20	42	42	0·021 ± 0·001	7·170	0·0281	– 12·203 (0·000)	51220 (48780–53659)

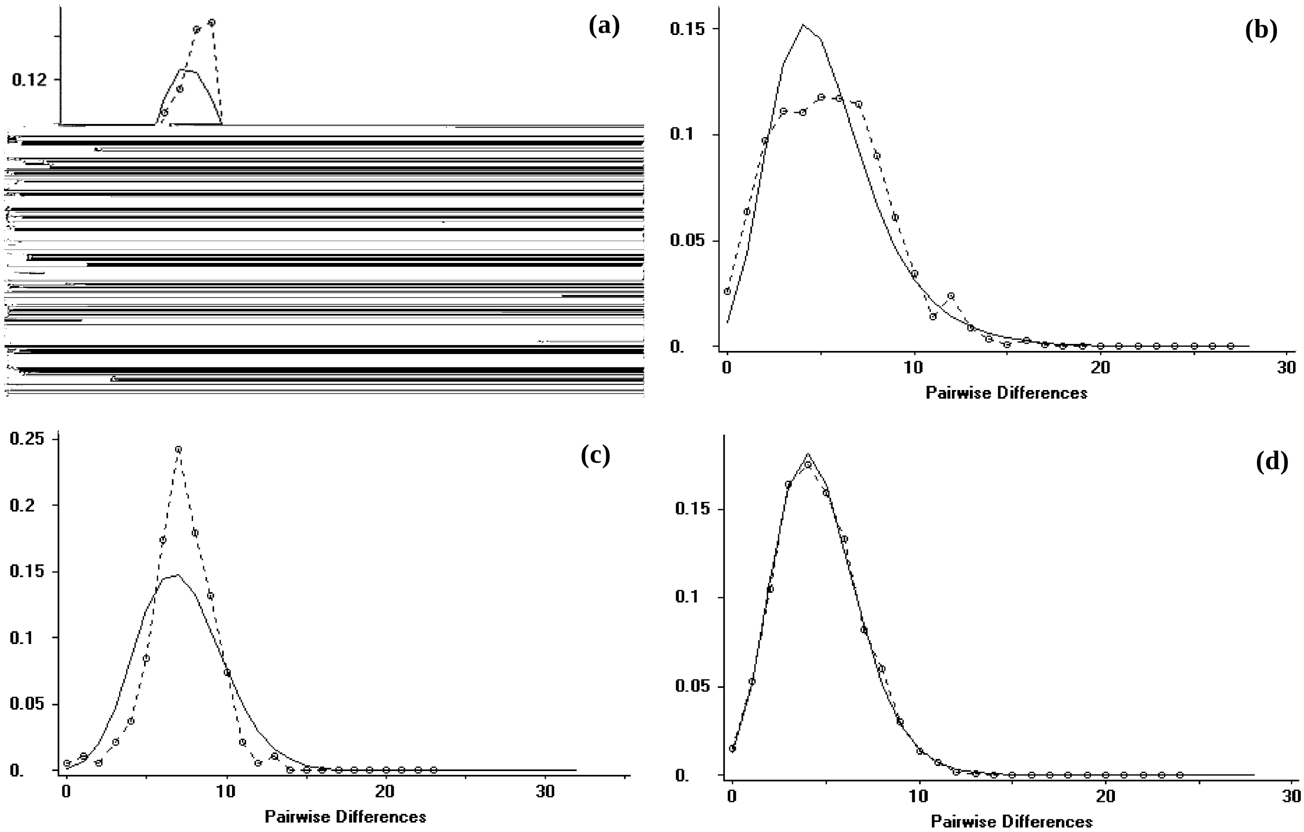

Figure 2. Observed (--o--o--) and expected (——) mismatch distributions based on mtDNA HVS-1 sequences for (**a**) Austro-asiatic, (**b**) Dravidian, (**c**) Tibeto-Burman speaking tribals of India, and (**d**) tribals belonging to mtDNA haplogroup M.

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 taken to be twice the standard deviation of the sampling variance of nucleotide diversity. The estimated expansion time of the Austro-Asiatics (≈56,000 ybp) is much older than that of the Dravidians (≈39,000 ybp). This difference is significant as indicated by the disjoint 95% confidence intervals of the estimates. Our tentative estimate, in view of the limited sample size, of the expansion time of the Tibeto-Burmans is ≈52,000 ybp.

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 (Risley 1915; Thapar 1966; Pattanayak 1998) 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 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 compared to 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. This haplogroup is known to occur in high frequencies among Caucasian populations, including those of central and west Asia. The frequency of this haplogroup in our pooled tribal 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. Kivisild *et al* (1999) found that there are several subclusters of haplogroup U, of which they had found 6 to be present in their sample of Indians. We have found only 2 of these subclusters to be present among the tribals in India. These are subclusters U2i and subcluster U1, with frequencies 77·3% and 9·1%, respectively. Interestingly, we have found that all the 6 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. 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 Kivisild *et al*'s (1999) conclusion 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 subcluster U1, K, U4, U5 with frequencies between 1%–5% in their samples from India, we found only the subcluster U1 in our tribal samples with a frequency of 9%. The subcluster U7, found at a frequency of about 13% in Kivisild *et al*'s (1999) samples but not found in our tribal samples, may also be western-Eurasian. Since the samples included in Kivisild *et al*'s (1999) study 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 may have been introduced in India with the entry of Aryan speakers from west and central Asia. This is contrary to Kivisild *et al*'s (1999) suggestion that all of the western-Eurasian subclusters of haplogroup U were present in India before the entry of the Aryan speakers. Thus, mtDNA provides signatures of population movements into India from central and west Asia.

Since many contend that immigrants from central and west Asia were predominantly males, we also sought to find similar signatures on Y-chromosomal DNA (Mukherjee *et al* 2001). Prehistoric, historic and linguistic evidences have suggested that Middle Eastern/west Asian and central Asian gene pools have contributed to the Indian gene pool. The northern exit route of humans from Africa to India was through the Middle East and west Asia. Subsequently, with the development of agriculture in the fertile crescent region that extends from Israel through northern Syria to western Iran, there was possibly migration of humans from this region into India. More recently, pastoral nomads originating in the central Asian

Table 3. Haplogroup frequencies among 8 tribal populations of India.

Population name	Geographical region of sampling	Haplogroup frequency* (%)					
		A	C†	D†	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)

*Haplogroups B, H and L were not observed in the study samples.
†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.

steppes may also have contributed to the gene pool of India. The entry of humans from these regions into India was through the northwest corridor of India (Thapar 1975). We have, therefore, chosen to investigate gene pools of contemporary population groups inhabiting northern India, since traces of ancient admixture are likely to be more easily detected in northern India than elsewhere. We have studied four groups inhabiting the northern Indian state of Uttar Pradesh. The ethnic groups were: Brahmin (BRA), Chamar (CHA), Muslim (MUS) and Rajput (RAJ). The Brahmins, Rajputs and Chamars all belong to the Hindu caste fold and occupy upper, middle and lower ranks, respectively, in the caste hierarchy. The Muslim is an Islamic religious group. Most individuals belonging to this group are religious converts from various other populations that inhabited this geographical location. The study is based on Y-chromosomal polymorphisms; our inferences, therefore, reflect male population movements and admixture. We have collated data, mostly published and some unpublished, from several Middle Eastern population groups (Hammer 2000; Nebel 2000) and have performed comparative statistical analyses to draw inferences.

DNA samples were typed in respect of 12 binary polymorphic markers – YAP, 92r7, SRY 4064, sY81, SRY+465, TAT, M9, M13, M17, M20, SRY10831 and p12f2. Based on the UEP markers, we have classified Y chromosomes of each population into haplogroups (HGs) as defined by Rosser *et al* (2000). The haplogroup frequencies are presented in table 4. Because the set of markers screened in this study were not exactly the same as those screened in the published studies (Hammer *et al* 2000; Nebel *et al* 2000), some of the haplogroups could not be resolved and had to be pooled. It is seen that there is substantial overlap in the types of haplogroups observed in the north Indian and in the Middle Eastern regions.

The distributions of Y haplogroups, defined (Rosser *et al* 2000) on the basis of 12 biallelic UEPs reveal many interesting patterns. The haplogroup diversities in the populations of northern India and the Middle East are quite high, which is indicative of large long-term effective population sizes or high rates of gene flow from disparate populations or both. Among the north Indian populations, the differences in the frequency distributions of haplogroups are not statistically significant, but these differences among the Middle Eastern populations are significant. Although several haplogroups are common to the north Indian and Middle Eastern populations, the haplogroup frequency distributions in these two regions are substantially different. In northern India, HG-3 is the most frequent (35%–58%), while HG-9 is the most frequent

Table 4.

(33%–57%) in Middle Eastern populations. Globally, the peak of HG-9 frequency is in the Caucasus-Anatolia region (Rosser *et al* 2000). This haplogroup is thought to have arisen about 5,500–17,400 ybp (Hammer *et al* 2000; Quintana-Murci *et al* 2001) in this region (south-western Iran). Our estimate (table 5) of the age of this haplogroup from data on the Middle Eastern populations is in fair agreement with this previous estimate. As noted in a previous study (Quintana-Murci *et al* 2001), this haplogroup may have been brought into India by Indo-European speakers from the Middle East. The frequency of this haplogroup is highest (23·5%) among the upper-ranked caste Brahmins and is lower (17·1%) among the middle-ranked caste Rajput. It is known that after the entry of the Aryan-speakers into India, the Brahmins were the torchbearers and promoters of Aryan rituals (Karve 1961). Therefore, it is likely that this group had the highest genetic contact with the Aryan-speaking peoples. This observation is consistent with the high frequency of HG-9 observed among them. This haplogroup may have percolated into the middle-ranked Rajputs, either through admixture with Brahmins or directly with the Aryan-speaking immigrants. It is noteworthy that HG-9 is absent among the low-ranked caste group, Chamar. A large section of the Muslims of Uttar Pradesh are known to be religious converts from both upper and middle-ranked caste groups. Our observation that HG-9 occurs in a lower frequency (10·5%) among the Muslim compared to the Brahmin and the Rajput is consistent with the known social history of this group.

Haplogroup-3, which is the most frequent haplogroup in India, is known to be widely found in Asia, except Eastern Asia, and is virtually absent in Africa and the Americas (Karafet *et al* 1999). HG-3 is found in high frequencies in central Asia (Russia and Altai region) and east Europe (Poland and Hungary). It appears that this haplogroup arose in central Asia about 7,500 ybp (Karafet *et al* 1999; Zerjal *et al* of the chemokine family of receptors serve as critical portals for the entry of HIV-1 into target cells. A mutant allele (Δ*ccr5*) of the *b*-chemokine receptor gene *CCR5* carrying a 32 base-pair deletion prevents cell invasion by the primary transmitting strain of HIV-1. The frequency of this mutant allele is known to be high among Caucasian populations, including populations of central and west Asia (Martinson *et al* 1997). We have screened about 1500 individuals, drawn from 40 diverse ethnic populations of India. We have found that the Δ*ccr5* allele is completely absent or only sporadically present in most populations. However, among the Muslims of Uttar Pradesh, the frequency of this allele was 5·36%, which may be due to admixture with immigrants from central and west Asia.

Acknowledgements

The studies described in this article were carried out with the active collaboration of scientists of various Indian institutions. Funding for these studies were obtained primarily from the Department of Biotechnology, New Delhi. I am grateful to my various collaborators, notably Susanta Roychoudhury, Nitai P Bhattacharyya, Bidyut Roy, A Ramesh, M V Usha Rani and Samir K Sil. I am also grateful to Sangita Roy, Namita Mukherjee, Badal Dey, Monami Roy, Sanat Banerjee, Madan Chakraborty, H Vishwanathan, N Prabhakaran and Analabha Basu, for help in data collection, DNA analyses, statistical analyses and other aspects of these studies. Ariella Oppenheim, Almut Nebel and Marina Faerman actively collaborated with me on an Indo-Israeli project on genomic continuity from west and central Asia to India.

References

Ballinger S W, Schurr T G, Torroni A, Gan Y Y, Hodge J A, Hassan K, Chen K-H and Wallace DC 1992 Southeast Asian mitochondrial DNA analysis reveals continuity of ancient mongoloid migrations; *Genetics* **130** 139–152

Beteille A 1998 The Indian heritage – a sociological perspective; in *The Indian human heritage* (eds) D Balasubramanian and N A Rao (Hyderabad: Universityies Press) pp 27–94

Bhattacharyya N, Basu P, Das M, Pramanik S, Banerjee R, Roy B, Roychoudhury S and Majumder P P 1999 Negligible geneflow across ethnic boundaries in India, revealed by analysis of Y-chromosomal DNA polymorphisms; *Genome Res.* **9** 711–719

Bonatto S L and Salzano F M 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 L H D 1925 *The peoples of Asia* (London)

Cann R L 2001 Genetic clues to dispersal of human populations: Retracing the past from the present;

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

Majumder P P and Dey B 2001 Absence of the HIV-1 protective Δ*ccr5* allele in most ethnic populations of India; *Eur. J. Hum. Genet.* (in press)

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

Martinson J, Chapman N H, Rees D C, Liu Y-T and Clegg J B 1997 Global distribution of the CCR5 gene 32-basepair deletion; *Nature Genet.* **16** 100–103

Meenakshi K 1995 Linguistics and the study of early Indian history; in *Recent perspectives of early Indian history* (ed.) R Thapar (Bombay: Popular Prakashan) pp 53–79

Misra V N 1992 Stone age in India: an ecological perspective; *Man Env.* **14** 17–64

Misra V N 2001 Prehistoric human coonization of India; *J. Biosci. (Suppl.)* **26** 491–531

Mountain J L, Hebert J M, Bhattacharyya S, Underhill P A, Ottolenghi C, Gadgil M and Cavalli-Sforza L L 1995 Demographic history of India and mtDNA sequence diversity; *Am. J. Hum. Genet.* **56** 979–992

Mukherjee N, Nebel A, Oppenheim A and Majumder P P 2001 High resolution mapping of Y-chromosomal polymorphisms reveals signatures of population movements from central and west Asia into India; *Hum. Genet.* (submitted)

Nebel A, Filon D, Weiss D A, Weale M, Faerman M, Oppenheim A and Thomas M G 2000 High-resolution Y chromosome haplotypes of Israeli and Palestinian Arabs reveal geographic substructure and substantial overlap with haplotypes of Jews; *Hum. Genet.* **107** 630–641

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

Passarino G, Semino O, Modiano G, Santachiara-Benerecetti A S and Wallace D C 1993 COII/tRNA^{lys} intergenic 9-bp deletion and other mtDNA markers clearly reveal that the Tharus (southern Nepal) have Oriental affinities; *Am. J. Hum. Genet.* **53** 609–618

Pattanayak D P 1998 The language heritage of India; in *The Indian human heritage* (eds) D Balasubramanian and N A Rao (Hyderabad: Universities Press) pp 95–99

Quintana-Murci L, Krausz C, Zerjal T, Sayar S H, Hammer M F, Mehdi S Q, Ayub Q, Qamar R, Mohyuddin A, Radhakrishna U, Jobling M A, Tyler-Smith C and McElreavey K 2001 Y-chromosome lineages trace diffusion of people and languages in southwestern Asia; *Am. J. Hum. Genet.* **68** 537–542

Rapson E J 1955 Peoples and languages; in *Cambridge history*