High-resolution analysis of Y-chromosomal polymorphisms reveals signatures of population movements from Central Asia and West Asia into India

NAMITA MUKHERJEE[1], ALMUT NEBEL[2], ARIELLA OPPENHEIM[2] and PARTHA P. MAJUMDER[1]*

[1]*Anthropology and Human Genetics Unit, Indian Statistical Institute, B.T. Road, Kolkata 700 108, India*
[2]*Hebrew University–Hadassah Medical School, Jerusalem, Israel 91120*

Abstract

Linguistic evidence suggests that West Asia and Central Asia have been the two major geographical sources of genes in the contemporary Indian gene pool. To test the nature and extent of similarities in the gene pools of these regions we have collected DNA samples from four ethnic populations of northern India, and have screened these samples for a set of 18 Y-chromosome polymorphic markers (12 unique event polymorphisms and six short tandem repeats). These data from Indian populations have been analysed in conjunction with published data from several West Asian and Central Asian populations. Our analyses have revealed traces of population movement from Central Asia and West Asia into India. Two haplogroups, HG-3 and HG-9, which are known to have arisen in the Central Asian region, are found in reasonably high frequencies (41.7% and 14.3% respectively) in the study populations. The ages estimated for these two haplogroups are less in the Indian populations than those estimated from data on Middle Eastern populations. A neighbour-joining tree based on Y-haplogroup frequencies shows that the North Indians are genetically placed between the West Asian and Central Asian populations. This is consistent with gene flow from West Asia and Central Asia into India.

[Mukherjee N., Nebel A., Oppenheim A. and Majumder P. P. 2001 High-resolution analysis of Y-chromosomal polymorphisms reveals signatures of population movements from Central Asia and West Asia into India. *J. Genet.* **80**, 125–135]

Introduction

Single-nucleotide and insertion/deletion polymorphisms in the nonrecombining portion of the human Y chromosome have been termed as unique event polymorphisms (UEP). These UEPs define related groups of chromosomes, termed haplogroups. Together with rapidly evolving short tandem repeat (STR) markers, Y chromosome polymorphisms have proved to be a very powerful tool in tracing movements of males in human population history (Zerjal *et al*. 1997; Casalotti *et al*. 1999; Hill *et al*. 2000; Stumpf and Goldstein 2001).

India occupies a centrestage in human evolution because one of the early waves of migration of modern humans from out of Africa, through West Asia, was into India (Cann 2001). More recently, about 15,000–10,000 years before present (ybp), when agriculture developed in the Fertile Crescent region that extends from Israel through northern Syria to western Iran, there was another eastward wave of human migration (Cavalli-Sforza *et al*. 1994; Renfrew 1987), a part of which also appears to have entered India. This wave has been postulated to have brought the Dravidian languages into India (Renfrew 1987). Subsequently, the Indo-European (Aryan) language family was introduced into India about 4,000–3,000 ybp from the Iranian plateau, where this language is thought to have been brought by pastoral nomads from Central

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

Keywords. unique event polymorphism; short tandem repeat; haplotype; haplogroup; genome diversity.

Asian steppes (Renfrew 1990). Therefore, primarily on the basis of linguistic evidence, it appears that West Asia and Central Asia may have been two major geographical sources of genes in the Indian gene pool. We therefore sought to test the nature and extent of similarities in the gene pools of these regions. We have used a set of 18 Y chromosome polymorphisms in this study, of which 12 are UEP and six are STR markers.

Materials and methods

***Population samples*:** We collected blood samples, with consent, from 89 males belonging to four endogamous ethnic populations inhabiting the state of Uttar Pradesh in northern India. The individuals from whom samples were collected were unrelated at least to the first cousin level. The names of the ethnic groups are: Brahmin (BRA), Chamar (CHA), Muslim (MUS) and Rajput (RAJ). The Brahmin, Rajput and Chamar 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.

We have extensively compared the data generated for these four populations with data published by us earlier on Israeli and Palestinian Arabs sampled from Israel and the Palestinian Authority Area (Nebel *et al*. 2000), and 11 other populations, of which eight (Kurdish Jew, Yemenite Jew, Palestinian, Syrian, Lebanese, Druze, Saudi Arabian and Turk) are from Middle and West Asia (Hammer *et al*. 2000) and three (Armenian, Georgian and Ossetian) are from Central Asia (Rosser *et al*. 2000).

***DNA markers*:** DNA samples were typed using 18 markers, 12 of which were diallelic UEP markers while six were short tandem repeat markers. The 12 UEP markers are YAP, 92r7, SRY4064, sY81, SRY+465, TAT, M9, M13, M17, M20, SRY10831 and p12f2. Primer sequences and amplification protocols for the first 11 DNA markers are described by Thomas *et al*. (1999). The primer sequences for p12f2 (Casanova *et al*. 1985) were 12f2D and 12f2G and the amplification protocol was as described by Rosser *et al*. (2000) to amplify an 88-bp product. As an internal control, a protocol standardized by A. Oppenheim and A. Nebel was used. A product of size 148 bp encompassing the M172 polymorphism was coamplified using the primers M172-F (5′-TCCCCCA AACCCATTTTGATGCAT-3′) and M172-R (5′-GGATC CATCTTCACTCAATGTTG-3′). PCR amplification was carried out in 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 2.5 mM $MgCl_2$, 0.2 mM of each dNTP, 0.2 μM of each p12f2-primer, 0.3 μM of each M172 primer, and 0.2 U of AmpliTaq Gold (Perkin Elmer, Roche Molecular Systems, USA). Cycling conditions were as follows: initial denaturation for 5 min; 30 cycles of denaturation at 94°C for 30 s, annealing at 58°C for 45 s and extension at 72°C for 45 s. The final cycle ended with an additional extension of 10 min at 72°C.

The six STR markers were DYS19, DYS388, DYS390, DYS391, DYS392 and DYS393, all of which were amplified using markers and protocols as described by Thomas *et al*. (1999). Restriction products and the amplified DNA were electrophoresed on an ABI 377 automated DNA sequencer and genotyped by Genescan version 3.1 and Genotyper version 2.1. In this study we have used the haplogroup definitions as given in Rosser *et al*. (2000).

***Statistical analysis*:** Tests of null hypotheses of equality of proportions of various haplogroups across populations were performed by exact tests as implemented in Genepop (Raymond and Rousset 1995; Rousset 2001). Haplogroup diversity was estimated using the method of Nei (1973).

We estimated the STR haplotype diversity within a population or haplogroup by computing the average Euclidean distance (AED) among the haplotypes, defined as

$$\mathrm{AED} = \frac{1}{\binom{N}{2}} \sum_{i<k}^{N} \sqrt{\sum_{j=1}^{6} (l_{ij} - l_{kj})^2},$$

where l_{ij} and l_{ki} are repeat numbers at the jth STR locus for two distinct haplotypes i and k ($i, k = 1, 2, \ldots, N$ = total number of sampled individuals), and $j = 1, 2, \ldots, 6$.

The age (A) of a haplogroup were estimated as: $A = g \times s^2/m$, where g is generation time (assumed to be 30 years); s^2 = the variance of STR repeat number among haplotypes belonging to the haplogroup, averaged over all six STR loci; and m = the mutation rate per generation at an STR locus (taken to be 0.18%, as previously estimated (Quintana-Murci *et al*. 2001) for the six STR loci under consideration). The 95% confidence interval of estimated A was calculated from the previously estimated (Quintana-Murci *et al*. 2001) 95% CI of m = (0.31%–0.098%). A genetic distance matrix (using Cavalli-Sforza and Edwards's chord distance) among the populations was computed and a neighbour-joining tree based on this distance matrix was constructed using Phylip version 3.5 (http://www.evolution.genetics.washington.edu/phylip.html).

Results

UEP haplogroups

On the basis of the UEP markers, we have classified Y chromosomes of each population into haplogroups (HG) as defined by Rosser *et al*. (2000). The haplogroup frequencies are presented in table 1. Because the set of markers screened in this study was not exactly

the same as that screened in published studies (Hammer *et al*. 2000; Nebel *et al*. 2000; Rosser *et al*. 2000), some of the haplogroups could not be resolved and had to be pooled. We have used the evolutionary relationships among the haplogroups to make this pooling meaningful. HG-22 is derived from HG-1 by a C-to-T transition at the SRY-2627 (M-167) locus (Hurles *et al*. 1999). Since in the present study we did not screen the SRY-2627 locus, we pooled the HG-1 and HG-22 frequencies in the Central Asian and West Asian populations. Similarly, HG-16 is a derivative of HG-26 (Zerjal *et al*.1997), and therefore we have pooled the HG-26 and HG-16 frequencies. In spite of this limitation of pooling, it is seen that there is substantial overlap in the types of haplogroups observed in the North Indian and in the Middle Eastern, Central Asian and West Asian regions. There are, however, some differences, the most notable of which is that HG-21 which is observed in substantial frequencies in all the Middle Eastern, Central Asian and West Asian populations is not found in the North Indian populations. In the Middle East, Central Asia and West Asia, HG-7 is generally absent, but it is observed in low frequencies among some populations (e.g. Israeli and Palestinian Arabs). This haplogroup is also absent in North India. The frequencies of haplogroups vary considerably across North Indian or Middle Eastern, Central Asian and West Asian populations. The frequency of HG-3 is the highest in pooled North Indian sample (41.7%), and also in the individual populations. The frequency distributions of haplogroups among the four Indian populations were not statistically significant ($P = 0.059$) at the 5% level. The differences in haplogroup frequencies among the 12 Middle Eastern, Central Asian and West Asian populations were, however, statistically significant at the 5% level.

The most frequent (41.4%) haplogroup in the Middle East, Central Asia and West Asia is HG-9. The phylogenetic relationships among the observed haplogroups and their frequencies in northern India and the Middle East, Central Asia and West Asia are depicted in figure 1(a and b). It is obvious from these figures that the haplogroup frequencies in these two geographical regions are considerably different (the difference is statistically significant at the 5% level).

Table 1. Haplogroup frequencies (%) in four populations of North India and 12 populations of the Middle East, West Asia and Central Asia.

	Haplogroup[a]							
Population	HG-1	HG-2	HG-3	HG-9	HG-21	HG-26	HG-28	Other
North India								
Brahmin (n = 17)	11.8	23.5	35.3	23.5	0.0	0.0	5.9	0.0
Chamar (n = 18)	11.1	44.4	44.4	0.0	0.0	0.0	0.0	0.0
Muslim (n = 19)	15.8	0.0	57.9	10.5	0.0	15.8	0.0	0.0
Rajput (n = 35)	11.4	25.7	37.1	17.1	0.0	2.9	5.7	0.0
Total (n = 89)	13.2	23.1	41.7	14.3	0.0	4.4	3.3	0.0
Middle East, West Asia and Central Asia (Caucasus region)								
Israeli and Palestinian Arab[e] (n = 143)	8.4	6.3	1.4	55.2	20.3	7.0	0.0	1.4[b]
Kurdish Jew[f] (n = 50)	16.0[c]	4.0	4.0	44.0	8.0	24.0[d]	0.0	0.0
Yemenite Jew[f] (n = 30)	27.0[c]	0.0	3.0	43.0	17.0	7.0[d]	0.0	3.0
Palestinian[f] (n = 73)	9.0[c]	5.0	0.0	51.0	19.0	10.0[d]	0.0	5.0
Syrian[f] (n = 91)	10.0[c]	3.0	9.0	57.0	10.0	11.0[d]	0.0	0.0
Lebanese[f] (n = 24)	0.0	13.0	4.0	46.0	29.0	4.0[d]	0.0	4.0
Druze[f] (n = 21)	5.0	0.0	0.0	38.0	19.0	38.0[d]	0.0	0.0
Saudi Arabian[f] (n = 21)	5.0	5.0	19.0	33.0	5.0	29.0[d]	0.0	5.0
Armenian[g] (n = 89)	25.0	31.0	6.0	29.0	3.0	5.0	0.0	0.0
Georgian[g] (n = 64)	19.0	48.0	6.0	23.0	2.0	2.0	0.0	0.0
Ossetian[g] (n = 47)	43.0	11.0	2.0	34.0	6.0	4.0	0.0	0.0
Turks[f] (n = 98)	23.0	12.0	5.0	26.0	6.0	21.0	0.0	6.0
Total (n = 751)	16.4	12.9	4.4	41.4	11.4	11.3	0.0	1.9

[a]Haplogroup definitions are as given in Rosser *et al*. (2000).
[b]All belong to HG-7.
[c]Includes HG-1 and HG-22 (see text for explanation).
[d]Includes HG-26 and HG-16 (see text for explanation).
[e]Nebel *et al*. (2000).
[f]Hammer *et al*.(2000).
[g]Rosser *et al*. (2000).

There are substantial differences in Y haplogroup diversities among the populations (figure 2). Among the Indian populations, the Brahmin and Rajput show high diversities, while the Chamar and the Muslim show much lower diversities. Most of the Middle Eastern, Central Asian and West Asian populations show high diversities. The diversities observed among the Saudi Arabian, Brahmin and Rajput are similar.

STR haplotypes

Considerable variations in allele frequencies were observed at the six STR loci among the four North Indian populations. These allele frequencies are depicted in figure 3. The differences in allele frequencies at DYS19, DYS388, DYS392 and DYS393 among the four populations were statistically significant at the 5% level, while the differences at DYS390 and DYS391 were not statistically significant. The frequency distributions are given in table 2. A total of 43 haplotypes were observed among the 83 individuals who were genotyped at all the six STR loci. It can be observed from table 2 that there are two modal haplotypes, 15-12-22-10-11-12 on HG-2 background and 15-12-25-11-11-13 on HG-3 background, with frequencies 12% and 10% respectively. The remaining haplotypes occur at low frequencies. The four populations carry nearly disjoint sets of haplotypes.

To examine diversities and commonalities of STR haplotypes within and across UEP haplogroups, we classified the STR haplotypes by haplogroups (table 3). In the Indian populations, the haplogroups carry nearly disjoint sets of haplotypes. Further, there is virtually no haplotype sharing between the North Indian populations and the Israeli and Palestinian Arabs (Nebel *et al*. 2000). In fact the two most frequent haplotypes found in northern India (15-12-22-10-11-12 and 15-12-25-11-11-13) are not found among the Israeli and Palestinian Arabs. Similarly, the modal haplotype (14-17-22-11-11-12) found among the Israeli and Palestinian Arabs, and the Cohen (14-16-23-10-11-12) and Galilee (14-17-23-11-11-12) modal haplotypes found among many populations whose ancestries can be traced back to the West Asian region are not present in northern India.

The average Euclidean distance (AED) values of the STR haplotypes among the Brahmin and Rajput are of similar magnitude (3.22 and 3.07, respectively). AEDs for Chamar and Muslim are also similar, but smaller (2.35 and 2.26, respectively). In the North Indian populations, Y chromosomes belonging to HG-1 have the highest AED (1.98), while those belonging to HG-2, HG-3 and HG-9 are smaller and have the same AED (1.65). Among the Israeli and Palestinian Arabs, however, Y chromosomes belonging to various haplogroups have dissimilar AEDs: HG-2, 2.52; HG-26, 2.17; HG-9, 2.12; and HG-1, 1.86.

We have estimated the ages of the various common haplogroups found in the Indian and Arab samples (table 4). It may, however, be noted that although the ages of the common haplogroups HG-1, HG-2 and HG-9 estimated from the data for North Indians and Israeli and Palestinian Arabs are different, the differences are not statistically significant as seen from the 95% confidence intervals.

On the basis of the UEP haplogroup frequencies we have constructed a neighbour-joining tree depicting relationships among the North Indian, Central Asian and West Asian populations (figure 4). It is seen that the North Indian populations cluster together, and this cluster is embedded in the clusters of populations that comprise the Central Asian and the Middle Eastern and West Asian populations. This is consistent with gene flow from the

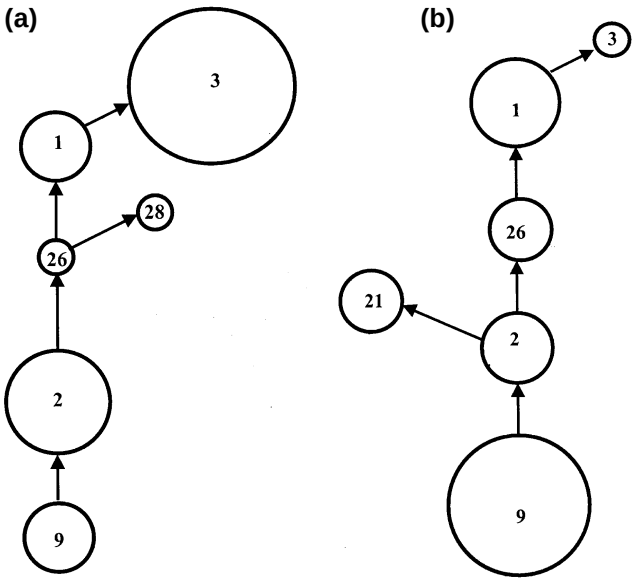

Figure 1. Evolutionary network of Y-chromosomal haplogroups in (a) northern India, and (b) the Middle East, Central Asia and West Asia. The area of each circle is proportional to observed haplogroup frequency.

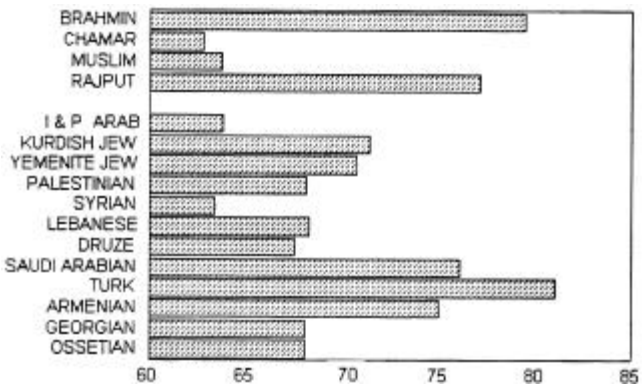

Figure 2. Y-chromosomal haplogroup diversities in four population groups of North India and 12 population groups of the Middle East, Central Asia and West Asia.

Middle East and West Asia, and also from Central Asia, into India.

Discussion

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, there was, possibly, migration of humans from this region into India. More recently, pastoral nomads originating in the Central Asian steppes may also have contributed to the gene pool of India. We therefore sought to find signatures of Middle Eastern and Central Asian gene pools in the gene pool of India. The entry of

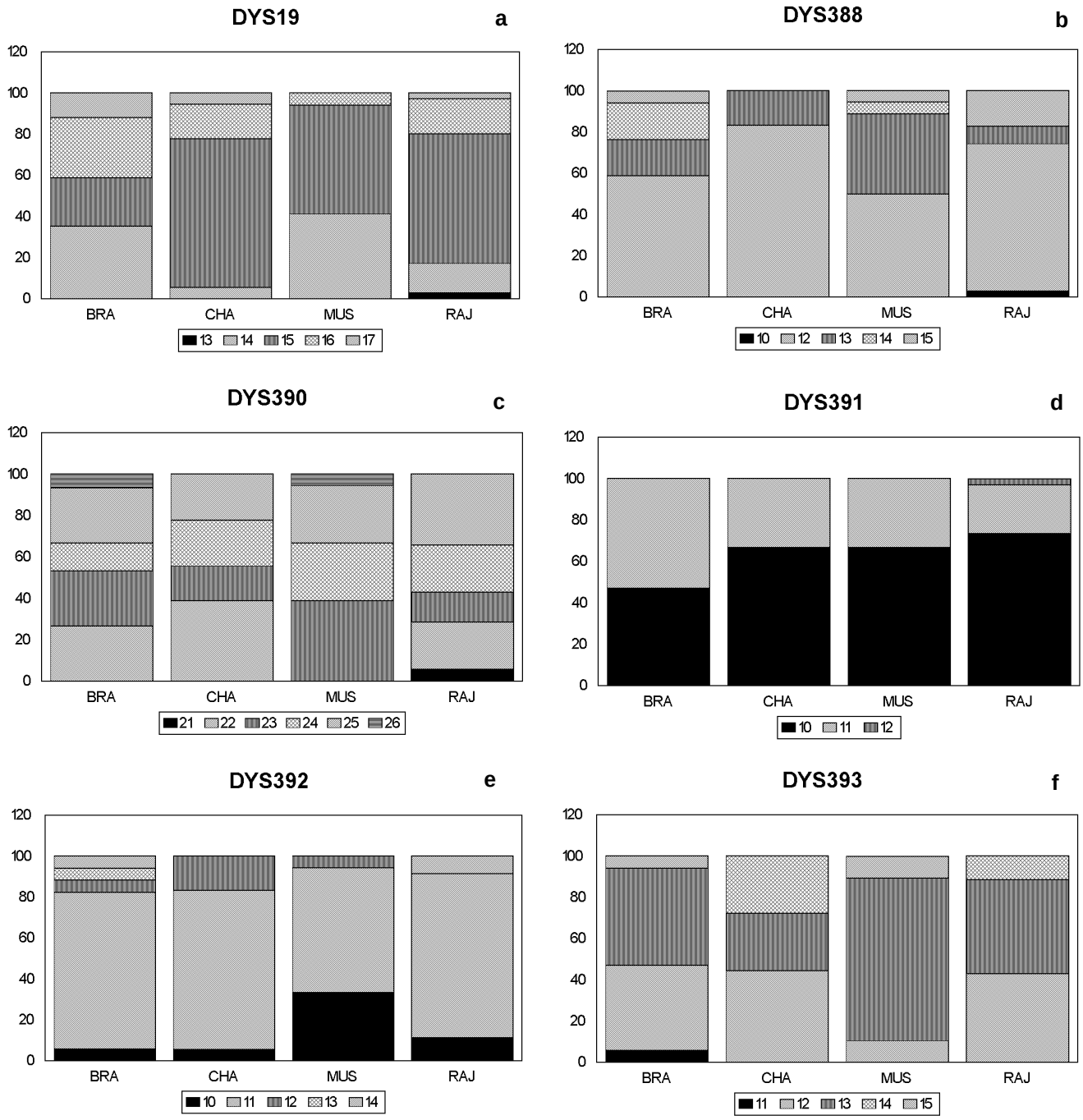

Figure 3. Allele frequencies at six Y-chromosomal STR loci in four population groups of North India.

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 in other parts of India. We have studied four groups inhabiting the northern Indian state of Uttar Pradesh. 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, Central Asian and West Asian population groups (Hammer *et al.* 2000; Nebel 2000; Rosser *et al.* 2000), and performed comparative statistical analyses to draw inferences.

The distributions of Y haplogroups, defined (Rosser *et al.* 2000) on the basis of 12 diallelic UEPs, reveal many interesting patterns. The haplogroup diversities in the populations of North 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

Table 2. Distributions of Y-chromosomal STR haplotypes in four populations of Uttar Pradesh, North India.

Haplotype*	Brahmin	Chamar	Muslim	Rajput	Total
14-12-22-11-14-11	1				1
14-12-25-10-13-13	1				1
14-14-23-11-11-12	3				3
14-15-24-10-11-12	1			2	3
15-12-22-10-11-12	3	4		3	10
15-13-25-10-11-13	1		2	1	4
16-12-23-10-10-15	1				1
16-12-24-10-12-13	1				1
16-12-25-11-11-13	2		1		3
17-13-26-11-11-13	1				1
14-12-23-10-10-14		1		2	3
15-12-22-10-12-12		1			1
15-12-23-10-11-12		1			1
15-12-25-10-11-14		1			1
15-12-25-11-11-13		3		5	8
15-13-22-10-12-12		1			1
15-13-23-10-12-14		1			1
15-13-24-11-11-14		1			1
16-12-22-10-11-12		1		1	2
16-12-24-10-11-14		1			1
16-12-24-11-11-13		1			1
17-12-24-11-11-13		1			1
14-12-23-10-10-15			2		2
14-13-23-10-10-13			4		4
15-12-24-11-11-13			4	1	5
15-12-25-10-11-12			1		1
15-12-26-10-11-13			1		1
15-15-23-10-11-12			1	1	2
13-10-24-10-14-12				1	1
14-13-23-10-10-14				1	1
15-12-21-11-11-12				1	1
15-12-22-10-14-12				2	2
15-12-23-10-10-13				1	1
15-12-24-10-11-13				1	1
15-12-24-10-11-14				1	1
15-12-25-10-11-13				3	3
15-13-22-11-11-13				1	1
16-12-21-10-11-13				1	1
16-12-25-12-11-13				1	1
16-15-24-10-11-12				2	2
16-15-25-10-11-12				1	1
17-12-25-10-11-13				1	1
Total	15	18	16	34	83

*Order of loci: DYS19-DYS388-DYS390-DYS391-DYS392-DYS393.

Table 3. Distributions of Y-chromosomal STR haplotypes in North Indians (I) and Israeli and Palestinian Arabs (A) in individuals belonging to various haplogroups.

	HG-1		HG-2		HG-3		HG-9		HG-26	
Haplotype*	I	A	I	A	I	A	I	A	I	A
14-12-25-10-13-13	1	1								
14-12-23-10-10-14	3									
14-13-23-10-10-13	3								1	
14-13-23-10-10-14	1									
15-12-23-10-10-13	1									
15-13-23-10-12-14	1									
16-12-23-10-10-15	1									
13-12-24-10-14-13		1								
14-12-23-10-10-12		1								
14-12-24-11-13-12		3								
14-12-24-11-14-12		1								
14-12-25-11-13-12		1								
15-12-24-10-13-12		1								
15-12-24-10-13-13		2								
15-12-24-11-13-13		1								
14-12-25-10-13-13	1	1								
14-12-23-10-10-14	3									
15-12-21-11-11-12			1							
15-12-22-10-11-12			10							
15-12-22-10-12-12			1							
15-12-24-10-11-14			1							
15-13-22-10-12-12			1							
15-13-22-11-11-13			1							
15-13-24-11-11-14			1							
15-13-25-10-11-13			2		2					
16-12-22-10-11-12			2							
14-12-22-11-11-14				1						
15-10-23-10-12-13				2						
15-12-22-11-11-14				2						
15-13-23-10-12-12				1						
15-13-25-10-11-12				1						
16-12-22-10-11-14				2						
16-12-23-10-11-12				1						
16-13-23-10-12-14				1						
16-13-14-11-11-13				1						
16-12-25-11-11-13					3	1				
15-12-23-10-11-12					1					
15-12-24-10-11-13					1					
15-12-24-11-11-13					5					
15-12-25-10-11-12					1					
15-12-25-10-11-13					3					
15-12-25-10-11-14					1					
15-12-25-11-11-13					8					
15-12-26-10-11-13					1					
16-12-21-10-11-13					1					
16-12-24-10-11-14					1					
16-12-24-10-12-13					1					
16-12-24-11-11-13					1					
16-12-25-12-11-13					1					
17-12-24-11-11-13					1					
17-12-25-10-11-13					1					
17-13-26-11-11-13					1					
15-12-25-11-11-12						1				
16-12-25-11-11-13					3	1				
15-12-23-10-11-12					1					
14-14-23-11-11-12							3	1		
14-15-24-10-11-12							3			
15-15-23-10-11-12							2			

(**Table 3.** *continued*)

	HG-1		HG-2		HG-3		HG-9		HG-26	
Haplotype*	I	A	I	A	I	A	I	A	I	A
16-15-24-10-11-12							2			
16-15-25-10-11-12							1			
14-14-23-10-11-12								1		
14-14-24-10-11-12								1		
14-14-25-11-11-12								1		
14-15-22-9-11-12								1		
14-15-22-11-11-13								2		
14-15-23-10-11-12								2		
14-15-23-10-11-13								2		
14-15-23-10-11-14								1		
14-15-26-10-11-12								2		
14-16-22-10-11-12								1		
14-16-22-11-11-12								1		
14-16-23-9-11-16								1		
14-16-23-10-11-12								3		
14-17-22-10-11-12								3		
14-17-22-11-11-12								20		
14-17-22-11-12-12								1		
14-17-22-12-11-13								1		
14-17-23-10-11-12								3		
14-17-23-11-11-12								12		
14-17-23-11-11-11								1		
14-17-23-13-11-12								1		
14-17-24-11-11-12								1		
15-15-22-10-8-12								1		
15-15-24-10-11-12								5		
15-15-24-11-11-12								1		
15-16-24-10-11-12								1		
15-17-22-11-11-12								1		
15-17-23-11-11-12								1		
15-18-23-12-11-13								1		
16-14-24-10-11-12								1		
16-15-23-10-11-12								1		
13-10-24-10-14-12									1	
14-12-23-10-10-15									2	
13-12-24-10-13-13										2
14-12-22-10-13-13										1
14-12-22-10-14-11										1
14-12-23-10-14-12										1
14-12-23-10-14-13										1
15-12-23-9-13-13										1
15-12-23-10-13-13										1
15-12-23-11-13-15										1
15-12-24-10-14-13										1

*Order of loci: DYS19-DYS388-DYS390-DYS391-DYS392-DYS393.

populations, the differences in the frequency distributions of haplogroups are not statistically significant, but the 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 (23%–57%) in Middle Eastern, West Asian and Central Asian 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 5500–17,400 ybp (Hammer *et al*. 2000; Quintana-Murci *et al*. 2001) in this region (southwestern Iran). Our estimate (table 4) of the age of this haplogroup from data on the Middle Eastern populations is in fair agreement with the earlier estimate. As noted in an earlier 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 Brahmin 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 Rajput either through admixture with Brahmins or directly with the Aryan-speaking immigrants. Since historians (Thapar 1975) have noted that some of the Central Asian pastoral nomads are ancestors of Rajputs, it is more likely that this haplogroup (HG-9) was introduced into the Rajputs directly by the Central Asians than indirectly through admixture with the Brahmins. 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-ranked 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.

HG-3, which is the most frequent haplogroup in India, is known to be widely found in Asia, except East 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 7500 ybp (Karafet *et al*.1999; Zerjal *et al*. 1999; Rosser *et al*. 2000)**,** and the distribution of this haplogroup reflects a recent and major population expansion within Eurasia. HG-3 shows a decreasing frequency cline from Central Asia westward into Europe (Rosser *et al*. 2000) and from Iran towards India (Quintana-

Table 4. Estimated ages of common Y-chromosomal haplogroups found among North Indians and Israeli and Palestinian Arabs.

Population	Haplogroup	Estimated age in years (95% CI)
North Indians	HG-1	7200 (4200–13,160)
	HG-2	5800 (3400–10,700)
	HG-3	5200 (3000–9500)
	HG-9	5200 (3000–9500)
Israeli and Palestinian Arabs	HG-1	6000 (3500–11,000)
	HG-2	10,000 (5800–18,400)
	HG-9	8300 (4800–15,300)
	HG-26	7300 (4250–13,400)

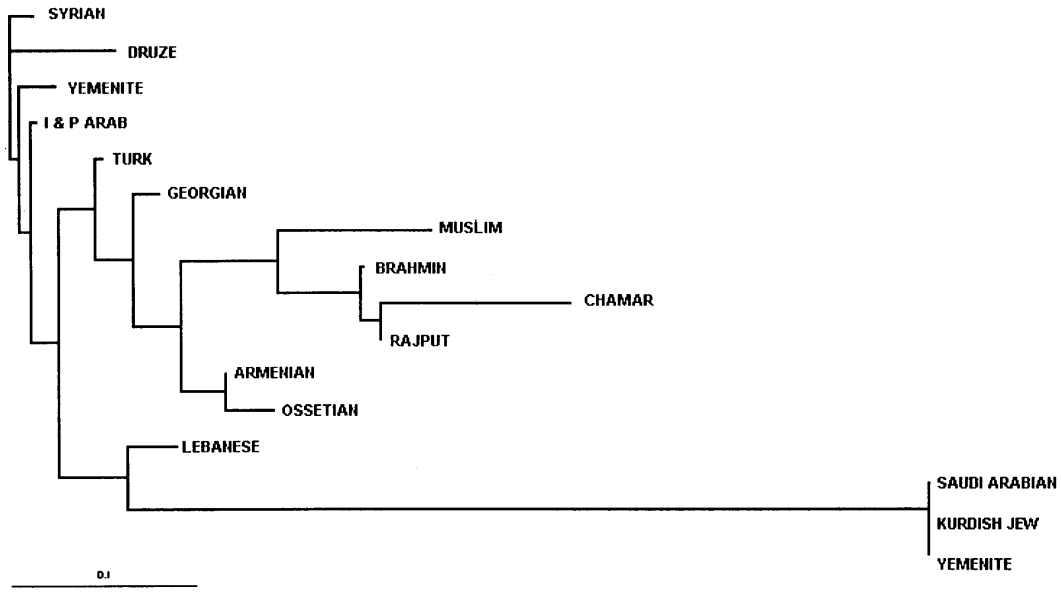

Figure 4. Neighbour-joining tree based on Y-haplogroup frequencies depicting relationships among populations of North India, the Middle East, Central Asia and West Asia.

Murci *et al*. 2001). Our data, however, are somewhat at variance with previous reports (Rosser *et al*. 2000; Quintana-Murci *et al*. 2001). We have found (table 1) that the frequency of HG-3 in Uttar Pradesh is quite high (35%–58%). Although some published data from India are available (Zerjal *et al*. 1999; Rosser *et al*. 2000), the earlier reported frequencies of HG-3 from other parts of India are much lower. Our present data, and the data for Sindhis of Pakistan (Zerjal *et al*. 1999) among whom HG-3 frequency is 52%, indicate that the decreasing clinal pattern of HG-3 from Iran to India may not be as smooth as previously reported (Rosser *et al*. 2000). However, our estimate of the age of this haplogroup (5200 ybp) from the data on North Indian populations does not contradict the higher estimated age (Karafet *et al*. 1999) of 7500 ybp from data on Central Asian populations.

In addition to the data on HG-9 and HG-3 that provide clear indications of population movements from Iran and Central Asia into India, some other haplogroups (HG-1 and HG-2) that are common in Europe (Rosser *et al*. 2000) are also found in North Indian and also in the Middle Eastern and Central Asian populations, indicating genomic closeness of populations of these regions to the Caucasoids. It is historically known that a major influx of people into India was of the Central Asian tribal Huns, who at the end of the fifth century broke through into northern India (Thapar 1975; Kochhar 2000) after several aborted attempts. The Hun dominion extended from Persia right across to Khotan, the main capital being Bamiyan in Afghanistan. Together with the Huns came a number of other Central Asian tribes and peoples, some of whom remained in northern India and others moved further to the south and the west. Some of these tribal people became the ancestors of the Rajput families (Thapar 1975). It is interesting that HG-21, which is thought to have originated in Africa about 31,000 ybp (Hammer *et al*. 1998), is present in the Middle East in moderately high frequencies, but is completely absent in northern India (table 1). This haplogroup is on the YAP(+), presence of the Y *Alu* polymorphic element, background. We have previously noted (Bhattacharyya *et al*. 1999) that the YAP element is absent in India. HG-21 is very common (Rosser *et al*. 2000) in many northern African populations and is largely confined to the southern region of Europe (Greece and Cyprus). Rosser *et al*. (2000) have contended that this haplogroup may reflect a barrier to gene flow between Africa and Europe.

The observed haplogroup diversities in northern India are highly variable. The Brahmin and Rajput, upper-ranked and middle-ranked caste groups, show much higher haplogroup diversities than the lower-ranked Chamar, and the Muslims. This observation is consistent with the historical view that there have been a greater inflow of genes from various Caucasoid groups into the Brahmin and Rajput. In the Middle East, Central Asia and West Asia, haplogroup diversities among the populations considered are generally high (63% for Syrian to 81% for Turk).

The sets of STR haplotypes present in the population groups of northern India are virtually disjoint. This feature has been reported by us earlier (Bhattacharyya *et al*. 1999) and is possibly a reflection of the low male gene flow across ethnic barriers resulting from social norms governing the institution of marriage. We have found that the dispersion of repeat frequencies among STR haplotypes is higher among Brahmins and Rajputs compared to Chamars and Muslims. This is consistent with the notion that there has been higher inflow of Caucasoid genes into the upper-ranked and middle-ranked caste groups than into other groups. It is noteworthy that all individuals possessing the two modal haplotypes 15-12-22-10-11-12 and 15-12-25-11-11-13, which are separated by a total of four repeats at three STR loci, belong to HG-2 and HG-3, respectively. Since the two modal haplotypes belong to different haplogroup backgrounds, they are not genealogically related and probably represent different source populations. Remarkably, the modal haplotypes are found across the three Hindu populations of this study but not in the Muslim. It is also remarkable that Indians and Arabs belonging to the same haplogroup have virtually no commonality of STR haplotypes. In fact the two most frequent haplotypes found in northern India (15-12-22-10-11-12 and 15-12-25-11-11-13) are not found among the Israeli and Palestinian Arabs. Similarly, the modal haplotypes found in the Middle Eastern populations are not present in northern India. This observation is not clearly interpretable, because it is known that there has been male immigration (warriors) into India from West Asia/Central Asia/Middle East in historical times (Spear 1975). The most probable reason for this is that because the STRs evolve rapidly because of high mutation rates, signatures of ancient population movements are often erased in a short time span. Within HG-3, the most frequent haplotypes (at loci DYS19, DYS390, DYS391, DYS392, DYS393) in Central Asia are 16-25-10-11-13 and 16-25-11-11-13. In our pooled sample of North Indians, the most common haplotype within this haplogroup is 15-25-11-11-13, which is only one or two mutational steps away from the common Central Asian haplotypes.

The neighbour-joining tree based on Y-haplogroup frequencies (figure 4) shows that the North Indian populations cluster together. Although the Middle Eastern and West Asian populations do not belong to a single cluster, the cluster of North Indian populations is embedded within the clusters of populations that comprise the Central Asian and the Middle Eastern and West Asian populations. This is consistent with gene flow from the Middle East and West Asia, and also from Central Asia, into India.

Acknowledgements

We are grateful to Mrs Monami Roy, Mr Badal Dey and Mr Madan Chakraborty for help in collection of blood samples. This study has been supported in part by Indo-Israeli Science Collaboration grants to P.P.M. from the Department of Biotechnology, Government of India, and to A.O. from the Israeli Ministry of Science, Culture and Sport. We wish to thank Dr Mira Korner at the National Genome Center, Hebrew University, for advice and assistance in part of the DNA analyses.

References

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

Cann R. L. 2001 Genetic clues to dispersal of human populations: retracing the past from the present. *Science* **291**, 1742–1748.

Casalotti E., Simoni L., Belledi M. and Barbujani G. 1999 Y-chromosome polymorphisms and the origins of the European gene pool. *Proc. R. Soc. London* **B266**, 1959–1965.

Casanova M., Leroy P., Boucekkine C., Weissenbach J., Bishop C., Fellous M., Purrello M., Fiori G. and Siniscalco M. 1985 A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. *Science* **230**, 1403–1406.

Cavalli-Sforza L. L., Piazza A. and Menozzi P. 1994 *The history and geography of human genes.* Princeton University Press. Princeton.

Hammer M. F., Karafet T., Rasanayagam A., Wood E. T., Altheide T. K., Jenkins T., Griffiths R. C., Templeton A. R. and Zegura S. L. 1998 Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol. Biol. Evol.* **15**, 427–441.

Hammer M. F., Redd A. J., Wood E. T., Bonner M. R., Jarjanazi H., Karafet T. *et al.* 2000 Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. *Proc. Natl. Acad. Sci. USA* **97**, 6769–6774.

Hill E. Q., Jobling M. A. and Bradley D. G. 2000 Y-chromosome variation and Irish origins. *Nature* **404**, 351–352.

Hurles M. E., Veitia R., Arroyo E., Armenteros M., Bertranpetit J., Perez-Lezaun A. *et al.* 1999 Recent male-mediated gene flow over a linguistic barrier in Iberia, suggested by analysis of a Y-chromosomal DNA polymorphism. *Am. J. Hum. Genet.* **65**, 1437–1448.

Karafet T. M., Zegura S. L., Posukh O., Osipova L., Bergan A., Long J. *et al.* 1999 Asian sources of New World Y-chromosome founder haplotypes. *Am. J. Hum. Genet.* **64**, 817–831.

Karve I. 1961 *Hindu society: an interpretation*, Deshmukh Prakashan, Pune.

Kochhar R. 2000 *The Vedic people*. Orient Longman, Hyderabad.

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.

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

Quintana-Murci L., Krausz C., Zerjal T., Sayar S. H., Hammer M. F., Mehdi S. Q. *et al.* 2001Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. *Am. J. Hum. Genet.* **68**, 537–542.

Raymond M. and Rousset F. 1995 Genepop (version 1.2): Population genetics software for exact tests and ecumenicism. *J. Hered.* **86**, 248–249.

Renfrew C. 1987 Language families and the spread of farming. In *The origins and the spread of agriculture and pastoralism in Eurasia* (ed. D. R. Harris), pp. 70–92. Smithsonian Institution Press, Washington.

Renfrew C. 1990 *Archaeology and language: the puzzle of Indo-European origins*. Jonathan Cape, London.

Rosser Z. H., Zerjal T., Hurles M. E., Adojaan M., Alavantic D., Amorim A. *et al.* 2000 Y-chromosomal diversity in Europe is clinal and is influenced primarily by geography, rather than by language. *Am. J. Hum. Genet.* **67**, 1526–1543.

Rousset F. 2001 Inferences from spatial population genetics. In *Handbook of statistical genetics* (ed. D. Balding, M. Bishop and C. Cannings) pp. 239–269. Wiley, London.

Spear P. 1975 *A history of India, Volume 2*. Penguin, London.

Stumpf M. P. H. and Goldstein D. B. 2001 Genealogical and evolutionary inference with the human Y chromosome. *Science* **291**, 1738–1742.

Thapar R. 1975 *A history of India, Volume 1*. Penguin, London.

Thomas M., Bradman N. and Flinn H. 1999 High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. *Hum. Genet.* **105**, 577–581.

Zerjal T., Dashnyam B., Pandya A., Kayser M., Roewer L., Santos F. R. *et al.* 1997 Genetic relationships of Asians and northern Europeans, revealed by Y-chromosomal DNA analysis. *Am. J. Hum. Genet.* **60**, 1174–1183.

Zerjal T., Pandya A., Santos F. R., Adhikari R., Tarazona E., Kayser M. *et al.* 1999 The use of Y-chromosomal DNA variation to investigate population history: recent male spread in Asia and Europe. In *Genomic diversity: applications in human population genetics* (ed. S. S. Papiha, R. Deka and R. Chakraborty), pp. 91–102, Plenum, New York.

Received 12 December 2001