Fundamental genomic unity of ethnic India is revealed by analysis of mitochondrial DNA††

Susanta Roychoudhury†, Sangita Roy†, Badal Dey*, Madan Chakraborty*, Monami Roy*, Bidyut Roy*, A. Ramesh#, N. Prabhakaran#, M. V. Usha Rani‡, H. Vishwanathan‡, Mitashree Mitra, Samir K. Sil§ and Partha P. Majumder*,##**

†Human Genetics and Genomics Division, Indian Institute of Chemical Biology, Calcutta 700 032, India
*Anthropology and Human Genetics Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700 035, India
#Department of Genetics, University of Madras, Chennai 600 013, India
‡Department of Environmental Science, Bharathiar University, Coimbatore 641 046, India
**School of Studies in Anthropology, Pandit Ravishankar Shukla University, Raipur 492 010, India
§Department of Life Science, Tripura University, Agartala 799 004, India

Mitochondrial DNA (mtDNA) profiles of 23 ethnic populations of India drawn from diverse cultural, linguistic and geographical backgrounds are presented. There is extensive sharing of a small number of mtDNA haplotypes, reconstructed on the basis of restriction fragment length polymorphisms, among the populations. This indicates that Indian populations were founded by a small number of females, possibly arriving on one of the early waves of out-of-Africa migration of modern humans; ethnic differentiation occurred subsequently through demographic expansions and geographic dispersal. The Asian-specific haplogroup *M* is in high frequency in most populations, especially tribal populations and Dravidian populations of southern India. Populations in which the frequencies of haplogroup *M* are relatively lower show higher frequencies of haplogroup *U*; such populations are primarily caste populations of northern India. This finding is indicative of a higher Caucasoid admixture in northern Indian populations. By examining the sharing of haplotypes between Indian and south-east Asian populations, we have provided evidence that south-east Asia was peopled by two waves of migration, one originating in India and the other originating in southern China. These findings have been examined and interpreted in the light of inferences derived from previous genomic and historical studies.

INDIA occupies a centrestage in human evolution. India appears to have been on the path of migration of humankind that started out of Africa about 100,000 years ago[1]. Upper palaeolithic (40,000 ybp) cultures flourished in different parts of the country and probably represent the activities of modern man[2]. Using data on human-specific insertion/deletion polymorphisms, we[3] have recently provided evidence that a major population expansion of modern humans took place within India. Although the period of this demographic expansion remains uncertain, it has been speculated[4] that this event had taken place 60,000 to 85,000 ybp. Perhaps this demographic expansion, followed by subsequent migration, resulted in the peopling of south-east Asia and later (50,000 to 60,000 ybp) of Australia[5]. About 60,000 ybp, there was perhaps another independent expansion of modern humans in southern China[5,6], which may have resulted in human migration into India and also into south-east Asia.

Contemporary India is a land of enormous human genetic, cultural and linguistic diversity. Using data on blood group, serum protein and red-cell enzyme markers, it has been shown that, with the exception of Africa, India harbours more genetic diversity than other comparable global regions[7]. The enormous diversity in social and cultural beliefs and practices has been well documented and emphasized[8,9]. The social structure of the Indian population is dominated by the Hindu caste system. Most contemporary populations belong to the Hindu religious fold and are hierarchically arranged in four main caste classes, viz. Brahmin (priestly class), Kshatriya (warrior class), Vysya (business class) and Sudra (menial labour class). The tribals are predominantly ancestor worshippers. In addition, there are several religious communities, who practice different religions, viz. Islam, Christianity, Sikhism, Judaism, etc. The Indian linguistic area is one of the larger areas involving hundreds of languages from four major language families – Indo-Aryan (a branch of Indo-European), Dravidian, Austro-Asiatic and Tibeto-Burman (Sino-Tibetan). Of these, the first two have been the major contributors to the development of Indian culture and society[10]. Indian culture and society are also known to have been affected by multiple waves of migration that took place in historic and prehistoric times[11,12]. A section of Indo-Aryan speakers are believed to have migrated first to Iran and from there to the north-west of India where they encountered the indigenous people who spoke non-

††Dedicated to the memory of Jyotirmoy Das who provided invaluable encouragement to the initiation of this study.
##For correspondence. (e-mail: ppm@isical.ac.in)

Aryan languages[10]. The linguistic affinity of the indigenous people of India is a matter of debate. The contemporary tribal populations are largely Dravidian or Austro-Asiatic speakers. The Austro-Asiatic speakers in India are exclusively tribal. In view of the persistent survival of Dravidian languages in pockets of Iran, Baluchistan and Afghanistan, some linguists[13] believe that Dravidian speakers came into India from outside. Some[14], however, believe that since the Dravidian speakers are largely restricted to India, this language may have developed within India. The Austro-Asiatic family is a fragmented language group. It is most widely spoken in Vietnam and Cambodia. Within India, only a small number of ethnic groups speak Austro-Asiatic languages. It is, however, noteworthy that the Indian Austro-Asiatic speakers are exclusively tribal, which may be indicative of their being the oldest inhabitants of India[13,14]. Some believe that the Austro-Asiatic linguistic family evolved in southern China[15]. If indeed this is true, then Indian Austro-Asiatic speakers must have entered India from southern China through the north-east. Mitochondrial DNA (mtDNA) provides long-range probes of the past. Since the seminal study of Cann *et al*.[16], 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[17]. In view of the foregoing discussion that indicates conflicting and uncertain views about the peopling of India and south-east Asia, we undertook a study of variation in mtDNA among several ethnic groups of north-eastern, eastern, north-central and southern India with a view to partially reconstructing the population history of this region. The present paper provides the results of this study.

Subjects and methods

Study populations

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 Islamic group of Muslims 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 are Indo-Aryan speakers. Further details about the populations along with number of individuals sampled from each population are provided in Table 1. The geographical locations of sampling of individuals from the study populations are provided in Figure 1.

From each population, a set of individuals, unrelated at least to the first cousin level, was chosen. From each individual, 5 to 10 ml of blood was drawn by venipuncture with appropriate informed consent.

DNA analysis

DNA was isolated from each individual using a standard protocol[18]. Each DNA sample was screened in respect of a set of seven loci using PCR amplification followed by restriction digestion (where necessary), agarose gel electrophoresis, ethidium bromide staining and band visualization under UV light. The following seven RFLP loci were screened: HincII loss at nt 13259 (primers used for PCR amplification: Forward [F] – nt 13208–13232, Reverse [R] – nt 13413–13393 of Cambridge reference sequence; annealing temperature; 49°C), HaeIII gain at nt 663 (F – nt 577–598, R – nt 743–721, 58°C), AluI loss at nt 5176 (F – nt 5099–5122, R – nt 5333–5310; 61°C), DdeI and AluI gains at nt 10394 and 10397, respectively (F – nt 10284–10309, R – nt 10484–10458; 61°C), HaeIII gain at nt 16517 (F – nt 16453–14473; R – nt 48–25; 55°C), HinfI gain at nt 12308 (F – nt 12104–12124; R – nt 12338–12309; 63°C). Additionally, we have analysed the COII/tRNA^{Lys} intergenic deletion of one of the two copies of the 9 bp tandem repeat sequence (CCCCCTCTA) that occurs between nt 8272 and 8289, which has been

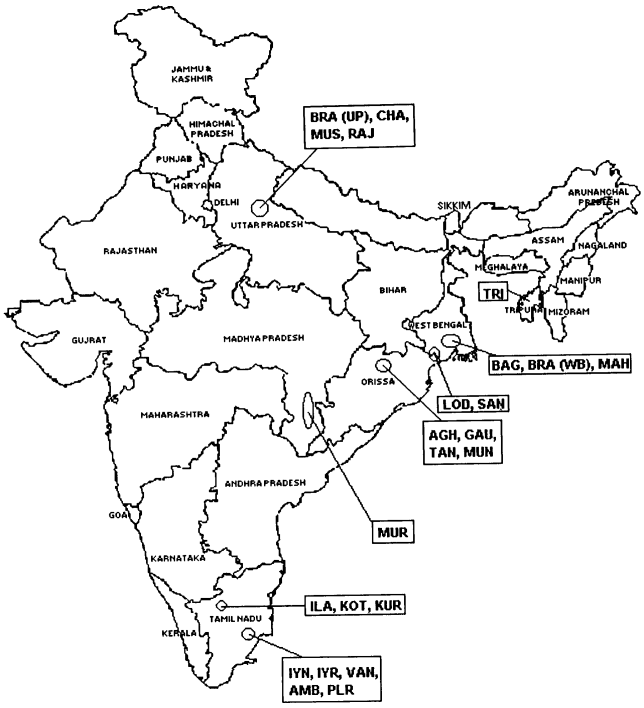

Figure 1. Geographical locations of sampling of study populations. (See Table 1 for explanation of abbreviations of study populations.)

Table 1. Study populations, sample sizes, geographical and ethnological information

Population (code)	Sample size	Geographical distribution	Linguistic affiliation	Social category	Occupation	Remarks
Agharia (AGH)	24	Eastern India	Indo-Aryan	Hindu caste	Agriculture	
Ambalakarer AMB)	30	Southern India – Tamil Nadu	Dravidian	Hindu caste	Agricultural labour	
Bagdi (BAG)	31	Eastern India	Indo-Aryan	Hindu caste	Agricultural labour	
Brahmins of Uttar Pradesh (BRA-UP)	27	Northern India – Uttar Pradesh	Indo-Aryan	Hindu caste	Traditionally priests, now various occupations	
Brahmins of West Bengal (BRA-WB)	22	Eastern India – West Bengal	Indo-Aryan	Hindu caste	Traditionally priests, now various occupations	
Chamar (CHA)	25	Northern India	Indo-Aryan	Hindu caste	Menial and agricultural labour	
Gaud (GAU)	13	Eastern India	Indo-Aryan	Hindu caste	Agriculture	
Irula (ILA)	30	Southern India – Primarily Tamil Nadu, including Nilgiri Hills	Dravidian	Tribe	Shifting cultivation	Possesses several Negrito morphological features
Iyengar (IYN)	30	Southern India – Tamil Nadu	Dravidian	Hindu caste	Traditionally priests, now various occupations	*Vaishnavites* – Worshippers of Lord Vishnu
Iyer (IYR)	30	Southern India – Tamil Nadu	Dravidian	Hindu caste	Traditionally priests, now various occupations	*Saivites* – Worshippers of Lord Shiva
Kota (KOT)	30	Southern India – Nilgiri Hills	Dravidian	Tribe	Artisans, musicians and agriculturists	One of the original inhabitants of Nilgiri Hills. Highly inbred group; significantly 'deviant' distributions of various polymorphic markers, including the near-absence of A blood group[41].
Kurumba (KUR)	30	Southern India – Nilgiri Hills	Dravidian	Tribe	Hunting and food gathering	Possesses several Negrito morphological features
Lodha (LOD)	32	Eastern India – West Bengal	Austro-Asiatic	Tribe	Hunting, food gathering and agircultural labour	
Mahishya (MAH)	33	Eastern India	Indo-Aryan	Hindu caste	Agriculture	
Munda (MUN)	7	Eastern India	Austro-Asiatic	Tribe	Hunting, food gathering and agriculture	
Muria (MUR)	49	Central India – Primarily Madhya Pradesh	Dravidian – Gondi dialect	Tribe	Agriculture, food gathering and hunting	There may have been considerable admixture with Austro-Asiatic tribals[42].
Muslim (MUS)	28	Throughout India	Indo-Aryan	Islamic religious group	Various occupations, including agriculture	The Muslims in India originated in one of two ways[43]; (a) one group came and settled during the various historic migrations and invasions, (b) the other group is believed to be formed through process of preselytization of the indigeneous Hindu population of a comparatively lower and middle order in the prevailing caste hierarchy, besides sections of a few tribes living in the fringe area of caste-dominated regions.
Pallar (PLR)	30	Southern India – Tamil Nadu	Dravidian	Hindu caste	Agriculture	
Rajput (RAJ)	51	Northern and Western India	Indo-Aryan	Hindu caste	Various occupations, including agriculture	
Santal (SAN)	20	Eastern India	Austro-Asiatic	Tribe	Agriculture, hunting and food gathering	
Tanti (TAN)	16	Eastern India	Indo-Aryan	Hindu caste	Weaving and agricultural labour	
Tripuri (TRI)	45	North-eastern India – Primarily Tripura	Tibeto-Burman	Tribe	Agriculture	One of the 10 authochthonous tribal groups of Tripura[44]; may be migrants from Central Asia[45].
Vanniyar (VAN)	30	Southern India – Tamil Nadu	Dravidian	Hindu caste	Agriculture	

used as a marker for populations of Asian and Asian-derived origins, such as Polynesians and Native Americans[19]. The primers used for PCR amplification of this region were: F – nt 8195–8215, R – nt 8317–8297; annealing temperature, 55°C. A random subset of the samples was also sequenced, for the purpose of cross-checking, using an ABI-377 automated DNA sequencer and the ABI Prism dideoxyterminator system (Applied Biosystems).

Statistical analysis

Allele and haplotype frequencies were calculated. Haplotype diversities were estimated according to Nei[20]. For estimating inter- and intra-populational genetic diversities, a gene diversity analysis[21] was performed and appropriate parameters were estimated. DNA sequences were aligned and analysed using CLUSTAL-W. AMOVA was performed using ARLEQUIN 1.1 (ref. 22). Phylogenetic relationships among haplotypes were estimated using the median network analysis[23] as implemented in NETWORK 1.1. Phylogenetic relationships among DNA sequences were estimated using the neighbour-joining algorithm[24].

Results

Variability at restriction sites

Screening for the 9-bp COII/tRNA^{Lys} intergenic length mutation revealed that all populations were monomorphic; no sampled individual showed the presence of 9-bp deletion. The remaining seven RFLP loci were polymorphic in the pooled data set. However, while the DdeI (10394) and AluI (10397) loci were polymorphic in all populations, several populations were monomorphic at the HaeIII (663), AluI (5176), HinfI (12308) and HincII (13259) loci. Population-wise allele frequencies are presented in Table 2. The average heterozygosity pooled over all populations and loci was 0.2394 ± 0.1986. There was considerable variability (0.0306–0.4802) in heterozygosities across loci (Table 2). The AluI (10397) locus exhibited the highest heterozygosity of 0.4802, which is close to the maximum attainable (0.5) for a biallelic locus.

Variation in haplotype frequencies

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 was 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 3. 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 3 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

Table 2. Allele frequencies at 7 mtDNA RFLP loci and heterozygosites in 23 ethnic populations of India

Population code	HaeIII (663) gain	AluI (5176) loss	DdeI (10394) gain	AluI (10397) gain	HinfI (12308) gain	HincII (13259) loss	HaeIII (16517) gain
AGH	0.000	0.000	0.750	0.750	0.208	0.000	0.792
AMB	0.000	0.000	0.900	0.900	0.067	0.000	0.567
BAG	0.000	0.000	0.645	0.613	0.129	0.000	0.774
BRA-UP	0.037	0.074	0.185	0.185	0.259	0.037	0.815
BRA-WB	0.000	0.000	0.727	0.901	0.046	0.000	0.864
CHA	0.000	0.080	0.640	0.640	0.320	0.000	0.880
GAU	0.000	0.000	0.615	0.462	0.385	0.000	0.692
ILA	0.000	0.000	0.533	0.533	0.233	0.033	0.766
IYN	0.000	0.000	0.533	0.500	0.100	0.000	0.833
IYR	0.033	0.000	0.500	0.500	0.233	0.033	0.767
KOT	0.000	0.000	0.967	0.967	0.000	0.000	1.000
KUR	0.033	0.000	0.767	0.761	0.067	0.000	1.000
LOD	0.000	0.000	0.812	0.812	0.188	0.000	0.656
MAH	0.091	0.000	0.485	0.454	0.212	0.000	0.758
MUN	0.000	0.143	0.714	0.714	0.000	0.000	0.714
MUR	0.000	0.000	0.800	0.800	0.067	0.000	0.933
MUS	0.143	0.071	0.393	0.357	0.321	0.000	0.643
PLR	0.000	0.000	0.733	0.733	0.133	0.000	0.867
RAJ	0.059	0.059	0.412	0.314	0.176	0.059	0.765
SAN	0.000	0.000	0.700	0.700	0.100	0.000	0.750
TAN	0.000	0.000	0.812	0.812	0.125	0.000	0.875
TRI	0.089	0.133	0.511	0.511	0.067	0.089	0.689
VAN	0.000	0.000	0.533	0.533	0.100	0.000	0.700
Heterozygosity	0.0514	0.0485	0.4707	0.4802	0.2580	0.0306	0.3367

Table 3. Frequencies of seven locus mtDNA haplotypes in 23 ethnic populations of India

Haplotype No.	Type*	AGH	AMB	BAG	BRA-UP	BRA-WB	CHA	GAU	ILA	IYN	IYR	KOT	KUR	LOD	MAH	MUN	MUR	MUS	PLR	RAJ	SAN	TAN	TRI	VAN	TOTAL
1	0111111							1 7.7											2 6.7						3 0.47
2	0110111							2 15.4												1 2.0					3 0.47
3	0100111	2 8.3		2 6.5	5 18.5	1 4.6	6 24.0	1 7.7	1 3.3		4 13.3		2 6.7	6 18.7	2 6.1		1 3.3	8 28.6	1 3.3	6 11.7		1 6.25	2 4.44	3 10.0	54 8.39
4	1100110														1 3.0										1 0.16
5	0100110	3 12.5	2 6.7	2 6.5	2 7.4		2 8.0	1 7.7	6 20.0	3 10.0					4 12.1		1 3.3	1 3.6	1 3.3	2 3.9	2 10.0	1 6.25	1 2.22		37 5.75
6	0111011	17 70.8	17 56.7	15 48.4	4 14.8	12 54.6	14 56.0	4 30.8	14 46.7	15 50.0	14 46.8	29 96.7	23 76.7	15 46.9	14 42.5	2 28.6	23 76.7	7 25.0	17 56.7	10 19.6	12 60.0	12 75.0	14 31.1	12 40.0	316 49.07
7	0011011				1 3.7		1 4.0									1 14.2				3 5.9			2 4.44		8 1.24
8	0110011			1 3.2		3 13.6									1 3.0					3 5.9					8 1.24
9	1100011												1 13.3												1 0.16
10	0100011			6 19.3	11 40.8	3 13.6	1 4.0	1 7.7	7 23.4	10 33.4	4 13.3	1 3.3	4 13.3		8 24.2	2 28.6	4 13.4	3 10.6	6 20.0	13 25.4	3 15.0	1 6.25	11 24.4	6 20.0	105 16.3
11	0111001								1 3.3		1 3.3									2 3.9			1 2.22	4 13.3	5 0.78
12	0100001				1 3.7															1 2.0			1 2.22		3 0.47
13	0111010	1 4.2	10 33.3	4 12.9		1 4.6		1 7.7	1 3.3					11 34.4	1 3.0	2 28.6	1 3.3	1 3.6	3 10.0	1 2.0	2 10.0	1 6.25	4 8.89		49 7.61
14	0011010						1 4.0											2 7.1					2 4.44		5 0.78
15	0110010									1 3.3								1 3.6		1 2.0					3 0.47
16	1100010				1 3.7						1 3.3				2 6.1			4 14.3		3 5.9			4 8.89		15 2.33
17	0100010	1 4.2	1 3.3	1 3.2	1 3.7	2 9.0		2 15.3		1 3.3	3 10.0							1 3.6		5 9.8	1 5.0		1 2.22	5 16.7	25 3.88
18	0000010				1 3.7																				1 0.16
19	0000000																						2 4.44		2 0.31

*Order of loci: HaeIII (663), AluI (5176), DdeI (10397), HinfI (12308), HincII (13259), HaeIII (16517).
1, Presence of restriction site; 0, Absence of restriction site.

all inhabitants of Uttar Pradesh in northern India – Brahmins, Rajputs and Muslims. Among Brahmins and Rajputs of Uttar Pradesh, the most frequent haplotype is 0100011. Among Muslims of Uttar Pradesh the most frequent haplotype is 0100111. However, all these three populations also harbour the 0111011 haplotype, which is modal in the remaining 20 populations, in fairly high frequencies (15% to 25%).

Population-specific haplotype diversities are presented in Figure 2. The haplotype diversity in most populations is quite high; the Rajputs exhibit the highest diversity (86%) and the Kotas the lowest (6%). There is no clear patterning of the extent of haplotype diversity by socio-religious category or by geographical zone.

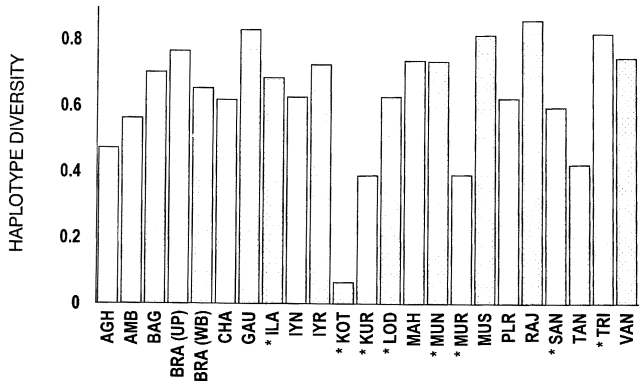

Figure 2. mtDNA haplotype diversities in 23 ethnic populations of India. Tribal populations are indicated by an asterisk.

We performed an analysis of molecular variance (AMOVA) using haplotype frequency data[25] for examining variability within and between (a) caste and tribal populations, (b) geographical zones, and (c) language families. This analysis of genomic structure of populations is similar to that based on data on allele frequencies at unlinked loci[26,27], but is particularly appropriate for haplotype data because it takes into account the number

of mutations between haplotypes. When this analysis was performed for caste and tribes defined as two groups of populations (14 castes, 8 tribal, Muslims excluded; $n = 616$), it was found that the percentages of haplotype diversity attributable to (i) between groups was 1.84%, (ii) among populations within groups was 7.65%, and (iii) among individuals within populations was 90.51%. Thus, we find that most of the mtDNA haplotype diversity occurs within populations, and that interpopulational or intergroup variation is small. The population structure F-statistics were: F_{CT} = Diversity between groups relative to the total diversity = 0.0183; F_{SC} = Diversity among populations within groups = 0.0779 and F_{ST} = Diversity among populations relative to the total diversity = 0.0949. Permutation tests of significance revealed that the variation among populations within the caste and tribal groups is significant.

For examining variation within and across geographical zones, we grouped the 23 populations into 4 zones: North, South, East and Central (see Table 1). AMOVA results indicated that 90% of the haplotype diversity is attributable to differences between individuals within populations, 5.6% to interpopulational differences within zones and 4.3% to zonal differences. The values of the F-statistics were: $F_{CT} = 0.0427$ (non-significant), $F_{SC} = 0.0995$ (non-significant) and $F_{ST} = 0.0995$ (non-significant).

In the AMOVA analysis for examining variation within and across language boundaries, we grouped the 23 populations as: Indo-Aryan, Dravidian, Austro-Asiatic and Tibeto-Burman. We found that 90% of the haplotype diversity is attributable to differences between individuals within populations, 6.9% to interpopulational differences within linguistic groups and 2.8% to language-group differences. The values of the F-statistics were: $F_{CT} = 0.0279$ (statistically non-significant), $F_{SC} = 0.0707$ (significant) and $F_{ST} = 0.0966$ (non-significant).

The above results show that most of the mtDNA diversity observed in Indian populations is between individuals within populations; there is no significant structuring of haplotype diversity by socio-religious affiliation, geographical location of habitat or linguistic affiliation.

Table 4. Frequencies (%) of various haplotypes in 23 ethnic populations of India

Population code	Haplogroup			
	A	*D*	*M*	*U*
AGH			18 (75.0)	5 (20.8)
AMB			27 (90.0)	2 (6.67)
BAG			19 (61.3)	4 (12.9)
BRA-UP	1 (3.7)	2 (7.4)	5 (18.5)	7 (25.9)
BRA-WB			13 (59.1)	1 (4.5)
CHA		2 (8.0)	16 (64.0)	8 (32.0)
GAU			6 (46.2)	2 (15.3)
ILA			16 (53.3)	7 (23.3)
IYN			15 (50.0)	3 (10.0)
IYR	1 (3.3)		15(50.0)	7 (23.3)
KOT			29 (96.7)	
KUR	1 (3.3)		23 (76.7)	2 (6.7)
LOD			26 (81.3)	6 (18.7)
MAH	3 (9.1)		15 (45.5)	7 (21.2)
MUN		1 (14.3)	5 (71.4)	
MUR			24 (80.0)	2 (6.7)
MUS	4 (14.3)	2 (7.1)	10 (35.7)	9 (32.2)
PLR			22 (73.3)	2 (6.7)
RAJ	3 (5.9)	3 (5.9)	16 (31.4)	8 (15.7)
SAN			14 (70.0)	2 (10.0)
TAN			13 (81.3)	2 (12.5)
TRI	4 (8.9)	6 (13.3)	23 (51.1)	3 (6.7)
VAN			16 (53.3)	3 (10.0)
Total	17 (2.6)	16 (2.5)	386 (59.9)	92 (14.3)

Haplogroup frequencies

Based on co-occurrence of mutations, mtDNA molecules have been classified into several haplogroups[28]. The frequencies of these haplogroups show significant geographical variation, as also variation across major human morphological groups[29] (Caucasoid, Mongoloid, etc.). The RFLP sites examined by us permit classification of our data into the following haplogroups: *A* defined by the presence of HaeIII (663) site, *B* defined by the presence of COII/tRNA^{Lys} intergenic 9-bp deletion and the presence of HaeIII (16517) site, *D* defined by the absence of the AluI (5176) site, *M* defined by the simultaneous presence of DdeI (10394) and AluI (10397) sites, and *U* defined by the absence of the DdeI (10394) site and the presence of the HinfI (12308) site.

Table 4 gives the frequencies of these haplogroups in the 23 study populations. Of particular interest are the frequencies of haplogroups *M* and *U*. Haplogroup *M* has been proposed to be an ancient east-Asian marker[6] and is virtually absent among African and Caucasoid populations[28,30,31]. However, Quintana-Murci *et al.*[32] have proposed that the origin of haplogroup *M* is in Africa, in view of its high frequency in Ethiopia. Haplogroup *U* is found in high frequencies among Caucasoid populations, making it suitable for identifying Caucasoid admixture in Indian populations. Figure 3 presents the frequencies of +/+, +/–

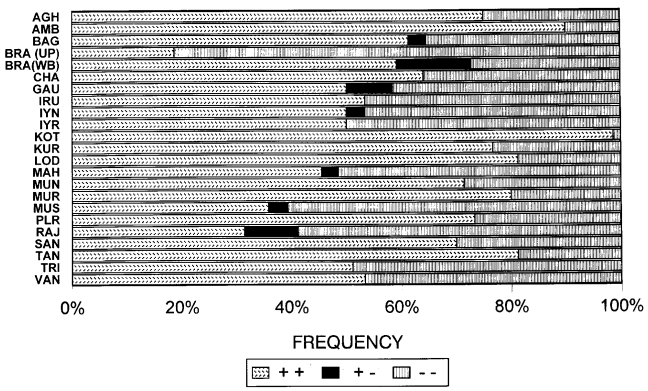

Figure 3. Haplotype frequencies at DdeI (10394) and AluI (10397) mtDNA RFLP loci in 23 ethnic populations of India.

and –/– mtDNA molecules when the sites DdeI (10394) and AluI (10397) are jointly considered. From this figure, we find that the frequency of haplogroup *M* in Indian populations is very high (overall: 59.9%; range: 18.5% [Brahmin of Uttar Pradesh] to 96.7% [Kota]). The +/– haplotype is very common in Africa and the –/– haplotype is very common among Caucasoids. The +/– haplotype is, however, absent or rather infrequent in Indian populations. On the other hand, the –/– haplotype is quite frequent in most Indian populations. The haplogroup *M* frequency is higher among tribal populations than among caste populations; among caste populations there is a decline in frequency with increase in social status of the population in caste hierarchy. Since the +/– haplotype is absent or infrequent, there is a reversal of trend of the –/– haplotype frequency. Further, the frequency of the +/+ haplotype is higher in southern India than in northern India. Since the +/+ haplotype is the most frequent in most populations, one can infer that this is the most ancient of the three haplotypes +/+, +/– and –/–. If this is indeed true, then it is also expected that the diversity among individuals carrying the +/+ haplotype will be higher than among individuals carrying the +/– or the –/– haplotypes. To test this, we computed the haplotype diversities (based on frequencies of haplotypes constructed on the basis of the remaining 5 polymorphic loci) among individuals belonging to these three classes. The results are presented in Figure 4. Contrary to expectations, it is seen that the haplotype diversity among individuals belonging to haplogroup *M* (+/+) is the lowest and that among –/– individuals is the highest. If indeed the +/+ haplogroup is the most ancient, then this finding is probably indicative of strong founder effects (a small group of individuals primarily belonging to haplogroup *M* founding an ancient Indian population, which later expanded and subdivided into multiple populations) or that multiple waves of more recent admixture with individuals of diverse ethnic, but predominantly Caucasoid (–/–), backgrounds or both. In view of the fact that the –/– haplotype is more frequent among most populations of northern India (Brahmins of Uttar Pradesh, Rajput, Muslim) who probably have had more admixture with individuals of Caucasoid background through multiple invasions of India in historical times than populations of the eastern or southern region, leads us to believe that these findings are more probably a result of multiple waves of admixture.

This inference is also bolstered by the fact that the frequency of haplogroup *U*, found predominantly among Caucasoids, is higher among populations of northern India (Muslim, Chamar, Brahmins of Uttar Pradesh) than among southern Indian populations (Figure 5). However, it is striking that the Austro-Asiatic speaking tribal populations of India (Lodha, Santal) also harbour this haplogroup in fairly high frequencies. The overall frequency of haplogroup *U* is 14.28%.

Haplogroup *B* is not present in any of the study populations. Haplogroups *A* and *D* occur with intermediate frequencies (about 5 to 10%) in Indian populations.

A median network[23] was constructed to investigate the relationships among the haplotypes. The network diagram is provided in Figure 6. It is seen from this Figure that there is only one dominant haplotype (# 6) and that the entire network is connected and compact. Any haplotype can be derived by single-step mutation from one of the pre-existing haplotypes.

We have also compared the distributions of haplotypes found in the populations included in the present study with those of other populations of south-east Asia. For this purpose, we collated and compacted the haplotype data presented in Ballinger *et al.*[6]. Since Ballinger *et al.*[6] did not study the RFLP site at nt 12308 studied by us, this locus had to be excluded for purposes of comparison. The results are presented in Figure 7, which shows that there is considerable sharing of haplotypes between Indian and south-east Asian populations. The distributions of haplotype frequencies are also similar. There is, however, one notable difference. The south-east Asian populations harbour a set of haplotypes, albeit with low or medium fre-

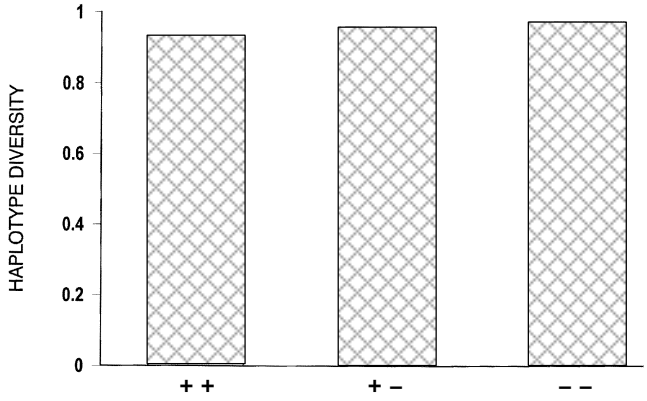

Figure 4. Diversities within three haplotype classes defined by DdeI (10394) and AluI (10397) mtDNA restriction sites in Indian populations.

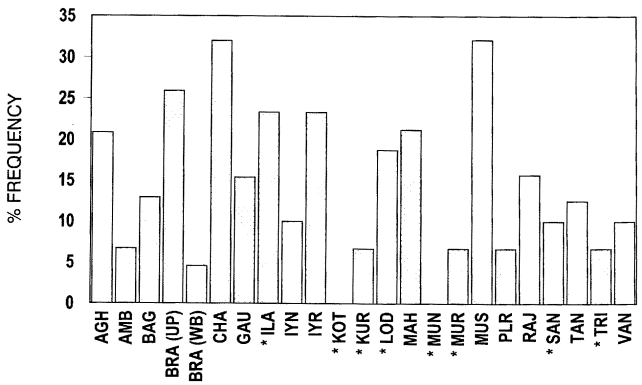

Figure 5. Percentage frequencies of haplogroup *U* in 23 ethnic populations of India.

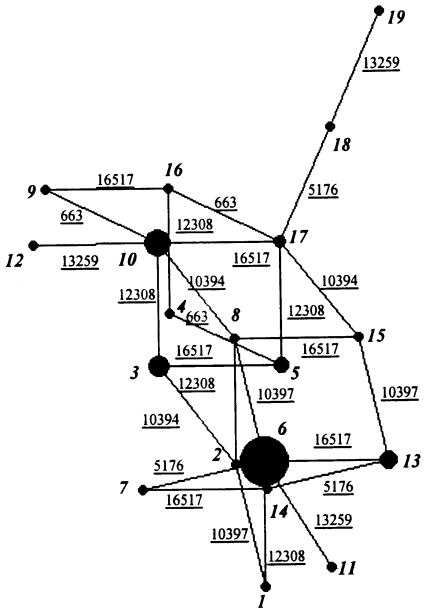

Figure 6. Median network of mtDNA haplotypes in Indian populations. (Note: Boldfaced numbers on nodes are haplotype numbers; see Table 3. Sizes of nodes are proportional to haplotype frequencies. Numbers on edges connecting nodes represent the site number at which the corresponding haplotypes differ.)

quencies, on the 9-bp deletion background, which is completely absent in the present study populations.

To examine the genomic affinities among the 23 population groups, we constructed a neighbour-joining tree using Nei's D_A distance based on haplotype frequencies. The tree is presented in Figure 8, from which it is seen that there is no clear clustering of populations either by geographical proximity of habitat or by linguistic similarity or by social contiguity.

Discussion

Ballinger *et al.*[6] proposed that the COII/tRNA^{Lys} 9-bp deletion originated in central China and spread to the south-east Asian populations and to coastal and island populations of the Pacific. None of the Indian populations included in this study was found to possess this deletion. Thus, the possibility[5] of modern human migration from central/southern China into India seems small. On the other hand, since the African populations also do not generally possess the 9-bp deletion, our finding is consistent with an early migration of modern humans from Africa to India. It may, however, be mentioned that Watkins *et al.*[33]

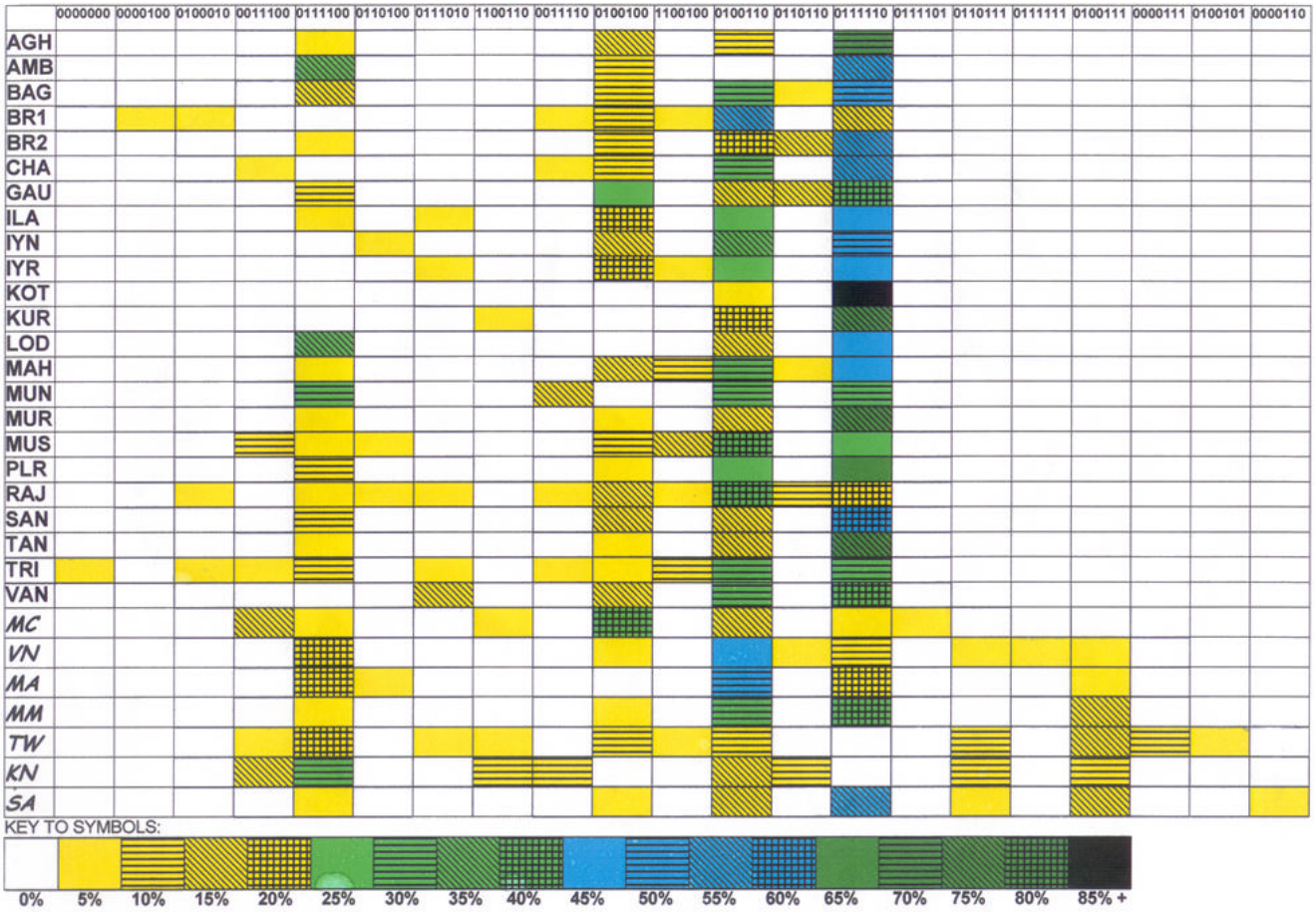

Figure 7. mtDNA haplotype frequencies in 23 ethnic populations of India (top 23 rows) and 7 south-east Asian populations (last 7 rows; MC = Malayan Chinese, VN = Vietnamese; MA = Malayan aborigines, MM = Malays, TW = Taiwanese, KN = Koreans, SA = Sabah aborigines). Order of loci: HaeIII (663), AluI (5176), DdeI (10394), AluI (10397), HincII (13259), HaeIII (16517), 9-bp deletion; 1 = Presence, 0 = Absent.

found high frequencies of the 9-bp deletion in several tribal groups of India. One of these groups, Irula, among whom Watkins *et al.*[33] estimated the frequency of 9-bp deletion to be 44%, has also been included in the present study. We did not find any 9-bp deletion, even though the sample sizes of both these studies were approximately the same. (To exclude the possibility of errors in genotyping performed by polyacrylamide gel electrophoresis of PCR-amplified samples, we carried out DNA sequencing of a subset of our samples, but did not find any genotyping error.) We are unable to offer any clear explanation of this discrepancy of estimated frequencies, except to state that we have sampled the Irulas from their original habitat, Nilgiri Hills, while the samples of Watkins *et al.*[33] were drawn from the coastal area of Andhra Pradesh, where the Irulas are presumably migrants.

The DdeI(10394)-AluI(10397) (+ +) haplotype, haplogroup *M*, which was recognized to be an ancient east Asian marker[6], was subsequently hypothesized[30,31] to have arisen before the split between proto-Indians and proto-Orientals and almost certainly predated the invasion of India by Indo-Aryan speakers. We have now shown that indeed haplogroup *M* occurs with a high frequency, averaging about 60%, across most Indian population groups, irrespective of geographical location of habitat. We have also shown that the tribal populations have higher frequencies of haplogroup *M* than caste populations. As the tribal populations are accepted by most anthropologists to be the indigenous populations, this confirms the hypothesis that haplogroup *M* is an ancient haplogroup in India. In those populations (primarily the caste populations) in which haplogroup *M* is less frequent, the frequency of haplogroup *U* is higher, perhaps indicating a higher extent of Caucasoid admixture. We have also argued that the higher level of diversity observed among individuals with a haplogroup *U* background, compared to those with a haplogroup *M* background, is probably because of multiple waves of immigration of Caucasoid peoples, originating from a large geographical area, into India over a relatively long period of historical time. Quintana-Murci *et al.*[32] have proposed that haplogroup *M* originated in eastern Africa approximately 60,000 ybp, and that it was carried into India through an eastern African exit route by an early dispersal event of modern humans out of Africa. However, we have found haplogroup *M* to be ubiquitous in India, although its frequency is somewhat higher in southern Indian populations than in northern Indian populations. Our findings, therefore, do not rule out the possibility of haplogroup *M* arising in India and being carried to Ethiopia from India.

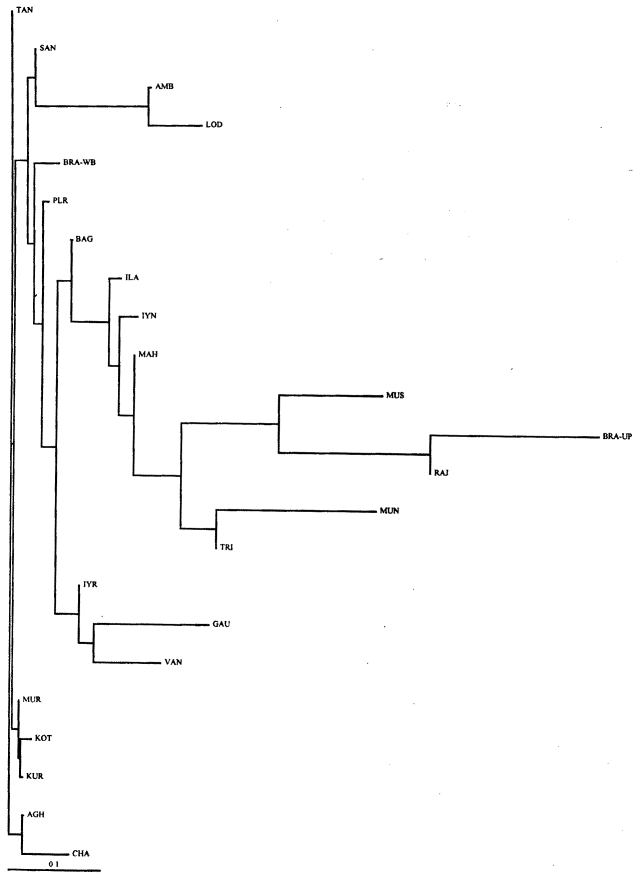

Figure 8. Neighbour-joining tree representing affinities among 23 ethnic populations of India based on mtDNA haplotype frequencies.

The results of comparison of our haplotype data across populations within India and with south-east Asian populations reveal several interesting features. First, we note that there is extensive sharing of one or two haplotypes across population groups within India, irrespective of their geographical location of habitat, linguistic affinity or social proximity. This reveals a fundamental unity of mtDNA lineages in India, in spite of the extensive cultural and linguistic diversity. We, therefore, propose that there was a relatively small founding group of females in India; ethnic differentiation took place subsequently through a series of demographic expansions, geographic dispersal and social groupings. 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. This further supports the hypothesis of a small female founding group in India. Our finding of the lack of correspondence of clusters based on mtDNA haplotype frequencies with either geographical location of habitat, language or social proximity is also consistent with our proposed model of peopling of India. It may be mentioned that our present observation of extensive sharing of mtDNA haplotypes across populations is in contrast to the finding of our earlier study[34] using Y-chromosomal DNA markers, in which we have provided evidence of extensive haplotype variation within ethnic populations of India, with virtually no sharing of haplotypes between populations. Thus, our observations on mtDNA haplotype sharing are

also compatible with the hypothesis that irrespective of the size of the female founding lineages, subsequent to population differentiation, there has been so much admixture across populations that there has been a homogenization of female lineages. While this hypothesis cannot be ruled out on the basis of our present data, the possibility of large-scale admixture across ethnic boundaries is unlikely. Our present findings complement some of the earlier findings from India based on samples drawn from a restricted geographical area[35] or of mixed backgrounds[36]. Also, among the south-east Asian populations, several haplotypes possessed the 9-bp deletion, most of which were on a DdeI(10394)-AluI(10397) –/– background. Ballinger *et al.*[6] have hypothesized that the 9-bp deletion arose in central China and radiated out from this region as migrants moved to populate parts of south-east Asia. If indeed India was also populated by migrants radiating out from central China, then one would have expected that a significant proportion of these migrants would carry the (–/– 9-bp-del) haplotype and 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[19]. On the other hand, a significant proportion of the south-east 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 south-east Asian populations (see Figure 7), one of which is completely absent in Indian populations, leads us to believe that south-east 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[37] and Japanese[16,38] – populations that are postulated to have arisen from human migrations originating from southern China. It is known[9,15] that there were two waves of human migration from mainland Asia through south-east 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 to 3500 ybp. Although we cannot be certain, we postulate that the early wave of migration was from India which carried the +/+ haplotype into south-east Asia. The second wave of migration from south China carried the (–/– 9-bp-del) haplotype into this region. An early wave of migration from India, actually from Africa through India, to south-east Asia has also been proposed in a recent study[39] using nuclear DNA microsatellite markers, which was subsequently supported by a study using Y-chromosomal DNA markers[40].

1. Clark, J. D., in *The Human Revolution: Behavioral and Biological Perspectives on the Origin of Modern Humans* (eds Mellars, P. and Stringer, C.), Edinburgh University Press, Edinburgh, 1989, pp. 565–588.
2. Misra, V. N., *Man Environ.*, 1989, **14**, 17–64.
3. Majumder, P. P., Roy, B., Banerjee, S., Chakraborty, M., Mukherjee, N., Roy, M., Guha Thakurta, P. and Sil, S. K., *Eur. J. Hum. Genet.*, 1999, **7**, 435–446.
4. Mountain, J. L., Hebert, J. M., Bhattacharyya, S., Underhill, P. A., Ottolenghi, C., Gadgil, M. and Cavalli-Sforza, L. L., *Am. J. Hum. Genet.*, 1995, **56**, 979–992.
5. Crow, T. J., in Abstracts of Contributions to the Dual Congress 1998, University of Witwatersrand Medical School, Johannesburg, 1998, p. 109
6. Ballinger, S. W., Schurr, T. G., Torroni, A., Gan, Y. Y., Hodge, J. A., Hassan, K., Chen, K-H. and Wallace, D. C., *Genetics*, 1992, **130**, 139–152.
7. Majumder, P. P., *Evol. Anthropol.*, 1998, **6**, 100–110.
8. Karve, I., *Hindu Society – An Interpretation.* Deshmukh Prakashan, Poona, 1968.
9. Beteille, A., in *The Indian Human Heritage* (eds Balasubramanian, D. and Rao, N. A.), Universities Press, Hyderabad, 1998, pp. 87–94.
10. Meenakshi, K., in *Recent Perspectives of Early Indian History* (ed. Thapar, R.), Popular Prakashan, Mumbai, 1995, pp. 53–79.
11. Ratnagar, S., in *Recent Perspectives of Early Indian History* (ed. Thapar, R.), Popular Prakashan, Mumbai, 1995, pp. 1–52.
12. Thapar, R., in *Recent Perspectives of Early Indian History* (ed. Thapar, R.), Popular Prakashan, Mumbai, 1995, pp. 80–141.
13. Pattanayak, D. P., in *The Indian Human Heritage* (eds Balasubramanian, D. and Rao, N. A.), Universities Press, Hyderabad, 1998, pp. 95–99.
14. Gadgil, M., Joshi, N. V., Manoharan, S., Patil, S. and Prasad, U. V. S., in *The Indian Human Heritage* (eds Balasubramanian, D. and Rao, N. A.), Universities Press, Hyderabad, 1998, pp. 100–129.
15. Diamond, J., *Guns, Germs and Steel: The Fates of Human Societies*, Jonathan Cape, London, 1997.
16. Cann, R. L., Stoneking, M. and Wilson, A. C., *Nature*, 1987, **325**, 31–36.
17. Sherry, S. T., Rogers, A. R., Harpending, H., Soodyall, H., Jenkins, T. and Stoneking, M., *Hum. Biol.*, 1994, **66**, 761–775.
18. Miller, S. A., Dykes, D. D. and Polesky, H. F., *Nucleic Acids Res.*, 1988, **16**, 1215.
19. Rickards, O., *Hum. Evol.*, 1995, **10**, 5–16.
20. Nei, M., *Molecular Evolutionary Genetics*, Columbia University Press, New York, 1987.
21. Nei, M., *Proc. Natl. Acad. Sci. USA*, 1973, **70**, 3321–3323.
22. Schneider, S., Kueffer, J-M., Roessli, D. and Excoffier, L., *Arlequin ver. 1.1: A Software for Population Genetic Data Analysis*, Genetics and Biometry Laboratory, University of Geneva, Switzerland, 1997.
23. Bandelt, H-J., Forster, P., Sykes, B. C. and Richards, M. B., *Genetics*, 1995, **141**, 743–753.
24. Saitou, N. and Nei, M., *Mol. Biol. Evol.*, 1987, **4**, 406–425.
25. Excoffier, L., Smouse, P. and Quattro, J., *Genetics*, 1992, **131**, 479–491.
26. Cockerham, C. C., *Evolution*, 1969, **23**, 72–83.
27. Cockerham, C. C., *Genetics*, 1973, **74**, 679–700.
28. Torroni, A., Lott, M. T., Cabell, M. F., Chen, Y-S., Lavergne, L. and Wallace, D. C., *Am. J. Hum. Genet.*, 1994, **55**, 760–776.
29. Chen, Y-S., Torroni, A., Excoffier, L., Santachiara-Benerecetti, A. S. and Wallace, D. C., *Am. J. Hum. Genet.*, 1995, **57**, 133–149.
30. Passarino, G., Semino, O., Modiano, G., Bernini, L. F. and Santachiara-Benerecetti, A. S., *Ann. Hum. Biol.*, 1996a, **23**, 121–126.
31. Passarino, G., Semino, O., Bernini, L. F. and Santachiara-Benerecetti, A. S., *Am. J. Hum. Genet.*, 1996b, **59**, 927–934.
32. Quintana-Murci, L., Semino, O., Bandelt, H-J., Passarino, G., McElreavey, K. and Santachiara-Benerecetti, A. S., *Nature Genet.*, 1999, **23**, 437–441.
33. Watkins, W. S., Bamshad, M., Dixon, M. E., Rao, B. B., Naidu,

J. M., Reddy, P. G. and Prasad, B. V. R., *Am. J. Phys. Anthropol.*, 1999, **109**, 147–158.
34. Bhattacharyya, N. P., Basu, P., Das, M., Pramanik, S., Banerjee, R., Roy, B., Roychoudhury, S. and Majumder, P. P., *Genome Res.*, 1999, **9**, 711–719.
35. Bamshad, M., Fraley, A. E., Crawford, M. H., Cann, R. L., Busi, B. R., Naidu, J. M. and Jorde, L. B., *Hum. Biol.*, 1996, **68**, 1–28.
36. Barnabas, S., Apte, R. V. and Suresh, C. G., *Ann. Hum. Genet.*, 1996, **60**, 409–422.
37. Passarino, G., Semino, O., Modiano, G., Santachiara-Benerecetti, A. S. and Wallace, D. C., *Am. J. Hum. Genet.*, 1993, **53**, 609–618.
38. Horai, S. and Matsunaga, E., *Hum. Genet.*, 1986, **68**, 324–332.
39. Chu, J. Y., Huang, W., Kuang, S. Q., Wang, J. M., Xu, J. J., Chu, Z. T. and Yang, Z. Q., *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 11763–11768.
40. Su, B., Xiao, J., Underhill, P., Deka, R., Zhang, W., Akey, J. and Huang, W., *Am. J. Hum. Genet.*, 1999, **65**, 1718–1724.
41. Ghosh, A. K., *Hum. Hered.*, 1973, **23**, 78–82.
42. Mehta, B. H., *Gonds of Central Indian Highlands: A Study of the Dynamics of Gond Society*, Concept Publishing Co, New Delhi, 1984.
43. Rizvi, S. H. M. and Roy, S., *Muslims: Bio-cultural Perspective*, B.R. Publishers, New Delhi, p. 84.
44. Sen, R., *Introducing Malsom*, Tribal Research Institute, Agartala, 1996.
45. Deb Barman, D., *Treatise on Traditional Social Institutions of Tripuri Community*, Directorate of Research, Department of Welfare for Scheduled Tribes and Scheduled Castes, Agartala, 1983.

ACKNOWLEDGEMENTS. We are grateful to Antonio Torroni, Andy Merriwether and Chris Tyler-Smith for suggestions at various stages of this work. This study has been partially supported by a grant from the Department of Biotechnology, Government of India.

Received 21 July 2000; accepted 19 August 2000

Multiple criteria decision making: Assigning teachers – an example

S. B. Siyambalapitiya
Department of Engineering Mathematics, Faculty of Engineering, University of Peradeniya, Peradeniya, Sri Lanka

Assignment problems when more than one criterion are to be used in assigning personnel have drawn little attention in the literature, even though there are important practical problems coming under this category. This paper addresses a problem of assigning personnel giving due consideration to preferences and other relevant criteria. An algorithm to satisfactorily solve a problem encountered in the educational sector in Sri Lanka is presented.

SOME areas in a country may be considered as congenial while others as difficult because of lack of certain basic facilities. People prefer to serve in congenial areas due to various reasons. It is assumed that in congenial areas health, educational, transportation, housing and entertainment facilities are better. Also sometimes certain areas become difficult due to adverse climatic conditions alone. Normally congenial areas are assumed to have a higher population density than in uncongenial or difficult areas. This is especially true in developing countries where resources are limited in order to develop all the regions with equal infrastructure facilities.

As a result, more and more educated people originate from congenial areas because the educational facilities in these areas are better compared to those in uncongenial areas. It is also assumed that the majority of the professionals prefer to work in congenial areas. Here we take the teaching profession as an example. Let us assume that a certain number of candidates have been selected to be employed as probationary teachers by the relevant authorities. They will have to be posted to schools in different areas of a country. It is also assumed that, a country is divided into several educational regions or districts. It appears that there is a natural tendency for a majority of them to prefer appointments in places closer to their home towns or in congenial areas. Sometimes, even the candidates coming from difficult areas seem to prefer to work in congenial areas due to the availability of certain facilities which enable them for advancement in their careers. We assume that the prospective candidates are given the opportunity to indicate their preferences for districts in their application forms.

However, it has been found that it is almost impossible to heed to these requests, as too many candidates are opting to serve in congenial areas. Usually, the number of qualified candidates hailing from congenial areas is more than the number of vacancies existing in these areas. On the other hand, very few qualified candidates originate from uncongenial areas, far short of those required in