PRINT: ISSN 0972-3757 ONLINE: 2456-6360
Int J Hum Genet, 8(1-2): 97-118 (2008)
DOI: 10.31901/24566330.2008/08.01-2.09

Genetic Imprints of Pleistocene Origin of Indian Populations: A Comprehensive Phylogeographic Sketch of Indian Y-Chromosomes

R. Trivedi[1], Sanghamitra Sahoo[1,2], Anamika Singh[1], G. Hima Bindu[1], Jheelam Banerjee[1], Manuj Tandon[1], Sonali Gaikwad[1], Revathi Rajkumar[1], T. Sitalaximi[1], Richa Ashma[1], G.B.N. Chainy[3] and V. K. Kashyap[*1,2]

1. National DNA Analysis Centre, Central Forensic Science Laboratory, 30, Gorachand Road, Kolkata 700 014, West Bengal, India
2. National Institute of Biologicals, A-32, Sector 62, Institutional Area, Noida 201307, Uttar Pradesh, India
3. PG Department of Biotechnology,Vani Vihar, Utkal University, Bhubaneswar 751 004, Orissa, India

KEYWORDS Population genetics, people of india, linguistic groups, migration

ABSTRACT Paleoanthropological evidence indicates that modern humans reached South Asia in one of the first dispersals out of Africa, which were later followed by migrations from different parts of the world. The variation of 20 microsatellite and 38 binary polymorphisms on the non-recombining part of the uniparental, hapliod Y-chromosome was examined in 1434 male individual of 87 different populations of India to investigate various hypothesis of migration and peopling of South Asia Sub-continent. This study revealed a total of 24 paternal lineages, of which haplogroups H, R1a1, O2a and R2 portrayed for approximately 70% of the Indian Y-Chromosomes. The high NRY diversity value (0.893) and coalescence age of approx. 45-50 KYA for H and C haplogroups signified an early settlement of the subcontinent by modern humans. Haplogroup frequency and AMOVA results provide similar evidence in support of a common Pleistocene origin of Indian populations, with partial influence of Indo-European gene pool on the Indian society. The differential Y-chromosome and mt DNA pattern in the two Austric speakers of India signaled that an earlier male–mediated exodus from South East Asia largely involved the Austro-Asiatic tribes, while the Tibeto-Burman males migrated with females through two different routes; one from Burma most likely brought the Naga-Kuki-Chin language and O3e Y-chromosomes and the other from Himalayas, which carried the YAP lineages into northern regions of subcontinent. Based on distribution of Y-chromosome haplogroups (H, C, O2a, and R2) and deep coalescing time depths for these paternal lineages, we propose that the present day Dravidian speaking populations of South India are the descendants of earliest Pleistocene settlers while Austro-Asiatic speakers came from SE Asia in a later migration event.

INTRODUCTION

The origins of modern humans in South Asia have been obscure. Archeological and paleo-anthropological evidences are few and fragmentary. Human remains dating back to the Late Pleistocene provide limited but conclusive evidence for early human occupation in the Indian subcontinent (Deraniyagala 1992; Kennedy 2000; James et al. 2005). A number of artefacts of Middle and Upper Paleolithic cultures in Narmada Valley and the remains of Acheulian culture have been extensively found through out South Asia. Mesolithic microliths and evidences of Neolithic settlements found in diverse parts of the subcontinent also testify towards occupation of India by early humans (Misra 2001). Most of the prevailing genetic records further corroborate with the hypothesis that *Homo sapiens* colonized South Asia as a part of an early southern dispersal from Africa (Quintana-Murci et al. 1999; Cann 2001; Macaulay et al. 2005). This paper examines the current genetic diversity of Indian Y chromosomes in context to place the genetic origin/(s) and time of settlement of the earliest human populations in India.

The present-day populations of India belong to 4635 endogamous communities (Singh 1998) and speak as many as 350 living languages (ethnologue), which fall under the four major supra-language families, i.e., Indo-European, Dravidian, Sino-Tibetan and Austric. The nature of extensive diversity among varied groups reported with 54 classical markers showed a typically north-south geographic division of

[*]*Corresponding Author:* Dr. V K Kashyap, Director National Institute of Biologicals, A-32, Sector 62, Institutional Area, NOIDA 201307, India.
Telephone: +91-120-2400027, *Fax:* +91-120-2403014
E-mail: vkk2k@hotmail.com

populations and placed Indians closer to European populations than either with east-Asians or Africans in the genetic distance trees (Cavalli- Sforza et al. 1994). A number of studies based on mt DNA, Y-chromosome and other nuclear DNA markers have invariably supported these observations. Numerous surveys of genetic variation have generally portrayed the differences between caste and tribes, and the extent of gene flow among ranked caste clusters. Most of the studies conclude that maternal gene pool of Indian populations are proto-Asian in origin with limited west-Eurasian admixture. While the Y-chromo-somes of the caste populations were found to be more similar to Europeans than Asians; with greater west-Eurasian admixture in castes of higher rank (Bamshad et al. 2001), recent studies provide congruent evidence against any major influx of Indo-European speakers into the Indian gene pool and have ascertained a late Pleistocene South Asian origin for majority of Indian populations (Sahoo et al. 2006; Sengupta 2006). These new findings are consistent with archeobotanical evidences (Fuller 2003) and linguistic data (Renfrew 1989) which suggest a recent common root for Elamite and Dravidic languages. It is hypothesized that the same prehistoric gene pool of southern Asian Pleistocene coastal settlers from Africa provided inocula for both Indian castes and tribes, and subsequent diversification of the gene pools was probably due to the genetic imprints laid down by later migrants, such as Huns, Greeks, Kushans, Moghuls, and others (Kivisild et al. 2003). However, much speculation remains about which of the population groups are amongst the earliest settlers of the Indian subcontinent. While the Austro-Asiatic tribes have been presumed to be descendants of the early modern humans based on nucleotide diversity of mitochondrial M haplogroup (Roychoudhary et al. 2001; Basu et al. 2003), analysis of the Indian Y chromosomes undertaken in this study depicts a different scenario.

In this study, we have assessed a total of 1434 unrelated male individuals belonging to 87 different Indian populations, of which 936 Y-chromosomes have been previously analysed for 38 Y-SNP markers (Sahoo et al. 2006). Additional 216 samples are included in the present analysis, while Y-chromosomal haplogroup data for 282 additional samples from seven other Indian populations were collated from the literature (Kivisild et al. 2003; Cordaux et al. 2004). The present study is based on simultaneous analysis of 38 SNP and 20-STR markers on the Y-chromosome to provide the age estimates and describe their phylogeographic distribution. Apart from determining antiquity of various populations groups in South Asia, we also discuss the genetic structure and peopling of the subcontinent in light of present molecular evidences.

SUBJECTS AND METHODS

Populations Analyzed

A total of 1152 unrelated male individuals belonging to 80 different populations were analyzed in the present study. Samples include populations from various linguistic families (Indo-European, Austro-Asiatic, Dravidian and Tibeto-Burman) and sixteen geographical areas of India. Blood samples were collected with informed consent using a protocol approved by the ethical committee of CFSL, Kolkata. DNA was extracted using standard protocols (Sambrook 1989) from peripheral blood lymphocytes. Information concerning their geographic origin, linguistic and socio-ethnic affiliation for each population is given in table 1. Additional data on 282 samples from seven Indian populations (Punjab, Konkanstha Brahmin, Koya, Yerava, Mullukunan, Kuruchian, Koraga) and from 76 populations of Western Europe (51), Russia (281), Middle East (102), Caucasus (122), Central Asia (584), Siberia (66), North East Asia (334), South East Asia (552), Oceania (225), Pakistan (691) and Sri Lanka (39) were collated from literature and included in the genetic distance analysis.

Markers Analyzed

38 binary polymorphisms included in the present analysis have been previously described (Sahoo et al. 2006). Analysis of 20 Y-STRs (twelve tetranucleotide repeats DYS19, DYS385a/b, DYS 389I/II, DYS390, DYS391, DYS393, DYS460, H4, DYS437 and DYS439 and three trinucleotide repeats, DYS392, DYS388 and DYS426, two dinucleotide YCAIIa/b, two pentanucleotide repeat loci, DYS438 and DYS447 and a hexa-repeat nucleotide DYS448) was carried out in the same DNA samples by using an in-house standardized protocol using primers described elsewhere (Butler et al. 2002). Y-STR haplotypes were

constructed in a sequential order of loci keeping an ascending numerical order for the minimal haplotype to facilitate Y-chromosome comparisons with other world populations.

Statistical Analyses

Several population genetic parameters, including mean haplogroup and haplotype diversity and their standard errors, mean number of pair wise differences (MPD), pairwise F_{ST} values for haplogroups and associated p values were calculated using ARLEQUIN ver.2.0 software package (Schneider et al. 2000). To test for differences in the proportions, χ^2- test for significance was employed. Apportionment of genetic differences among various socio-ethnic, geographic and linguistic groups at different levels of hierarchical subdivisions; between individuals within populations, between populations within groups and between groups of populations were calculated using analysis of molecular variance (AMOVA) (Excoffier et al. 1992). To examine the factor/(s) responsible for genetic differentiation of Y-chromosomes, AMOVA was done both on binary markers as well as with Y-STRs within the lineages. Significance levels of the genetic variance components as well as $Ö_{ST}$ values were estimated by using 10000 iterations.

Median-joining network algorithm of haplogroup associated haplotypes (Bandelt et al., 1999; Forster et al., 2000) was performed using the software NETWORK 4.1.0.8 version (Life Sciences and Engineering Technology Solutions Web site), with epsilon value set to zero. For network calculation, seven Y-STR (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392 and DYS393) loci were used, where weightage to each locus was given according to the estimated variance. Y-STR loci with highest variance was given the lowest weights. To estimate the time to the most recent common ancestor (TMRCA), we calculated the ages to STR variation within the correponding haplogroup observed in the Indian populations using the average square difference (ASD) method. We used the same seven Y-STRs as those used in Network analysis and and a generation time of 25 years and mutation rate of 6.9×10^{-4} as described by Zhivotovsky et al. (2004).

Neighbor-joining tree based on F_{ST} values of 87 Indian populations were used to illustrate the genetic affinity between the studied groups using MEGA 3.0. Genetic relationship between populations of India and other parts of the world was estimated based on pairwise genetic distances (F_{ST} values) calculated from haplogroup frequencies. Multidimensional scaling (MDS) analysis of pairwise F_{ST} values was performed using XL STAT pro 7.5 to decipher the genetic affinities of populations. The Indian populations were pooled into their regional boundaries for comparison of genetic similarity with world populations, to obtain a better resolution in the MDS plot.

RESULTS

Approximately 1152 individuals belonging to 80 extant human populations from 16 geographical regions of India were analysed with 38 Y-SNPs and 20 Y-STR markers to evaluate the possibility that Austro-Asiatic speakers are the earliest settlers of the Indian subcontinent. 24 different paternal lineages were observed, out of which, haplogroups H-M69, R1a1-M17, R2-M124 and O2a-M95 together account for 69% of the paternal diversity in South Asia. Another 20.9% of the genetic variation in Indian males is described by haplogroups L-M11, J2-M172, O3e-M134, K2-M70, F-M89 and C-RPS4Y_{711}, while the presence of other haplogroups- R1b3-M269 and G-M201 could be attributed to recent admixture with Europeans.

Haplogroups and Extent of Y-Chromosome Diversity in Indians

The Y-SNPs used in this study were based on previous reports of polymorphisms in Eurasian and Oceanic populations. Overall haplogroup diversity among Indians was relatively high when compared to European or East Asian populations. Indian populations depicted diversity values from 0.133 to a high of 0.914, with Austro-Asiatic and Tibeto-Burman tribes generally showing reduced diversity (Table 1). Twenty-five Dravidian populations showed a higher mean haplogroup diversity (0.723± 0.083) compared to Indo-European speakers (0.684± 0.079) represented by thirty endogamous groups. South Indian groups; Andhra Brahmins, Kallar, Raju, Chenchu and Lambadi displayed high lineage diversity values, while populations of North India typically demonstrated lower mean haplogroup diversity

Table 1: Description of the Indian populations included in this study

	Population	*Code*	*State*	*Region*	*Language*	*Social Status*	*Hierarchy*	*Haplogroup Diversity*
1	ANDHRABRAHMIN	ANB	ANDHRA PRADESH	South	Dravidian	CASTE	Upper	0.8538 ± 0.0504
2	CHENCHU	CHU	ANDHRA PRADESH	South	Dravidian	TRIBE		0.8474 ± 0.0437
3	KAMMA CHAUDHARY	KMC	ANDHRA PRADESH	South	Dravidian	CASTE	Lower	0.6433 ± 0.1078
4	KAPPU NAIDU	KPN	ANDHRA PRADESH	South	Dravidian	CASTE	Lower	0.5737 ± 0.1213
5	KOMATI	KOM	ANDHRA PRADESH	South	Dravidian	CASTE	Lower	0.5000 ± 0.1222
6	LAMBADI	LMD	ANDHRA PRADESH	South	Indo-European	TRIBE		0.8789 ± 0.0432
7	NAIKPOD GOND	NPG	ANDHRA PRADESH	South	Dravidian	TRIBE		0.7421 ± 0.0584
8	RAJU	RAJ	ANDHRA PRADESH	South	Dravidian	CASTE	Middle	0.8596 ± 0.0393
9	REDDY	RDY	ANDHRA PRADESH	South	Dravidian	CASTE	Middle	0.8022 ± 0.0687
10	YERUKULA	YER	ANDHRA PRADESH	South	Dravidian	TRIBE		0.6316 ± 0.0875
11	ADI PASI	ADI	ARUNACHAL PRADESH	North-East	Tibeto-Burman	TRIBE		0.3556 ± 0.1591
12	BIHARBRAHMIN	BBH	BIHAR	East	Indo-European	CASTE	Upper	0.4510 ± 0.1174
13	BHUMIHAR	BHU	BIHAR	East	Indo-European	CASTE	Middle	0.6368 ± 0.1151
14	RAJPUT	BRJ	BIHAR	East	Indo-European	CASTE	Upper	0.4394 ± 0.1581
15	KAYASTHA	BKY	BIHAR	East	Indo-European	CASTE	Middle	0.7363 ± 0.0748
16	YADAV	YAV	BIHAR	East	Indo-European	CASTE	Lower	0.6786 ± 0.1220
17	KURMI	KUI	BIHAR	East	Indo-European	CASTE	Lower	0.8590 ± 0.0633
18	BANIYA	BAN	BIHAR	East	Indo-European	CASTE	Lower	0.7273 ± 0.0679
19	GUJ PATEL	PAT	GUJRAT	West	Indo-European	CASTE	Middle	0.7778 ± 0.1100
20	HP RAJPUT	HRJ	HP	North	Indo-European	CASTE	Upper	0.9143 ± 0.0425
21	OROAN	ORO	JHARKHAND	East	Dravidian	TRIBE		0.4091 ± 0.1333
22	HO	HO	JHARKHAND	East	Austro-Asiatic	TRIBE		0.0000 ± 0.0000
23	BHUMIJ	BHJ	JHARKHAND	East	Austro-Asiatic	TRIBE		0.4571 ± 0.1406
24	KHARIA	KRA	JHARKHAND	East	Austro-Asiatic	TRIBE		0.5333 ± 0.1801
25	MUNDA	MUN	JHARKHAND	East	Austro-Asiatic	TRIBE		0.1429 ± 0.1188
26	BIRHOR	BIR	JHARKHAND	East	Austro-Asiatic	TRIBE		0.1333 ± 0.1123
27	SANTHAL	SAN	JHARKHAND	East	Austro-Asiatic	TRIBE		0.0000 ± 0.0000
28	IYENGAR	IYN	KARNATAKA	South	Dravidian	CASTE	Upper	0.7485 ± 0.0610
29	LINGAYAT	LYN	KARNATAKA	South	Dravidian	CASTE	Upper	0.8333 ± 0.0691
30	GOWDA	GOW	KARNATAKA	South	Dravidian	CASTE	Lower	0.9524 ± 0.0955
31	BHOVI	BHV	KARNATAKA	South	Dravidian	CASTE	Lower	0.7524 ± 0.0918
32	CHRISTIAN	CHR	KARNATAKA	South	Dravidian	CASTE	Lower	0.8333 ± 0.0597
33	MUSLIM	MUS	KARNATAKA	South	Dravidian	CASTE	Lower	0.8333 ± 0.2224
34	KURUVA	KUR	KARNATAKA	South	Dravidian	TRIBE		0.7436 ± 0.0909
35	DESASTH BRAHMIN	DSB	MAHARASTRA	West	Indo-European	CASTE	Upper	0.8421 ± 0.0657
36	CHITPAVAN BRAHMIN	CHB	MAHARASTRA	West	Indo-European	CASTE	Upper	0.9048 ± 0.0405
37	MARATHA	MAT	MAHARASTRA	West	Indo-European	CASTE	Middle	0.8000 ± 0.0681
38	DHANGAR	DGR	MAHARASTRA	West	Indo-European	CASTE	Lower	0.8167 ± 0.0571
39	PAWARA	PWR	MAHARASTRA	West	Indo-European	TRIBE		0.7417 ± 0.1053
40	KATKARI	KTK	MAHARASTRA	West	Indo-European	TRIBE		0.8246 ± 0.0648
41	MADIA GOND	MGD	MAHARASTRA	West	Dravidian	TRIBE		0.6264 ± 0.1098

Table 1: Contd....

	Population	*Code*	*State*	*Region*	*Language*	*Social Status*	*Hierarchy*	*Haplogroup Diversity*
42	MAHADEO KOLI	MKL	MAHARASTRA	West	Indo-European	TRIBE		0.7636 ± 0.0833
43	MARA	MRA	MIZORAM	North-East	Tibeto-Burman	TRIBE		0.5333 ± 0.0515
44	HMAR	HMR	MIZORAM	North-East	Tibeto-Burman	TRIBE		0.2789 ± 0.1235
45	LAI	LAI	MIZORAM	North-East	Tibeto-Burman	TRIBE		0.5455 ± 0.0615
46	LUSEI	LUS	MIZORAM	North-East	Tibeto-Burman	TRIBE		0.4094 ± 0.1002
47	KUKI	KUK	MIZORAM	North-East	Tibeto-Burman	TRIBE		0.6818 ± 0.0910
48	MANIPURI MUSLIM	MMS	MIZORAM	North-East	Tibeto-Burman	CASTE	Lower	0.8333 ± 0.0980
49	ORIYABRAHMIN	OBH	ORISSA	East	Indo-European	CASTE	Upper	0.8043 ± 0.0697
50	KARAN	KRN	ORISSA	East	Indo-European	CASTE	Middle	0.6471 ± 0.0953
51	KHANDAYAT	KDY	ORISSA	East	Indo-European	CASTE	Middle	0.7564 ± 0.0974
52	GOPE	GPE	ORISSA	East	Indo-European	CASTE	Lower	0.8333 ± 0.0720
53	PAROJA	PRJ	ORISSA	East	Dravidian	TRIBE		0.8667 ± 0.0483
54	JUANG	JUN	ORISSA	East	Austro-Asiatic	TRIBE		0.0000 ± 0.0000
55	SAORA	SAR	ORISSA	East	Austro-Asiatic	TRIBE		0.6784 ± 0.0884
56	NEPALI	NEP	SIKKIM	North-East	Tibeto-Burman/ Indo-European	CASTE	Upper	0.9048 ± 0.1033
57	BHUTIA	BHT	SIKKIM	North-East	Tibeto-Burman	TRIBE		0.5000 ± 0.2652
58	CHAKKLIAR	CHK	TAMIL NADU	South	Dravidian	CASTE	Lower	0.7912 ± 0.0673
59	KALLAR	KAL	TAMIL NADU	South	Dravidian	CASTE	Middle	0.8788 ± 0.0751
60	VANNIYAR	VAN	TAMIL NADU	South	Dravidian	CASTE	Middle	0.8333 ± 0.0597
61	PALLAR	PAL	TAMIL NADU	South	Dravidian	CASTE	Lower	0.7000 ± 0.0896
62	GOUNDER	GOU	TAMIL NADU	South	Dravidian	CASTE	Upper	0.7124 ± 0.0650
63	IRULAR	IRU	TAMIL NADU	South	Dravidian	TRIBE		0.7576 ± 0.1221
64	KANYAKUBJ BRAHMIN	KKB	UTTAR PRADESH	North	Indo-European	CASTE	Upper	0.6909 ± 0.1276
65	UP JAT	UPJ	UTTAR PRADESH	North	Indo-European	CASTE	Upper	0.0000 ± 0.0000
66	UP THAKUR	UPT	UTTAR PRADESH	North	Indo-European	CASTE	Upper	0.5357 ± 0.1232
67	KHATRI	KHT	UTTAR PRADESH	North	Indo-European	CASTE	Middle	0.2857 ± 0.1964
68	BHOKSHA	BKS	UTTAR PRADESH	North	Tibeto-Burman/ Indo-European	TRIBE		0.6222 ± 0.1383
69	UP KURMI	UPK	UTTAR PRADESH	North	Indo-European	CASTE	Lower	0.0000 ± 0.0000
70	THARU	THR	UTTAR PRADESH	North	Tibeto-Burman/ Indo-European	TRIBE		0.8000 ± 0.1721
71	JAUNSARI	JUS	UTTAR PRADESH	North	Tibeto-Burman/ Indo-European	TRIBE		0.7333 ± 0.1552
72	MAHISHIYA	MSY	WEST BENGAL	East	Indo-European	CASTE	Middle	0.8684 ± 0.0489
73	NAMASUDRA	NMS	WEST BENGAL	East	Indo-European	CASTE	Lower	0.9000 ± 0.0355
74	BAURI	BAU	WEST BENGAL	East	Indo-European	CASTE	Lower	0.6526 ± 0.0648
75	MAHELI	MHL	WEST BENGAL	East	Austro-Asiatic	TRIBE		0.8211 ± 0.0586
76	KARMALI	KRM	WEST BENGAL	East	Austro-Asiatic	TRIBE		0.2924 ± 0.1274
77	KORA	KOR	WEST BENGAL	East	Indo-European	TRIBE		0.7579 ± 0.0495
78	LODHA	LOD	WEST BENGAL	East	Austro-Asiatic	TRIBE		0.8421 ± 0.0595
79	EZHAVA HINDU	EZH	KERALA	South	Dravidian	CASTE	Lower	0.8056 ± 0.0889
80	NAIR	NAR	KERALA	South	Dravidian	CASTE	Upper	0.0000 ± 0.0000

Table 2a: Comprehensive haplogroup frequency data among linguistic, geographic and social categories of India

	Sample Size	*C*	*D*	*F**	*G*	*H**	*H1*	*H2*	*J2**	*K**	*K2*	*L*	*L1*	*M*	*N*	*O**
TOTAL INDIA	1152	0.014	0.004	0.030	0.001	0.069	0.159	0.002	0.051	0.038	0.031	0.045	0.010	0.000	0.000	0.003
Language																
INDO-EUROPEAN	518	0.012	0.004	0.027	0.002	0.079	0.183	0.002	0.058	0.029	0.033	0.035	0.002	0.000	0.000	0.000
DRAVIDIAN	393	0.020	0.000	0.048	0.000	0.089	**0.209**	0.003	0.056	0.041	0.043	0.084	0.028	0.000	0.000	0.000
AUSTRO-ASIATIC	140	0.014	0.000	0.014	0.000	0.021	0.043	0.000	0.050	0.043	0.014	0.007	0.000	0.000	0.000	0.007
TIBETO-BURMAN	101	0.000	0.030	0.000	0.000	0.010	0.000	0.000	0.000	0.069	0.000	0.000	0.000	0.000	0.000	0.020
Geography																
NORTH	180	0.000	0.006	0.011	0.006	0.106	0.139	0.000	0.078	0.000	0.000	0.017	0.000	0.000	0.000	0.000
WEST	135	0.037	0.000	0.007	0.000	0.081	**0.356**	0.007	0.081	0.000	0.000	0.096	0.000	0.000	0.000	0.000
EAST	357	0.011	0.000	0.048	0.000	0.056	0.106	0.000	0.036	0.053	0.048	0.020	0.003	0.000	0.000	0.003
NORTH-EAST	108	0.000	0.037	0.000	0.000	0.009	0.000	0.000	0.000	0.083	0.000	0.000	0.000	0.000	0.000	0.019
SOUTH	372	0.019	0.000	0.040	0.000	0.078	**0.194**	0.003	0.056	0.043	0.051	0.078	0.030	0.000	0.000	0.000
Social Hierarchy																
UPPER CASTE	211	0.009	0.005	0.019	0.000	0.043	0.185	0.005	0.100	0.024	0.000	0.095	0.019	0.000	0.000	0.000
MIDDLE CASTE	175	0.006	0.000	0.051	0.000	0.040	0.171	0.000	0.097	0.040	0.017	0.034	0.023	0.000	0.000	0.000
LOWER CASTE	261	0.008	0.000	0.046	0.000	0.107	0.169	0.000	0.031	0.050	0.046	0.054	0.000	0.000	0.000	0.000
TRIBES	505	0.022	0.008	0.020	0.002	0.071	0.139	0.002	0.026	0.038	0.042	0.024	0.008	0.000	0.000	0.006

	Sample Size	*O2a*	*O2a1*	*O3*	*O3e*	*P**	*R**	*R1*	*R1a*	*R1a1*	*R1b3*	*R2*
TOTAL INDIA	1152	0.149	0.001	0.001	0.026	0.027	0.010	0.011	0.002	0.175	0.005	0.135
Language												
INDO-EUROPEAN	518	0.010	0.000	0.002	0.006	0.039	0.021	0.008	0.004	**0.297**	0.012	0.137
DRAVIDIAN	393	0.023	0.000	0.000	0.000	0.008	0.000	0.023	0.000	0.117	0.000	0.209
AUSTRO-ASIATIC	140	**0.729**	0.000	0.000	0.000	0.036	0.000	0.000	0.000	0.007	0.000	0.014
TIBETO-BURMAN	101	**0.554**	0.010	0.000	0.267	0.030	0.000	0.000	0.000	0.010	0.000	0.000
Geography												
NORTH	180	0.000	0.000	0.006	0.017	0.000	0.011	0.000	0.006	**0.483**	0.006	0.111
WEST	135	0.000	0.000	0.000	0.000	0.030	0.022	0.000	0.000	0.193	0.000	0.089
EAST	357	**0.325**	0.000	0.000	0.000	0.045	0.014	0.006	0.003	0.104	0.000	0.120
NORTH-EAST	108	**0.519**	0.009	0.000	0.250	0.046	0.000	0.009	0.000	0.019	0.000	0.000
SOUTH	372	0.000	0.000	0.000	0.000	0.016	0.003	0.027	0.000	0.134	0.013	0.215
Social Hierarchy												
UPPER CASTE	211	0.000	0.000	0.000	0.000	0.019	0.014	0.005	0.005	**0.360**	0.005	0.090
MIDDLE CASTE	175	0.000	0.000	0.000	0.000	0.029	0.011	0.029	0.000	**0.263**	0.000	0.189
LOWER CASTE	261	0.004	0.000	0.000	0.004	0.023	0.004	0.023	0.000	0.157	0.000	**0.276**
TRIBES	505	**0.339**	0.002	0.002	0.057	0.032	0.010	0.002	0.002	0.077	0.010	0.061

Table 2b: Comparative haplogroup frequency data in socio-ethnic groups of India

	Tribe				*Dravidian Caste*			*Indo-European Caste*				
	AA	*DR*	*TB*	*IE*	*Upper*	*Middle*	*Lower*	*Upper*	*Middle*	*Lower*	*Caste_DR*	*Caste _IE*
No. of populations	11	8	7	8	5	4	10	9	8	8	19	25
Sample size.	179	126	92	108	72	58	137	132	117	115		
							Total (in %)					
C	1.1	5.6		1.9			0.7	1.5	0.9	0.9	0.4	1.1
D			3.3	0.9								
F*	2.2	3.2		1.9	2.8	10.3	5.1	1.5	2.6	4.3	5.6	2.7
G				0.9								
H*	2.8	12.7		13.9	5.6	5.2	8.8	3.8	3.4	13.0	7.1	6.6
H1	7.3	22.2		26.9	31.9	10.3	18.2	12.1	20.5	16.5	20.2	16.2
H2				0.9	1.4						0.4	
J2*	3.9	3.2		1.9	8.3	15.5	2.2	11.4	6.8	4.3	6.7	7.7
K*	3.4	7.1	4.3		4.2	1.7	2.2		5.1	6.1	2.6	3.6
K2	2.2	11.1		2.8		5.2				10.4	1.1	3.3
L	0.6	4.8		4.6	15.3	5.2	9.5	6.8	2.6	0.9	10.1	3.6
L1		3.2			4.2	6.9		0.8			2.6	0.3
O*	0.6		2.2									
O2a	57.0	7.1	59.8	4.6								
O2a1			1.1									
O3				0.9								
O3e			28.3	2.8								
P*	5.6	0.8		4.6		1.7	0.7	1.5	3.4	1.7	0.7	2.2
R*	1.7			1.9				2.3	1.7	0.9		1.6
R1				0.9		5.2	4.4		1.7		3.4	0.5
R1a	0.6							0.8				0.3
R1a1	0.6	11.9	1.1	20.4	15.3	10.3	10.2	48.5	34.2	23.5	11.6	36.0
R1b3				4.6				0.8				0.3
R2	10.6	7.1		2.8	11.1	22.4	38.0	8.3	17.1	17.4	27.3	14.0

Table 3: Y-Chromosome microsatellite diversity within the twelve major lineages found in India

	P	*F*	*H*	*O3*	*O2a*	*L*	*K*	*K2*	*R1a1*	*C*	*J2*	*R2*
No of Males	29	24	221	26	162	55	36	36	191	16	52	136
No of Haplotypes	27	24	206	24	144	53	36	36	188	16	51	131
Lineage diversity	0.995± 0.010	1.000 ± 0.012	0.999 ± 0.000	0.993 ± 0.012	0.997 ± 0.001	0.998 ± 0.003	1.000 ± 0.006	1.000 ± 0.006	0.999 ± 0.000	1.000 ± 0.022	0.999 ± 0.004	0.999± 0.001
MPD	14.19 ± 6.55	13.91± 6.47	12.66± 5.73	7.37 ± 3.56	11.55± 5.26	13.13 ± 6.00	12.93 ± 5.96	14.18 ± 6.51	12.05± 5.47	13.45± 6.38	14.04 ± 6.40	13.04± 5.90
Av. gene diversity	0.709 ± 0.364	0.695 ± 0.360	0.633± 0.317	0.368 ± 0.198	0.577± 0.291	0.691 ± 0.350	0.718 ± 0.368	0.709 ± 0.361	0.602 ± 0.302	0.707 ± 0.376	0.702 ± 0.355	0.652± 0.327
Haplotype sharing												
Within	2	Non	10	2	10	2	Non	Non	3	Non	1	2
Between	Non	Non	2	Non	Non	Non	Non	Non	Non	Non	Non	1

(0.569 ± 0.104). The distribution of the lineages and Y-STR diversity within the haplogroups are described in detail.

Haplogroup H-M69

Majority of males analyzed from different geographic regions of India (23%) carried the M69C haplotype, which is additionally defined by M52C mutation. Distribution of Haplogroup H showed a north-south gradient (24.4% to 27.4%), however geographically its total frequency was highest (44.4%) in populations of western India (Table 2a). Among the 23% carrying H lineage in their Y-chromosomes, most of them were representatives of south India (27.4%), speaking Dravidian languages (30.0%). In socio-ethnic groups, the frequency was 27.6% in lower caste groups, while in the tribal groups it accounted for 21.2% of their paternal variation. However, the pattern of distribution did not vary statistically between the Dravidian and Indo-European speaking tribes or caste cluster (Table 2b). 206 distinct 20-Y-STR haplotype profiles deciphered out of 221 individuals carried a mean pairwise difference of 12.66 (Table 3), where none of the hapotypes were shared between groups. In the median–joining network analysis with 7-Y-STRs associated with M69C/ M52C lineage branch, majority of Y-chromosome STR haplotypes are connected by one-or two-step mutation events (Fig. 2a). Kora, an Indo-European speaking tribal group from eastern India, branched out of the network with more than three mutation steps. The Y-STR based coalescence time of haplogroup H1 chromosomes was estimated to be ~ 43,556 years (Table 4).

Haplogroup R1a1-M17

Haplogroup R1a1-M17 characterizes 17.5% of the Indians. The Indo-European speakers demonstrated a significantly higher proportion of this lineage as compared to populations belonging to Dravidian linguistic family (29.7% vs. 11.7%; χ^2= 7.82, p<0.05) (Table 2a). With the exception of Lodha, Nepali and Bhutia, all other Austro-Asiatic and Tibeto-Burman speakers lack this haplogroup in their Y-chromsomes. While the Indo-European and Dravidian caste group depict significant variation (χ^2= 12.5, p<0.01), the tribal groups are more akin. Distribution of M17 lineage also showed a decreasing geographic cline along the latitude; its frequency was highest (approx. 50%) among the populations of Bihar and Uttar Pradesh, where almost 60% of the Upper and Middle caste groups harbored R1a1 Y–chromosomes. Out of the 191 males that carried R1a1 haplogroup, 188 unique 20-YSTR haplotypes were observed (h=0.999). While no haplotype was shared between populations, intra-population variation was observed within Jat, Bhoksha and Yerukula and the mean pairwise difference between all the Y-STRs was found to be high (12.05) (Table 3). The median–joining network analysis, however, revealed that populations of neighbouring area shared few of the haplotypes (Fig. 2b). Passarino et al 2002 reported two region specific allele pattern associated within M17 among Europeans; DYS19=15 and YCA IIa,b=19,21 was specific to the R1a1s in Western Europe, while Eastern European R1a1s typically harbored allele16 for DYS19 and 19,23 for YCA IIa,b. In our dataset, although, allele 15 and 16 at DYS19 were the two most common alleles with significant difference in their frequency (χ^2= 4.66, p<0.05), they did not reveal any specific geographical, socio-ethnic or linguistic pattern in their distribution. The TMRCA of those individuals harboring R1a1 Y-chromosome is estimated around ~32 KYA (Table 4).

Table 4: Y-Chromosome haplogroup variances and TMRCA estimated on seven Y-STR loci

Locus -wise variance	*R1a1*	*H*	*C*	*O2a*	*R2*
DYS19	0.549	0.497	1.183	0.293	0.770
DYS3891	0.793	0.760	0.783	0.383	0.649
DYS3892	0.960	0.872	1.400	0.620	1.211
DYS390	1.173	2.246	1.983	2.314	1.375
DYS 391	1.081	2.533	1.762	0.791	0.966
DYS392	1.009	0.708	1.662	1.620	1.749
DYS393	0.710	0.921	0.917	0.995	1.051
Average	0.896	1.220	1.384	1.002	1.110
Age estimates in years	32,015.31	43,556.12	49,438.78	35,795.92	39,647.96

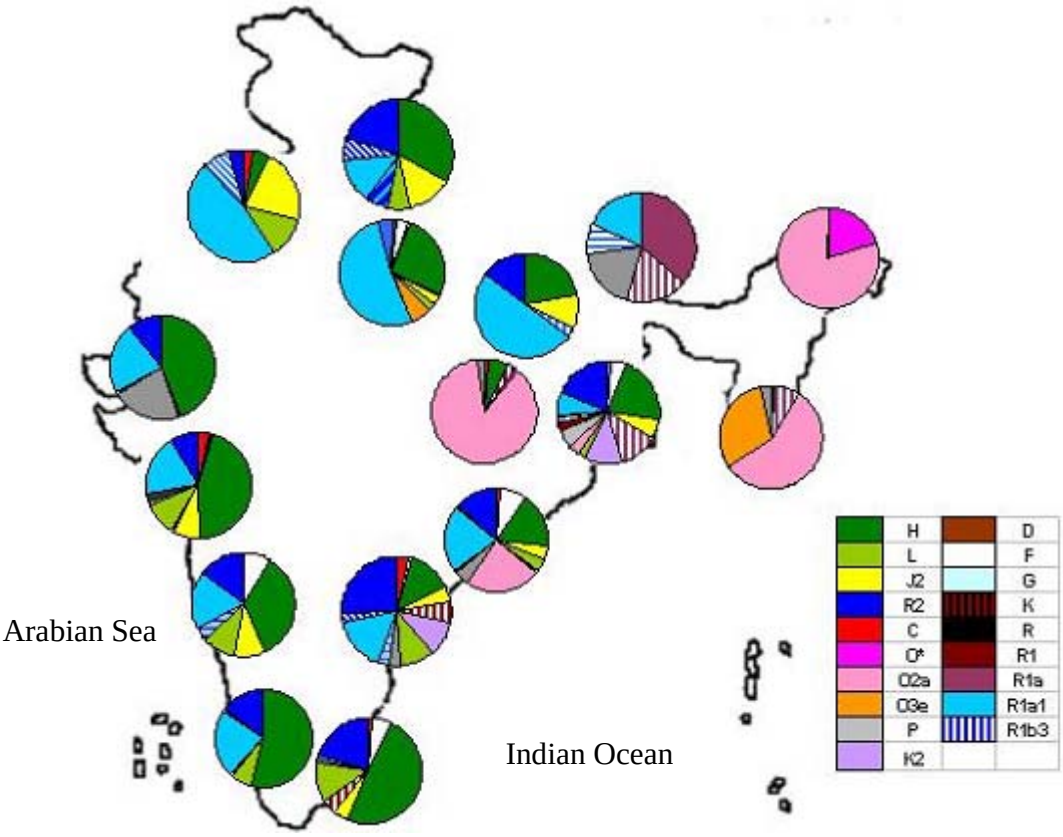

Fig. 1. Y-Chromosome haplogroups and their frequency distribution in different regional populations of India

Haplogroup O2a-M95

Haplogroup M95, which forms the major South-East Asian male lineage, (Su et al. 2000, Karafet et al 2001) accounts for 15% of the Y-chromosome variation in India. It is however, localized to the eastern part of the subcontinent, restricted among the Austro-Asiatic speakers (72.9%) and Tibeto-Burman speaking tribes (56.4%) of NE India (Table 2a). Although this haplogroup was also detected in Indo-European and Dravidian speakers (3.3% in total), its presence in them could be sufficiently attributed to admixture from Austro-Asiatic speaking neighbors living in close vicinity. While this lineage is completely fixed in Juang, Ho and Santhal, it is observed that the frequency in Tibeto-Burman tribes varied from 25% in Kuki to 80% in Hmar. Surprisingly however, none of the Himalayish branch of Tibeto-Burman speakers; Nepali, Bhutia, Tharu, Jaunsari and Bhoksha harbor this haplogroup (Fig. 1). Although the Y-chromosomes were rather similar (F_{ST} =0.03, $p<0.05$), none of the Y-STRs were shared between groups (Table 3). A comparison of Indian M-95 Y-STR haplotypes with populations of SE Asia including Java, Borneo, Taiwan and Malay revealed that the Austro-Asiatic speakers of Indian subcontinent showed closer affinity to the SE Asians than their Tibeto-Burman speaking neighbors (F_{CT} =0.43 vs 4.15, respectively) (data not shown). To further investigate the relationships between O2a Y-chromosome in the Austro-Asiatic and Tibeto-Burman speakers, a median-joining network of 27 discrete haplotypes of 151 individuals was constructed (Fig. 2c). This network exhibited two distinct clusters of haplotypes with considerable haplotype sharing between the two linguistic families; however the Austro-Asiatic speakers depicted more diverse

haplotypes compared to the Tibeto-Burmans. The TMRCA of all M95T chromosomes was estimated to be ~35,795 years (Table 4).

Haplogroup R2-M124

Our analysis revealed that haplogroup R2 characterizes 13.5% of the Indian Y-chromosomes and its frequency among Dravidian speakers was comparable to that of haplogroup H (20.9%) and significantly different from Indo-European and Austro-Asiatic speakers (χ^2= 16.2, d=3, p<0.05). While the distribution across various geographic regions was almost uniform, significant differentiation was observed along the social groups (χ^2= 18.7, d=3, p<0.05); a decreasing gradient was discernible as one moved up the caste hierarchy (Table 2a). Although tribes contributed only 7.4% of the total R2 lineage, it was proportionately distributed between the Austro-Asiatic and Dravidian tribes (Table 2b). Extensive analysis of its distribution between north and south Indian populations showed that while there was marginal difference among middle and lower caste groups of north India (17.1 and 17.4 % respectively), a clear gradient was observed among south Indians, where the frequency declined by more than one-half from lower to upper caste groups. Analysis of 20-Y-STRs within the R2 lineage revealed that three haplotypes were shared; one between Kamma Chaudhary and Kappu Naidu, both lower caste Dravidian speakers from Andhra Pradesh and two within Karmali and Pallar populations. Network analysis (Fig. 2d) depicted that a large number haplotypes were shared between populations of south India, while the populations of eastern India harbored more discrete Y-STR haplotypes. The TMRCA for M124T was estimated to be ~39,647 years (Table 4).

Haplogroup L-M11

The overall frequency of haplogroup L-M11 in the Indian populations was estimated to be 5.6%, while sporadic occurrence of this lineage has earlier been described among Indo-European speakers of Caucasus, Middle East, Europe and a maximum of 4.3% in Central Asia (Semino et al. 2000; Wells et al. 2001). Dravidian speaking populations harbored a significantly higher percentage of L haplogroup compared to the Indo-European speakers, 11.2 and 3.7% respectively (χ^2= 3.77, d=1, p=0.05). While frequencies were rather comparable in the lower caste groups of north and south, middle and upper caste populations of south India demonstrated relatively higher frequencies than northern caste groups (Table 2a). The L-network and high MPD (13.13) revealed results in congruence with AMOVA suggesting that no clear geographic, linguistic or social pattern could be discerned among the Y-STR haplotypes (data not shown).

Haplogroup J2-M172

Haplogroup J2-M172 is the major lineage of Middle East/Mediterranean and its frequency decreases into Europe. Among the studied Indian populations, M172G exhibited a total frequency of 5.1%, where it was uniformly distributed among the three major linguistic families. Except for the Tibeto-Burman speaking tribes of northeast India, where this lineage was totally absent, no specific cline could be deciphered among the other seventy-three mainland populations. In the social categories, upper and middle caste populations harbor a significantly higher percentage of J2 lineage (~ 10%) as compared to ~3% in lower caste populations and tribal groups (χ^2= 7.74, d=3, p=0.05). While the distribution was proportionate among upper caste groups of north and south India, the difference was more discrete (6.8% vs 15.5%) among middle caste groups of these regions (Table 2b). The Near Eastern populations harboring M172 Y-chromosomes are characterized by a very high frequency of DYS388 alleles with e”15 repeats, while more than 70% of the males examined under this study displayed alleles with repeat motifs d”14. Network analysis showed a large number of divergent haplotypes even with 7-YSTRs and only two reticulations. Lodha, the only Austro-Asiatic tribe that harbored J2 lineage also displayed very diverse haplotype profiles. The genetic structure estimated with AMOVA showed that although the extent of Y-STR differentiation in the populations carrying this lineage was approximately 3%, geography, language or position in the social hierarchy could not statistically delineate the studied Indian populations. Because this marker is associated with the spread of agriculture, we further estimated variance in the agricultural groups and observed a marginal difference of 1% in Y-chromosomes of landowner and labourer communities harboring the J2 lineage (data not shown).

Haplogroup O3-M122 and O3e-M134

The M122 haplogroup and its sub-lineage M134 are among the predominant and widespread lineages of East Asia (Su et al. 2000; Karafet et al. 2001; Shi et al. 2005). In the Indian males, it was detectable at frequencies less than 3% and was largely restricted among the Tibeto-Burman speakers of North-East. It was sporadically present among tribal groups of north India, particularly Tharu (Fig. 1), probably due to recent admixture with neighboring Tibeto-Burman speakers of Nepal and China. A clear delineation along the language family was observed in its distribution; where it was completely absent among the Tani speakers (Adi Pasi), while the Naga-Kuki-Chin branch of Tibeto-Burman speakers contributed the entire 26.7% of O3 lineage. The mean pairwise difference between Y-STRs and the lineage diversity was low at 7.37 and 0.993, respectively, compared to its sister clade O2a.

Haplogroup K2-M70

Haplogroup K2 occurs on a M9G background and is reported to occur in populations of Near East and Europe (Underhill et al. 2000). In our study, it was found only in the eastern and southern regions of the country, adding to an overall frequency of 3.1%. Although it was present in the three major linguistic families, the statistical difference in its distribution was insignificant. However, its distribution depicted an inverse relation as one moved up the social ladder, with the upper caste populations completely lacking the M70 lineage in their Y-chromosomes (Table 2a). This lineage was predominant amongst the lower caste groups of east (Bauri) and tribal groups of south, particularly Yerukula, contributing approx. 60% of the total K2 chromosomes. Within the lineage, 36 distinct Y-STR haplotypes, with a very high mean pairwise difference (14.18) between them, depicted the absence of population structure due to language, geography or ethnicity.

Haplogroup C-M130 (RPS4Y$_{711}$)

The RPS4Y$_{711}$T forms the second major cluster in Asia and Australo-Melanesia and has reportedly spread into North America (Wells et al. 2001; Underhill et al. 2000; Karafet et al. 1999; Underhill et al. 2001). In the present study, this lineage was found spread all along the coastal belt in populations of Maharastra, Tamilnadu, Andhra Pradesh, Orissa and West Bengal, at an average frequency of 1.4%, and noticeably absent in the populations of North and North-East. Although this lineage was present in high frequency in tribes compared to caste groups, its distribution in them was not statistically significant (χ^2= 2.25, d=1, p>0.05). We observed 16 discrete Y-STR haplotypes, with mean pairwise difference between haplotypes of 13.45 (Table 3). Although the total number of individuals carrying RPS4Y$_{711}$T was too small (n=16) to make accurate evolutionary inferences about its origin within South Asia, the TMRCA of Indian RPS4Y$_{711}$T individuals was estimated to be ~ 49,438 years (Table 4).

Other Haplogroups Observed in Indians

Haplogroup F, which is major and the most paraphyletic subcluster of M168 lineages was ubiquitous in its distribution along geographic, linguistic and socio-ethnic boundaries of India and was observed in approximately 3% of the studied males (Table 2a). Haplogroup D, a monophyletic branch of M168 lineage, defined by an Alu insertion and M174C mutation, on the other hand, was restricted in Bhutia and Tharu tribal groups. Its presence among them is most likely due to gene flow from Tibet, where this haplogroup has earlier been reported. Major haplogroups K*, P*, R*, R1, R1a contribute approximately 2-3% of the total Indian Y-chromosomes and there was no difference in its distribution pattern among castes or tribes or among different geographic regions. Although, we detected a few European –specific haplogroups G and R1b3 in Indians, none of our studied samples showed the presence of haplogroup K3-M147, N-M231 or I-M170, which are the other highly predominant haplogroups of Europe.

Genetic structure of the Indian populations

Analysis of molecular variance revealed that the extent of genetic differentiation was high among Indians; percent variation among different groups added up to 27.11%, suggesting that gene pool of India males was highly structured. To identify factor/(s) responsible for this compartmentalization of Y-chromosomes, population

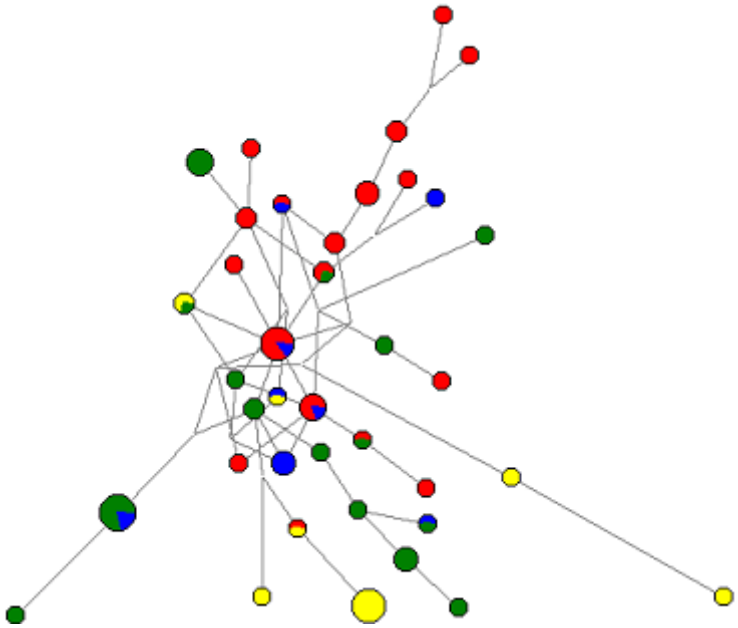

Fig. 2a. Median-Joining network of H haplogroup individuals, based on seven Y-STR haplotypes. Circles represent haplotypes and have an area proportional to frequency. Colour represents the four geographic regions of India (Red: South; Blue: North; Green: West; Yellow: East)

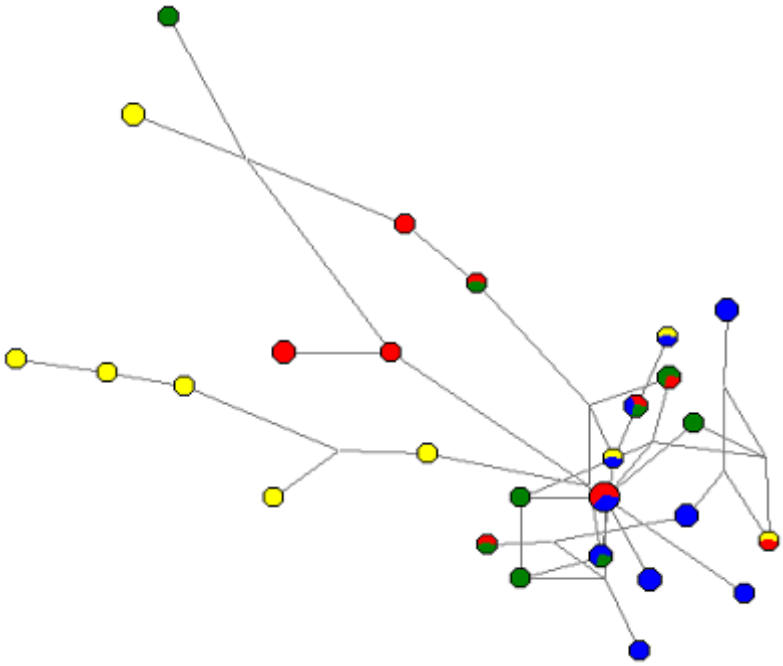

Fig. 2b. Median-Joining network of R1a1 haplogroup individuals, based on seven Y-STR haplotypes. Circles represent haplotypes and have an area proportional to frequency. Colour represents the four geographic regions of India (Red: South; Blue: North; Green: West; Yellow: East)

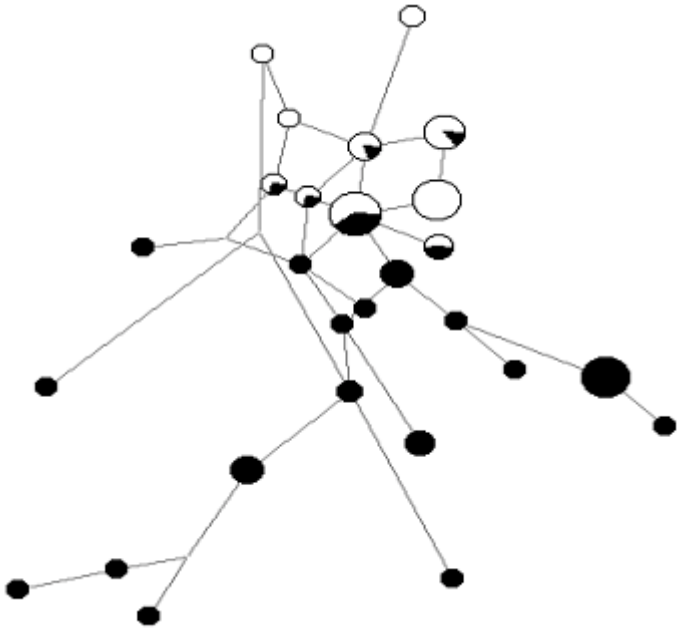

Fig. 2c. Median-Joining network of O2a haplogroup individuals, based on seven Y-STR haplotypes. Circles represent haplotypes and have an area proportional to frequency. Colour represents the Austro-Asiatic (Black) and Tibeto-Burman (White) linguistic families of India

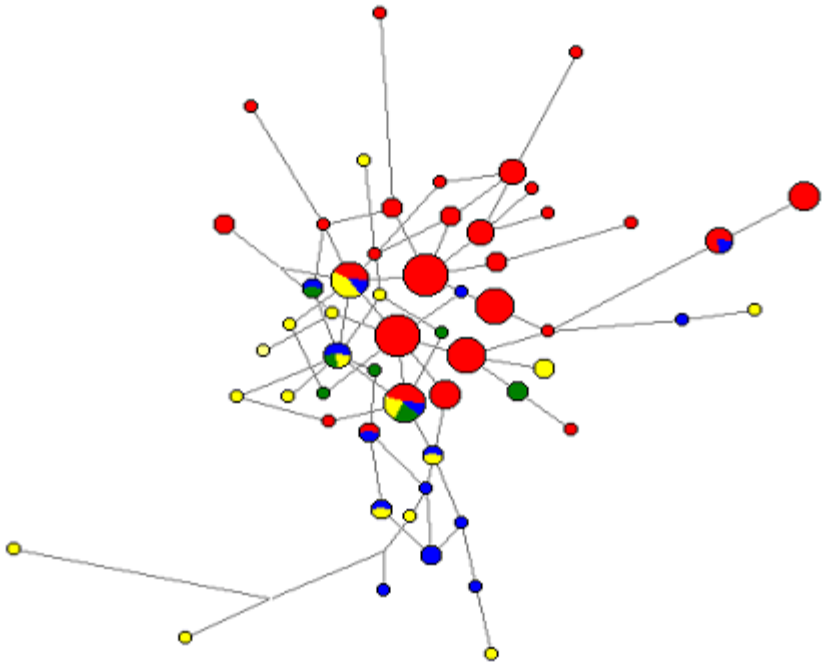

Fig. 2d. Median-Joining network of R2 haplogroup individuals, based on seven Y-STR haplotypes. Circles represent haplotypes and have an area proportional to frequency. Colour represents the four geographic regions of India (Red: South; Blue: North; Green: West; Yellow: East)

genetic structure was analyzed on haplogroup frequency data in a hierarchical mode: within populations, among populations and among groups of populations, pooled according to geography, linguistic family and their socio-ethnic position (Table 5). The amount of genetic variation among five major geographical regions was lesser than the percent due to variation among populations within regions (Öct= 0.096 and Ösc= 0.211, respectively). Further analysis revealed that almost 14.57% of among group variation was due to regional boundaries which also defined their linguistic affinities (language sub-families). The increase in "among group variation" and "among populations within groups" was non-significant when only four major linguistic families were used as a criterion for grouping Indian populations. High Fct and Fsc values suggest that although significant structuring occurs within the populations of India, they could not be partitioned either geographically or linguistically. Apportionment of populations into two broad socio-ethnic group; caste and tribes depicted only 8% difference between them, which further decreased to 6.63% when the caste populations were further resolved into upper, middle and lower groups. Among themselves, the caste populations were not very different, harboring only 1.6% variation among them.

Genetic Relationships among the Indians and with World Populations

To investigate the extent of genetic relationships among India populations, pairwise F_{ST} distances were estimated on the Y-haplogroup frequencies. On the whole, populations clustered according to their Y-chromosome lineages. Two distinct clusters of Indo-European and Dravidian speakers were discernible in the NJ tree on 87 Indian populations, where except for a few deviations most of the populations clustered within their linguistic family. Austro-Asiatic and Tibeto-Burman speakers harboring O2a Y-chromosome lineage described a separate cluster, while Tharu grouped with Mara, Lai and Kuki carrying O3e lineage, and formed a branch distant from other Tibeto-Burman tribes (Fig. 3). MDS plot also substantiated the genetic proximity of Austro-Asiatic and Tibeto-Burman speakers to the populations of South East Asia. Most of the other Indian populations were closer to the Indo-European speakers of Central Asia and Eastern Europe (Russia and Siberia) but distant from populations of Western Europe, while populations of Middle East and Caucasus region formed a separate cluster in the MDS plot. Populations of Uttar Pradesh, Bihar and Punjab were moderately distant from other Indo-European speakers, while those of Pakistan remained between Indians, Central Asia and Russia (Fig. 4).

DISCUSSION

This comprehensive study of Y-chromosome diversity within India aims to identify evolutionary events (founder effects, gene flow and

Table 5: Genetic Differentiation in Indians at different levels of hierarchy based on Y-SNP Data

			Within Population		*Among Population Within Groups*		*Among Groups*	
		No. of Groups	*%*	*FST**	*%*	*FSC**	*%*	*FCT**
Total			72.89	0.271			27.11	
Geography		5	71.22	0.287	19.09	0.211	9.69	0.096
Regional		14	71.95	0.28	13.48	0.157	**14.57**	0.145
Language		4	69.16	0.308	15.53	0.183	**15.31**	**0.153**
	a	4	69.88	0.301	17.17	0.197	12.96	0.129
Social	CS vs TR[$]	2	69.79	0.302	21.73	0.237	8.48	0.084
	b	4	71.49	0.285	21.89	0.234	6.63	0.066
	c	4	82.29	0.177	16.21	0.164	1.51	0.015
Castes	UP vs MD vs LW[#]	3	83.73	0.162	14.65	0.148	1.61	0.016

*All values are statistically significant at $p<0.05$

[$] CS: Castes: TR: Tribes

[#] UP: Upper castes; MD: Middle castes; LW: Lower Castes

a: Includes Karmali and Maheli under Austro-Asiatic; Tharu under Tibeto-Burman language family

b: Includes Upper, Middle, Lower Castes and Tribes

c: Includes Upper, Middle, Lower Castes and excludes Austro-Asiatic and Tibeto-Burman Tribes

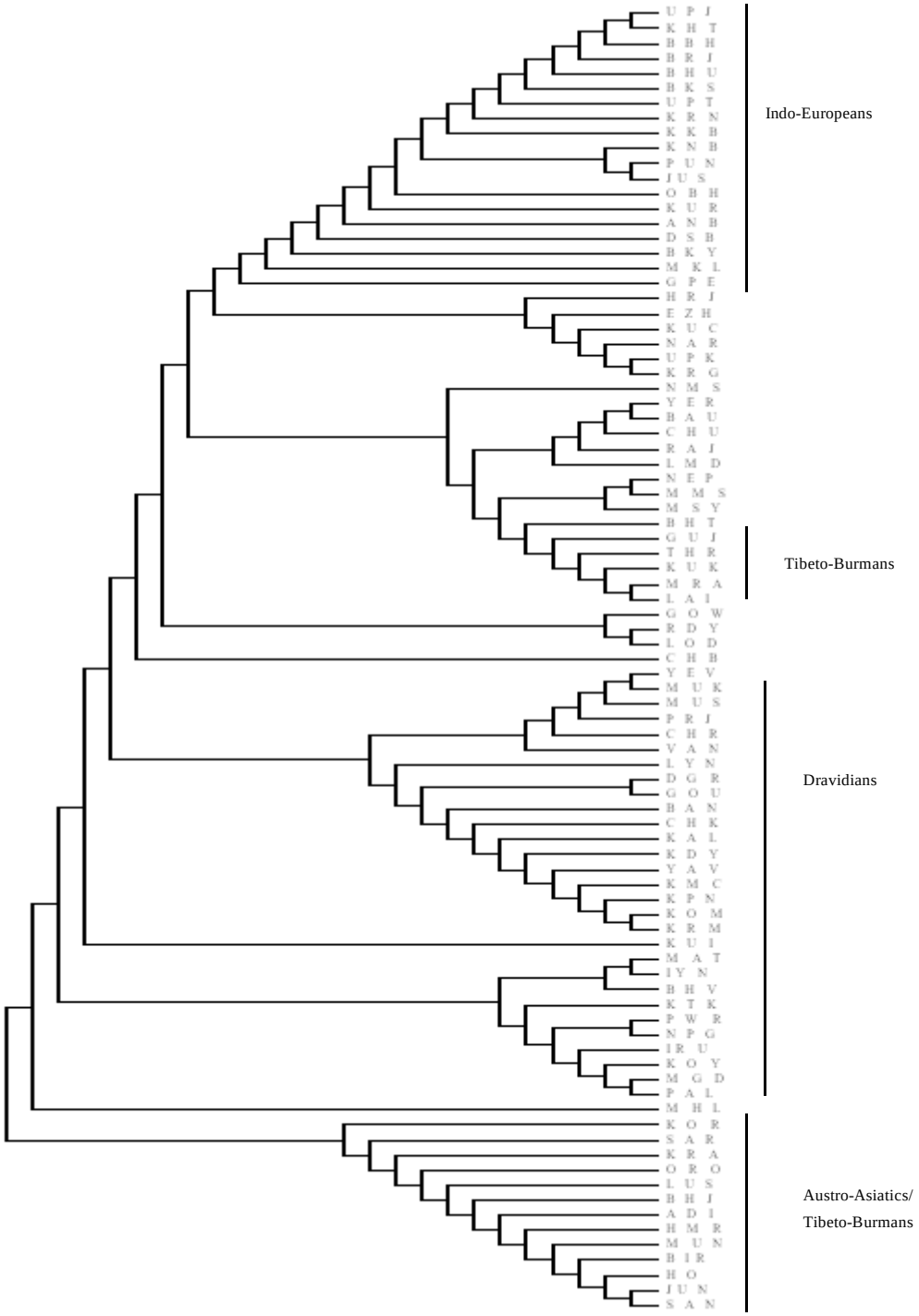

Fig. 3. Genetic relationship among populations of India based on F_{ST} distances estimated on Y-Haplogroup frequencies

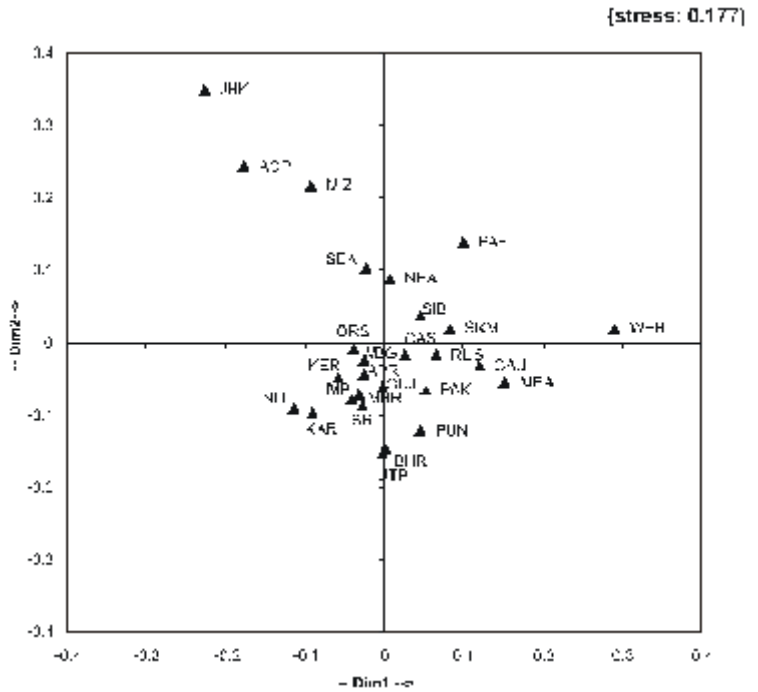

Fig. 4. Genetic relationship between populations of India and world estimated from Y-Chromosome haplogroup frequencies represented in MDS plot

genetic drift) and factors (geographical, linguistic and cultural barriers) that might have produced a high degree (27%) of genetic differentiation among the Indian patrilines. Here we also evaluate some of the suggested theories of occupation of Indian subcontinent by modern humans and population histories, in the light of current molecular genetic evidences.

Phylogeography of Indian Y-Chromosomes

India is a relict area, which is likely to have served as an incubator during the early dispersal of modern humans out of East Africa (Quintana-Murci et al. 1999; Cann 2001) and a treasure-house of ancient population genetic signatures in its gene pool. This is reflected in the 24 different haplogroups which were observed in the present Y-chromosome analysis of 1434 Indian males. Overall haplogroup diversity among Indian populations was relatively high (0.893) in contrast to other European or East Asian populations, but was closer to that of Central Asia. This pattern of high NRY diversity (Y-SNP and Y-STR) indicates an early settlement of the Indian subcontinent by anatomically modern humans. Four haplogroups; H= 23%; R1a1=17.5%; O2a=15% and R2=13.5%, form major paternal lineage of Indians and together account for ~70% of their Y-chromosomes. Being largely restricted to the Indian subcontinent, haplogroup H is assumed to be associated with the eastward expansion of M89 Y-chromosomes from the Leventine corridor, which also carried the two late Pleistocene mt DNA haplogroups, U2 and U7 into India. Although the M69 Y-chromosomes are particularly predominant among the Dravidian speakers of south India, its fairly uniform distribution across different regions and socio-ethnic groups of India suggests deep time depth for these lineage clusters.

Based on the predominance of M17 lineage among diverse linguistic families (Indo-European, Altaic, Uralic and Caucasian) and geographic regions (Central Asia, Europe, Caucasus, Middle East), (Wells et al. 2001; Underhill et al. 2000; Karafet et al. 1999; Rosser et al. 2000; Nebel et al. 2000) it has been associated with the Kurgan culture, domestication of horses and spread of Indo-European languages, all which supposedly

originated in southern Russia/Ukraine and subsequently extended to Europe, Central Asia around 3000 B.C (Wells et al. 2001). Its presence in India has been linked to the "Aryan" migration and subsequent spread of Indo-European languages, appearance of iron and Painted Grey Ware culture in North West frontier (Cavalli-Sforza et al. 1994). However, antiquity and geographic origin of this lineage still remains contentious. Our study reveals that this lineage is present in a significant proportion among the Indian Indo-European speakers, and is proportionately high among upper caste groups (Table 2a). The dispersion of this lineage into the southern tribal groups (Kivisild et al. 2003, Cordaux et al. 2004) and the fact that it is proportionately distributed between the Dravidian and Indo-European tribal groups provides significant evidence against any major influx of Indo-European speakers that could have drastically changed the Indian male gene pool (Sahoo et al. 2006; Sengupta et al. 2006). The high average STR variance (0.896) and TMRCA supports a rapid population growth and expansion of M17 Y-chromosomes, which contributed M17 lineages both to Central Asian nomads and South Asian tribes much before the Indo-European introgression into India. Another sub-lineage of M173, R1b3-M269 is present at appreciable frequencies 14.5% in Turkey (Cinnioglu et al. 2004) and at considerable frequency in Europe (Cruciani et al. 2002), while it is detected at relatively low frequency (1.9%) in India, substantiating a recent and limited admixture with west Europeans.

The observed high frequency of R2 Y-chromosomes in Indians, which is equivalent to that of haplogroup H among Dravidian speakers, corroborates previous reports suggesting its Indian origin (Cordaux et al. 2004). The deep coalescence time for R2 lineages, dating back to Late Pleistocene, supports its indigenous origin. Outside India, it is found in Iran and Central Asia (3.3%) and among Roma Gypsies of Europe, known to have historical evidence of their migration from India (Wells et al. 2001). Within India, while it is predominant in both eastern and southern regions, its distribution pattern is rather patchy in east (Sahoo et al. 2006). It is most likely that genetic drift or bottleneck has reduced the paternal diversity of Karmali, which contributes 28% of the eastern R2 lineages. This population although considered to be Austro-Asiatic speaker, does not present any evidence of O2a Y-chromosome lineage, portraying a distinctly different history.

On an average, the patterns of NRY haplogroup variation of Indians reflect that populations of the subcontinent are not very distinct from each and probably have share a few common paternal ancestors. The lineage diversity was small for Austro-Asiatic and Tibeto-Burman speakers and most of them harbored single lineage, indicating a founder paternal source for these endogamous groups, which are confined to the eastern and north-eastern regions of India. Haplogroup diversities were rather high for populations of south India (average of 0.740), giving concordant evidences of a relative early settlement, growth and expansion of populations living in southern India.

Traces of Ancient Migration of Modern Humans

Recent studies provide substantial evidences in favor of the southern route hypothesis for the dispersal of modern human ~ 60-75 kya from the horn of Africa along the tropical coast of Indian Ocean to reach insular South East Asia and Oceania (Cann 2001; Stringer 2000). A strong Y-chromosome support to this model is the distribution of haplogroup C lineages in Asia (Kivisild et al. 2003). Australo-Melanesia and North America (Karafet et al. 1999). Although present in low frequencies in Indian subcontinent, (Bamshad et al. 2001; Sengupta et al. 2006; Kivisild et al. 2003; Cordaux 2004; Wells 2006; Ramana et al. 2001) it is largely distributed along the coastal regions, with a few patchy occurrences in Punjab. However, the persistence of M130 lineages mostly among the south Indians, Pakistan (Qamar et al. 2002) and Sri Lanka (Kivisild et al. 2003) provides indirect evidence in support of the southern route of migration by early modern humans. We have previously suggested that lack of haplogroup C sub-lineages (M217, M38 and M8) is indicative of the indigenous origin of most Indian populations and argues against the theory of Aryan migration from Central Asia (Sahoo et al. 2006). The present analysis showed none of the deletions (DYS390.1 or DYS 390.3) associated with Australian or Polynesian C* chromosomes, contesting the claims of link between India and Australian aboriginals (Redd et al. 2002). . The age estimate of approximately 49KYA years in Indian samples indicates a probable Indian origin of this lineage. However, until further analysis

and age estimates in other world populations are known, it cannot be conclusively proven if RPS4YT mutation arose in India or arrived with the earliest migrants after it arose somewhere in west Asia, from where it was finally lost or diluted during the Upper Paleolithic expansion of modern humans (Underhill et al. 2001).

Genesis of Caste Structure and Influence of Migrations on the Indian Gene Pool

The main feature of Indian society is that it is highly structured by social factors such as caste system, in which birth determines the position in the society, mode of subsistence (occupation), and choice of marriage partners. However, genesis of caste system in India is ambiguous, since many of the caste groups are known to have tribal origins (Kosambi 1964). Further the migration of Greeks, Huns, Arabs, Chinese, Turks, Persians, Portuguese and others have made understanding the nature of population structure more complex. mt DNA analysis from different geographic region and social status showed that maternal haplogroups in India are derived from a limited number of founder lineages of M and N clades supporting a common proto-Asian ancestry with limited gene flow from later migrants (Kivisild et al. 2003; Basu et al. 2003). Our study reveals that there is virtually no genetic difference in the Y-chromosomes between the caste groups and tribes (Table 5). Whatever minor difference is present is largely due to haplogroup O2a, contributed exclusively by the Austro-Asiatic and Tibeto-Burman tribes. Our present analysis (AMOVA and haplogroup frequency distribution in populations excluding the Austro-Asiatic tribes of Jharkhand and Orissa and Tibeto-Burman tribes from Northeast) provides congruent evidence in support to the hypothesis that populations in India largely derive their gene pool from the common Pleistocene settlers. High frequency of J2 and R1a1 lineages mirror a greater influence of Indo-European migrants on upper caste populations of Gangetic plains compared to the peninsular southern regions. However, these skewed frequencies also suggest that the indigenous populations received limited external gene flow from Europe, Central and West Asia. This is also supported by mt DNA haplogroups that depict Indian-specific lineages with a limited contribution from both west and east Eurasian populations (Metspalu et al. 2004). A similar trend of J2 and R1a1 among caste populations would probably provide a simplistic assumption that agriculture was brought along with caste system by the Indo-European speakers as a result of demic diffusion of early farmers from southwestern Iran, Fertile Cresent and Anatolia (Quintana-Murci et al. 2001; Cordaux et al. 2004). However, the absence of other Neolithic markers of early farmers, M35 and M201; that are prevalent in Europe, Anatolia, South Caucasus and Iran (Semino et al. 2000; Underhill et al. 2001) among Indians, in addition to the frequency of M172 in southern and western India and its persistence in south India and tribal groups (Table 2a) questions the validity of this hypothesis. Agriculture in India probably arose as two independent events; one that was a consequence of earliest migration that brought the Dravidian speakers and another much later through spread of rice cultivators from SE Asia (Fuller 2003; Diamond et al. 2003).

Insights into Origin of Austro-Asiatic and Tibeto-Burman speakers

The origin of two language families, Austro-Asiatic and Tibeto-Burman in India is of particular interest and has received considerable attention (Basu et al. 2003; Cordaux et al. 2004). In the present study, analysis of eleven Austro-Asiatic and seven Tibeto-Burman tribes from the eastern and north-eastern region of India, establishes that the male gene pool of these groups are distinctly different from other mainland tribal populations (Table 2a). An overall low Y-STR haplotype diversity and complete fixation of O2a in some of the aboriginal tribes (Ho, Santhal, Juang, Birhor and Munda) suggests that these tribes probably experienced a major demographic event, such as common founder effect followed by a bottleneck that greatly reduced the Y-chromosome diversity in the Austro-Asiatic tribes of eastern India (Table 1). In contrast to Austro-Asiatic tribes, the Tibeto-Burman tribes harbor both O2a and O3e lineages in their Y-chromosomes. Interestingly, the two branches of the Tibeto-Burman language; Himalayish and Naga-Kuki-Chin, could be distinctly identified from their Y-chromosomes. The former depicts influence of Tibetan gene pool, marked by the presence of Haplogroup D lineages in Bhutia and Tharu (Sahoo et al. 2006), while the other linguistic branch harbors O3e lineages. The predominance of O haplogroup and its sub-

lineages in populations of East Asia suggest a SE Asian origin of Indian Austro-Asiatic and Tibeto-Burman speakers. We hypothesize that the Tibeto-Burman speakers came as a number of migratory events, while the Austro-Asiatic tribes probably arrived in India as a single event. The two groups probably migrated into India at different time period is evident from the absence of O3e lineages among Austro-Asiatic speakers, which probably are the earliest immigrants of the two. Presence of an Austro-Asiatic speaking tribe, Khasi, among the Tibeto-Burman speaking neighbors in the northeast corroborates this assumption. While the Tibeto-Burman speakers brought in a number of East Asian maternal lineages (A, B5b, F1b, M8c, M8z) (Metspalu et al. 2004), absence of these lineages in Austro-Asiatic tribes (Thangaraj et al. 2005; Sahoo 2006a) portrays two different scenarios. First, the earlier exodus from South East Asia was probably a major male–mediated migration into India, or that the female gene pool of the migrating East Asians is completely lost among the Austro-Asiatic tribes. Additional confirmation to this hypothesis is provided with evidences of agricultural expansions from their homelands in China, at different times and over different geographic ranges. Austro-Asiatics are presumed to have spread west and south from southern China into the Indian subcontinent and Malay Peninsula and brought rice cultivation with them (Higham 2003; Bellwood 2004). The genetic evidence revealed in this study is consistent with anthropological records (Guha 1935), which suggests that Sino-Tibetans dispersed from the Yellow River and came into India through two different routes; one from Burma probably brought the Naga-Kuki-Chin language and O3e Y-chromosomes and the other from Himalayas, which carried the YAP lineages into northern regions of subcontinent.

Age of Human Occupation in India-Austro-Asiatic or Dravidians as First Settlers?

Based on socio-cultural and linguistic evidences (Thapar 1995; Pattanayak 1998) and results based on mt DNA HVSI nucleotide diversity and highest frequencies of mitochondrial M haplogroup (Roychoudhary et al. 2001; Basu et al. 2003), it was asserted that Austro-Asiatic tribes are the earliest settlers in India. The present comprehensive Y-chromosome analysis, which includes populations of all linguistic and socio-ethnic affiliations, however, suggests people of south India as the original settlers of the subcontinent. The total lineage diversity and distribution of Indian-specific Y-chromosome haplogroups (H, L, C, R1a1 and R2) in different geographical and socio-linguistic layers of the Indian populations provides substantial support in favor of this hypothesis. This theory also gathers adequate evidence from presence of the coastal marker, RPS4Y, in the south Indian tribes, who probably represent remnants of the modern human migration out of Africa that took the southern route to Australia. Any possibility that Austro-Asiatic speakers could have dispersed from India is also eliminated based on the differential distribution of O2a Y-chromosomes in southern China and India and the complete absence of East-Asian specific mt DNA lineages in Austro-Asiatic and Dravidian speakers of India. mt DNA haplogroups of Indian Austro-Asiatic speakers are instead, probably a sub-group of their Dravidian neighbors (unpublished data, Kashyap et al.). Recent archeological and linguistic evidences corroborate a Neolithic expansion of Austro-Asiatic languages from Yangtze River basin (Higham 2003) and our present study supports an east-west clinal expansion of Austro-Asiatic males from South East Asia, which was not associated with any female gene flow. Further, deeper coalescence age for the Y-chromosome haplogroups C, H, R2 compared to O2a is consistent with hypothesis that Austro-Asiatic speakers cannot be considered as the earliest settlers of South Asia.

CONCLUSIONS

We find that genetic variation in India is characterized by a high Y-chromosome diversity, which is reflected by a greater correspondence with linguistic groups of India. Our results demonstrate India as a hotspot both as an important source and recipient of major Y-chromosome lineages of the world. Haplogroup distribution and AMOVA results provide tandem evidence in support a common Pleistocene origin of Indian populations, which was subsequently followed by migrations of Austro-Asiatic speaking tribal males from SE Asia. The Tibeto-Burman populations were later migrants who took two different routes and carried both male and female lineages specific to East Asia. Based on deep coalescence age estimates of H, R2 and C Y-

chromosome lineages, their diversity and distribution pattern, our data suggests an early Pleistocene settlement of South Asia by Dravidian speaking south Indian populations; the Austro-Asiatic speakers migrated much later from SE Asia and probably contributed only paternal lineages while amalgamating with the aboriginal populations of the region.

ACKNOWLEDGEMENTS

We express our appreciation to all the original donors who made this study possible. This study was made possible through facilities provided at CFSL, Kolkata. We acknowledge all researchers whose valuable data was used for this study. The SS, AS, JB, MT, SG, RR, RA are grateful to the Directorate of Forensic Sciences, MHA for the Senior Research Fellowship. GHB and TS are recipients of Senior Research Fellowship from CSIR, India. This research was supported by a financial grant to CFSL, Kolkata under the Xth Five Year Plan of the Govt. of India.

Electronic –Database Information

URLs for the data mentioned in this article are as follows:

XL STAT pro 7.5, http://www.xlstat.com

Network 4.1, http://www.fluxus-engineering.com

http://www.ethnologue.com

REFERENCES

Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, Naidu JM, Prasad BV, Reddy PG, Rasanayagam A, Papiha SS, Villems R, Redd AJ, Hammer MF, Nguyen SV, Carroll ML, Batzer MA, Jorde LB 2001. Genetic evidence on the origins of Indian caste populations. Genome Res 11: 994-1004

Bandelt HJ, Forster P, Rohl A 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol,* **16**: 37-48

Basu A, Mukherjee N, Roy S, Sengupta S, Banerjee S, Chakraborty M, Dey B, Roy M, Roy B, Bhattacharyya NP, Roychoudhury S, Majumder PP 2003. Ethnic India: A genomic view, with special reference to peopling and structure. *Genome Res,* **13**: 2277-2290.

Bellwood P 2004. Tracking the spreads of farming beyond the fertile Crescent: Europe and Asia. In: First Farmers: *The Origins of Agricultural Societies.* pp 87

Butler JM, Schoske R, Vallone PM, Kline MC, Redd AJ, Hammer MF 2002. A novel multiplex for simultaneous amplification of 20 Y chromosome STR markers. *Forensic Sci Int,* **129**: 10-24.

Cann RL 2001. Genetic clues to the dispersal of human populations: Retracing the past from the present. *Science,* **291**: 1742-1748

Cavalli-Sforza LL, Menozzi P, Piazza A 1994. *The History and Geography of Human Genes*. Princeton University Press, Princeton pp 208-213

Cinnioglu C, King R, Kivisild T, Kalfoglu E, Atasoy S, Cavalleri GL, Lillie AS, Roseman CC, Lin AA, Prince K, Oefner PJ, Shen P, Semino O, Cavalli-Sforza LL, Underhill PA 2004. Excavating Y-chromosome haplotype strata in Anatolia. *Hum Genet,* **114**: 127-148.

Cordaux R, Aunger R, Bentley G, Nasidze I, Sirajuddin SM, Stoneking M 2004. Independent origins of Indian caste and tribal paternal lineages. *Curr Biol,* **14**: 231-235.

Cordaux R, Deepa E, Vishwanathan H, Stoneking M 2004. Genetic evidence for the demic diffusion of agriculture to India. *Science,* **304**: 1125.

Cordaux R, Weiss G, Saha N, Stoneking, M 2004. The northeast Indian passageway: a barrier or corridor for human migrations? *Mol Biol Evol,* **21**: 1525-1533

Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V, Wallace DC, Oefner PJ, Torroni A, Cavalli-Sforza LL, Scozzari R, Underhill PA 2002. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet,* **70**: 1197-1214.

Deraniyagala SU 1992. *The Prehistory of Sri Lanka: An Ecological Perspective*. Colombo: Department of The Archeological Survey, Government of Sri Lanka

Diamond J, Bellwood P 2003. Farmers and their languages: The first expansions. *Science,* **300**: 597-602.

Excoffier L, Smouse PE, Quattro JM 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics,* **131**: 479-491.

Forster P, Rohl A, Lunnemann P, Brinkmann C, Zerjal T, Tyler-Smith C, Brinkmann B 2000. A short tandem repeat-based phylogeny for the human Y chromosome. *Am J Hum Genet,* **67**: 182-196

Fuller D 2003. An agricultural perspective on Dravidian historical linguistics: archaeological crop packages, livestock and Dravidian crop vocabulary. In: P Bellwood, C Renfrew (Eds.): *Examining the Farming/ Language Dispersal Hypothesis.* McDonald Institute for Archaeological Research, Cambridge. pp. 191-213.

Guha BS 1935. The racial affinities of the people of India. In: *Census of India, 1931,* Part III-Ethno-graphical

Higham C 2003. Languages and farming dispersals: Austro-Asiatic languages and rice cultivation. In: P Bellwood, C Renfrew (Eds.): *Examining the Farming/ Language Dispersal Hypothesis.* McDonald Institute for Archaeological Research, Cambridge

James HVA, Petraglia MD 2005. Modern Human origins and the evolution of behavior in the later Pleistocene Record of South Asia. *Curr Anthropol,* **46 Supp**: S3-S27

Karafet T, Xu L, Du R, Wang W, Feng S, Wells RS, Redd AJ, Zegura SL, Hammer MF 2001. Paternal

Population History of East Asia: Sources, Patterns and Microevolutionary Processes. *Am J Hum Genet,* **69**: 615-628

Karafet TM, Zegura SL, Posukh O, Osipova L, Bergen A, Long J, Goldman D, Klitz W, Harihara S, de Knijff P, Wiebe V, Griffiths RC, Templeton AR, Hammer MF 1999. Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. *Am J Hum Genet* **64**: 817-831

Kennedy K 2000. *God, Apes and Fossil Men: Paleoanthropology in South Asia*. Ann Arbor: University of Michigan Press

Kivisild T, Rootsi S, Metspalu M, Mastana S, Kaldma K, Parik J, Metspalu E, Adojaan M, Tolk HV, Stepanov V, Golge M, Usanga E, Papiha SS, Cinnioglu C, King R, Cavalli-Sforza L, Underhill PA, Villems R 2003. The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am J Hum Genet,* **72**: 313-332

Kosambi DD 1964. *The Culture and Civilization of Ancient India in Historical Outline,* New Delhi: Vikas Publishing House Pvt. Ltd.

Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F, Taha A, Shaari NK, Raja JM, Ismail P, Zainuddin Z, Goodwin W, Bulbeck D, Bandelt HJ, Oppenheimer S, Torroni A, Richards M 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science,* **308**: 1034-1036

Metspalu M, Kivisild T, Metspalu E, Parik J, Hudjashov G, Kaldma K, Serk P, Karmin M, Behar DM, Gilbert MT, Endicott P, Mastana S, Papiha SS, Skorecki K, Torroni A, Villems R 2004. Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet,* **5**: 26

Misra VN 2001 Prehistoric human colonization of India. *J Biosci,* **26**: 491-531

Nebel A, Filon D, Weiss DA, Weale M, Faerman M, Oppenheim A, Thomas MG 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

Passarino G, Cavalleri GL, Lin AA, Cavalli-Sforza LL, Borresen-Dale AL, Underhill PA 2002. Different genetic components in the Norwegian population revealed by the analysis of mtDNA and Y chromosome polymorphisms. *Eur J Hum Genet,* **10**: 521-529

Pattanayak DP 1998. The language heritage of India. In: Balasubramanian and NA Rao (Eds.): *The Indian Human Heritage.* Hyderabad. pp: 95-99

Qamar R, Ayub Q, Mohyuddin A, Helgason A, Mazhar K, Mansoor A, Zerjal T, Tyler-Smith C, Mehdi SQ 2002. Y-chromosomal DNA variation in Pakistan. *Am J Hum Genet,* **70**: 1107-1124

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

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

Ramana GV, Su B, Jin L, Singh L, Wang N, Underhill P, Chakraborty R 2001. Y-chromosome SNP haplotypes suggest evidence of gene flow among caste, tribe, and the migrant Siddi populations of Andhra Pradesh, South India. *Eur J Hum Genet,* **9**: 695-700.

Redd AJ, Roberts-Thomson J, Karafet T, Bamshad M, Jorde LB, Naidu JM, Walsh B, Hammer MF 2002. Gene flow from the Indian subcontinent to Australia: Evidence from the Y chromosome. *Curr Biol,* **12**: 673-677.

Renfrew C 1989. The origins of Indo-European languages. *Sci Am,* **261**: 82-90.

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

Roychoudhary S, Roy S, Basu A, Banerjee R, Vishwanathan H, Usha Rani MV, Sil SK, Mitra M, Majumder PP 2001. Genomic structures and population histories of linguistically distinct tribal groups of India. *Hum Genet,* **109**: 339-50.

Sahoo S and Kashyap VK 2006. Phylogeography of mitochondrial DNA and Y-Chromosome haplogroups reveal asymmetric gene flow in populations of Eastern India. *Am J Phys Anthropol,* (in press).

Sahoo S, Singh A, Himabindu G, Banerjee J, Sitalaximi T, Gaikwad S, Trivedi R, Endicott P, Kivisild T, Metspalu M, Villems R, Kashyap VK 2006. A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. *Proc Natl Acad Sci USA,* **103**: 843-848

Sambrook J, Fritsch EF, Maniatis T 1989. *Molecular Cloning. A Laboratory Manual.* 2nd Ed. CSHL Press, Cold Spring Harbor, NY

Schneider S, Roessli D, Excoffier L 2000. *ARLEQUIN ver 2.0.a software for Population Genetics Data Analysis.* Geneva: Genetics and Biometry Laboratory, University of Geneva.

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Underhill PA. 2000. The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science,* **290**: 1155-1159.

Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow CE, Lin AA, Mitra M, Sil SK, Ramesh A, Usha Rani MV, Thakur CM, Cavalli-Sforza LL, Majumder PP, Underhill PA 2006. Polarity and temporality of high-resolution Y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of central Asian pastoralists. *Am J Hum Genet,* **78**: 202-221.

Shi H, Dong YL, Wen B, Xiao CJ, Underhill PA, Shen PD, Chakraborty R, Jin L, Su B 2005. Y-chromosome evidence of southern origin of the East Asian-specific haplogroup O3-M122. *Am J Hum Genet,* **77**: 408-419

Singh, KS 1998. *India's Communities. National Series. People of India.* New Delhi: Oxford University Press.

Stringer C 2000. Coasting out of Africa. *Nature,* **405**: 24-25

Su B, Xiao C, Deka R, Seielstad MT, Kangwanpong D, Xiao J, Lu D, Underhill P, Cavalli-Sforza L, Chakraborty R, Jin L 2000. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. *Hum Genet,* **107**: 582-590

Thangaraj K, Sridhar V, Kivisild T, Reddy AG, Chaubey G, Singh VK, Kaur S, Agarawal P, Rai A, Gupta J, Mallick CB, Kumar N, Velavan TP, Suganthan R, Udaykumar D, Kumar R, Mishra R, Khan A, Annapurna C, Singh L 2005. Different population histories of the Mundari- and Mon-Khmer-speaking Austro-Asiatic tribes inferred from the mtDNA 9-bp deletion/ insertion polymorphism in Indian populations. *Hum Genet,* **116**: 507-517

Thapar R 1995 The first millennium B.C. in the northern India. In: R. Thaper (Ed.): *Recent Perspective of Early Indian History*. Bombay. pp. 80-141

Underhill P, Passarino G, Lin AA, Shen P, Mirazón Lahr M, Foley RA, Oefner PJ, Cavalli-Sforza LL 2001. The phylogeography of the Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet,* **65**: 43-62

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ 2000. Y chromosome sequence variation and the history of human populations. *Nat Genet,* **26**: 358-361

Wells RS, Yuldasheva N, Ruzibakiev R, Underhill P, Evseeva I, Blue-Smith J, Jin L, et al. 2001. The Eurasian heartland: A continental perspective on Y-chromosome diversity. *Proc Natl Acad Sci USA*, **98**: 10244-10249

Zhivotovsky LA, Underhill PA, Cinnioglu C, Kayser M, Morar B, Kivisild T, Scozzari R, Cruciani F, Destro-Bisol G, Spedini G, Chambers G., Herrera RJ, Yong KK, Gresham D, Tournev I, Feldman MW, Kalaydjieva L 2004. The Effective Mutation Rate at Y Chromosome Short Tandem Repeats, with Application to Human Population-Divergence Time. *Am J Hum Genet,* **74**: 50-61.