Ethnic India: A Genomic View, With Special Reference to Peopling and Structure

Analabha Basu, Namita Mukherjee, Sangita Roy, et al.

Genome Res. 2003 13: 2277-2290
Access the most recent version at doi:10.1101/gr.1413403

Ethnic India: A Genomic View, With Special Reference to Peopling and Structure

Analabha Basu,[1,4] Namita Mukherjee,[1,4] Sangita Roy,[2,4] Sanghamitra Sengupta,[1,4] Sanat Banerjee,[1] Madan Chakraborty,[1] Badal Dey,[1] Monami Roy,[1] Bidyut Roy,[1] Nitai P. Bhattacharyya,[3] Susanta Roychoudhury,[2] and Partha P. Majumder[1,5]

[1]Anthropology & Human Genetics Unit, Indian Statistical Institute, Calcutta 700 108, India; [2]Human Genetics & Genomics Department, Indian Institute of Chemical Biology, Calcutta, India; [3]Crystallography & Molecular Biology Division, Saha Institute of Nuclear Physics, Calcutta, India

We report a comprehensive statistical analysis of data on 58 DNA markers (mitochondrial [mt], Y-chromosomal, and autosomal) and sequence data of the mtHVS1 from a large number of ethnically diverse populations of India. Our results provide genomic evidence that (1) there is an underlying unity of female lineages in India, indicating that the initial number of female settlers may have been small; (2) the tribal and the caste populations are highly differentiated; (3) the Austro-Asiatic tribals are the earliest settlers in India, providing support to one anthropological hypothesis while refuting some others; (4) a major wave of humans entered India through the northeast; (5) the Tibeto-Burman tribals share considerable genetic commonalities with the Austro-Asiatic tribals, supporting the hypothesis that they may have shared a common habitat in southern China, but the two groups of tribals can be differentiated on the basis of Y-chromosomal haplotypes; (6) the Dravidian tribals were possibly widespread throughout India before the arrival of the Indo-European-speaking nomads, but retreated to southern India to avoid dominance; (7) formation of populations by fission that resulted in founder and drift effects have left their imprints on the genetic structures of contemporary populations; (8) the upper castes show closer genetic affinities with Central Asian populations, although those of southern India are more distant than those of northern India; (9) historical gene flow into India has contributed to a considerable obliteration of genetic histories of contemporary populations so that there is at present no clear congruence of genetic and geographical or sociocultural affinities.

[Supplemental Material is available online at www.genome.org. The following individuals kindly provided reagents, samples, or unpublished information as indicated in the paper: C.S. Chakraborty, R. Lalthantluanga, M. Mitra, A. Ramesh, N.K. Sengupta, S.K. Sil, J.R. Singh, C.M. Thakur, M.V. Usha Rani, L. Jorde, K. Kidd, A. Merriwether, A. Torroni, and C. Tyler-Smith.]

India has served as a major corridor for the dispersal of modern humans (Cann 2001). The date of entry of modern humans into India remains uncertain. By the middle Paleolithic period (50,000–20,000 years before present [ybp]), humans appear to have spread to many parts of India (Misra 1992). The migration routes of modern humans into India remain enigmatic, and whether there were also returns to Africa from India/Asia is unclear (Maca-Meyer et al. 2001; Roychoudhury et al. 2001; Cruciani et al. 2002). Contemporary ethnic India is a land of enormous genetic, cultural, and linguistic diversity (Karve 1961; Beteille 1998; Majumder 1998). The people of India are culturally stratified as tribals, who constitute 8.08% of the total population (1991 Census of India), and nontribals. There are ~450 tribal communities in India (Singh 1992), who speak ~750 dialects (Kosambi 1991) that can be classified into one of the following three language families: Austro-Asiatic (AA), Dravidian (DR), and Tibeto-Burman (TB). Most contemporary nontribal populations of India belong to the Hindu religious fold and are hierarchically arranged in four main caste classes, namely, Brahmin (priestly class), Kshatriya (warrior class), Vysya (business class), and Sudra (menial labor class). In addition, there are several religious communities, who practice different religions, namely, Islam, Christianity, Sikhism, Judaism, and so on. The nontribals predominantly speak languages that belong to the Indo-European (IE) or Dravidian families. The IE and DR groups have been the major contributors to the development of Indian culture and society (Meenakshi 1995). Indian culture and society are also known to have been affected by multiple waves of migration and gene flow that took place in historic and prehistoric times (Ratnagar 1995; Thapar 1995). In a recent study conducted on ranked caste populations sampled from one southern Indian State (Andhra Pradesh), Bamshad et al. (2001) have found that the genomic affinity to Europeans is proportionate to caste rank—the upper castes being most similar to Europeans, particularly East Europeans, whereas the lower castes are more similar to Asians. These findings are consistent with the migration of IE groups into India, the establishment of the caste system, and subsequent recruitment of indigenous people into the caste fold. Because the Indian samples for this study were drawn from one geographical area, whether we can safely generalize these findings needs to be investigated.

The tribals are possibly the original inhabitants of India (Thapar 1966; Ray 1973), although their evolutionary histories and biological contributions to the nontribal populations have been debated (Risley 1915; Guha 1935; Sarkar 1958). Therefore, it is crucial to carry out genetic investigations in geographically

[4]These authors have contributed equally to this work.
[5]Corresponding author.
E-MAIL ppm@isical.ac.in; FAX 91-33-2577 3049.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1413403.

and culturally disparate, but ethnically well-defined, populations, using data on a uniform set of mitochondrial (mt), Y-chromosomal, and autosomal DNA markers. Unfortunately, the vast majority of earlier studies on Indian populations have been conducted on ethnically ill-defined populations or have been restricted to a single geographical area or a single set of markers—primarily either mitochondrial or Y-chromosomal (e.g., Kivisild et al. 1999a; Bamshad et al. 2001). The objectives of the present study are to (1) provide a comprehensive view of genomic diversity and differentiation in India, and (2) to draw inferences on the peopling of India, and the origins of the ethnic populations, specifically in relation to the various competing hypotheses, such as whether the Austro-Asiatic or the Dravidian-speaking tribal groups were the original inhabitants of India (Risley 1915; Guha 1935; Sarkar 1958).

We analyzed genetic variation in 44 geographically, linguistically, and socially disparate ethnic populations of India (Table 1). These include 10 restriction site polymorphisms (RSPs), one insertion/deletion (InDel) polymorphism, and hypervariable segment 1 (HVS1) sequences on mtDNA; 11 RSPs, 1 InDel, and 10 short tandem repeat (STR) loci on Y-chromosomal DNA; and 8 InDel and 17 RSPs on autosomal DNA.

RESULTS

Distribution of mtDNA Lineages Indicates a Small Founding Group of Females

The *Hpa*I np 3592 mtDNA restriction site locus was monomorphic in all populations. We observed 32 distinct 10-locus mtDNA

Table 1. Names of Study Populations, Sample Sizes, Linguistic, and Ethnological Information

	Sample size					
	mt					
Population name[a] [code]	RSP	HVS1 seq	Y	Autosomal	Linguistic affiliation	Social category
1. Agharia [AGH][3]	24	10	9	24	Indo-European	Middle caste
2. Ambalakarer [AMB][4]	30	10	18	50	Dravidian	Middle caste
3. Bagdi [BAG][3]	31	10	11	31	Indo-European	Lower caste
4. Chakma [CHK][2]	10	10	4	10	Tibeto-Burman	Tribe
5. Chamar [CHA][1]	25	10	18	25	Indo-European	Lower caste
6. Gaud [GAU][3]	13	10	4	15	Indo-European	Middle caste
7. Gond [GND][6]	51	10			Dravidian—Gondi dialect	Tribe
8. Halba [HAL][6]	47	20	20	48	Indo-European Primarily Marathi	Tribe
9. Ho [HO][3]	54	10	20	54	Austro-Asiatic	Tribe
10. Irula [ILA][4]	30	14	18	50	Dravidian	Tribe
11. Iyengar [IYN][4]	30	10	20	51	Dravidian	Upper caste
12. Iyer [IYR][4]	30	10	20	50	Dravidian	Upper caste
13. Jamatiya [JAM][2]	55	10	16	55	Tibeto-Burman	Tribe
14. Jat Sikh [JSK][1]	48	15			Indo-European	Middle caste
15. Kamar [KMR][6]	54	10	19	57	Dravidian	Tribe
16. Khatris [KHT][1]	48	15			Indo-European	Middle caste
17. Konkan Brahmins [KBR][5]	31	10			Indo-European	Upper caste
18. Kota [KOT][4]	30	25	15	45	Dravidian	Tribe
19. Kurumba [KUR][4]	30	10	18	54	Dravidian	Tribe
20. Lodha [LOD][3]	32	14	17	32	Austro-Asiatic	Tribe
21. Mahishya [MAH][3]	33	10	9	34	Indo-European	Middle caste
22. Manipuri (Meitei) [MNP][2]	11	9			Tibeto-Burman	Upper caste
23. Maratha [MRT][5]	41	10			Indo-European	Middle caste
24. Mizo [MZO][2]	29	14	20	29	Tibeto-Burman	Tribe
25. Mog [MOG][2]	25	10	6	25	Tibeto-Burman	Tribe
26. Munda [MUN][3]	7	6		49	Austro-Asiatic	Tribe
27. Muria [MUR][6]	30	12	8	28	Dravidian—Gondi dialect	Tribe
28. Muslim [MUS][1]	28	10	19		Indo-European	Islamic religious group
29. Naba-Baudh [NBH][5]	40	10			Indo-European	Lower caste (recently adopted Buddhism)
30. Pallan [PLN][4]	30	10	15	50	Dravidian	Lower caste
31. Punjab Brahmins [PBR][1]	48	12			Indo-European	Upper caste
32. Rajput [RAJ][1]	51	10	35	52	Indo-European	Middle caste
33. Riang [RIA][2]	51	12	17	50	Tibeto-Burman	Tribe
34. Santal [SAN][3]	20	14	15	24	Austro-Asiatic	Tribe
35. Saryupari Brahmins [SBR][6]	26	19			Indo-European	Upper caste
36. Scheduled caste-Punjab [SCH][1]	48	15			Indo-European	Lower caste
37. Tanti [TAN][3]	16	10	6	16	Indo-European	Lower caste
38. Tripperah (Tripuri) [TRI][2]	51	20	17	50	Tibeto-Burman	Tribe
39. Toda [TOD][4]	50	10	8	50	Dravidian	Tribe
40. Toto [TTO][3]	30	20	12	30	Tibeto-Burman	Tribe
41. Uttar Pradesh Brahmins [UBR][1]	27	10	17	27	Indo-European	Upper caste
42. Vanniyar [VAN][4]	30	10	14	50	Dravidian	Middle caste
43. Vellala [VLR][4]	43	10	16	43	Dravidian	Middle caste
44. West Bengal Brahmins [WBR][3]	22	10	13	23	Indo-European	Upper caste

[a] Geographical location: 1 = North, 2 = Northeast, 3 = East, 4 = South, 5 = West, and 6 = Central.

RSP haplotypes (maximum number within a population = 13, for Rajput and Tipperah; minimum = 2, for Kota and Toda) among the 1490 individuals studied from the 44 populations, five (15.7%) of which were exclusive to the northeastern populations. (Haplotype frequencies are provided in Supplemental Table 1, available online at www.genome.org.) Although the frequency distributions of haplotypes among populations differed significantly ($p < 0.05$), one haplotype, belonging to haplogroup (HG) M, accounted for 46.4% of all mtDNA molecules. This modal haplotype in the pooled data set was also the modal haplotype in 34 (77%) of the 44 study populations. The 10 populations in which this haplotype was not the most frequent primarily comprised ethnic groups of the northern (Uttar Pradesh Brahmins, Punjab Brahmins, Rajput, Muslim) and the northeastern regions (Chakma, Jamatiya, Mog, Toto). In these populations, the modal haplotypes belong to non-M HGs, especially HG-U. However, even in these populations, the haplotype that was modal in the overall data set was generally the second most frequent haplotype. Thus, the distribution of mtDNA haplotypes shows that there is a strong uniformity of female lineages in India.

Among the 528 individuals for whom mtDNA HVS1 sequences were determined, the total number of distinct sequences observed was 323, of which 91 (28.2%) were shared by multiple individuals. In all, 298 (56.4%) individuals shared HVS1 sequences. Thus, there was a significant sharing both in the number of sequences and in the proportion of individuals sharing these sequences. This reinforces our earlier conclusion of uniformity of female lineages. The most parsimonious explanation of these findings is that there was a small number of founding female lineages in Indian populations. The small number of founding female lineages may have either resulted from a founder effect caused by a small number of women entering India or, possibly more likely, caused by the group of founding females, irrespective of the size of the group, being drawn from an ancestral population with a relatively homogeneous pool of mitochondrial haplotypes.

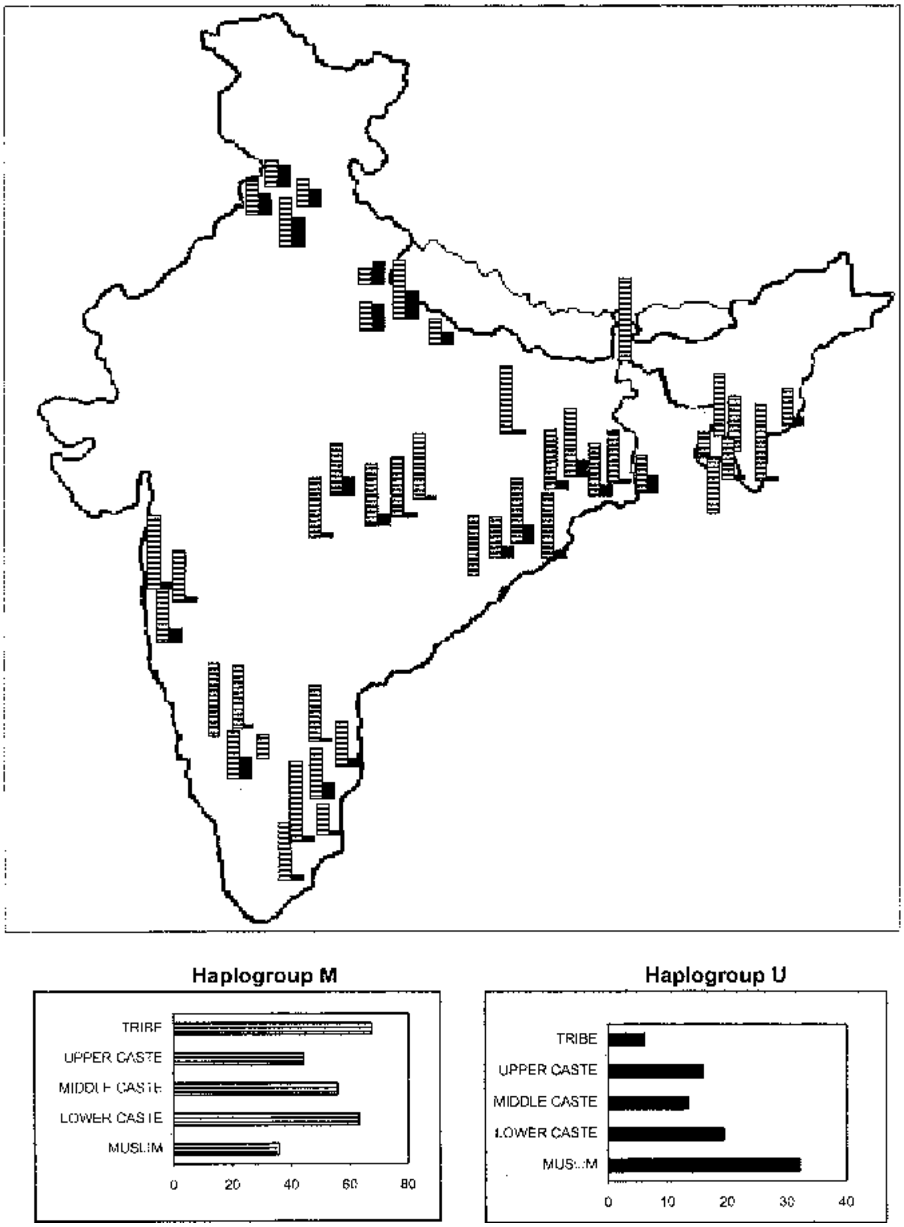


Figure 1 Frequencies (%) of mitochondrial haplogroups M (hatched) and U (solid black) in 44 ethnic populations, and among sociocultural groups of populations (*insets*).

Haplogroup Distributions Are Largely Concordant With Gene Flow Into India and Provide Insights Into Population Structure

The frequencies of the most predominant mtDNA HGs in India, M and U, are roughly inversely correlated (Fig. 1). HG-M frequency is very high (overall 59.9%: range 18.5% [Brahmins of Uttar Pradesh] to 96.7% [Kota]), confirming that it is an ancient marker in India. HG-M frequency is the highest among tribal groups, particularly in the AA tribals. Among HG-M individuals, 98.22% belong to subHG-M*, defined by the presence of T at np 16,223. Figure 2A presents the frequencies of various known (Bamshad et al. 2001) subHGs of M* among different sociocultural categories. (Detailed data are given in Suppl. Tables 2–4.) Individuals belonging to subHG-M2 had the highest nucleotide diversity in HVS1, indicating that M2 may be the most ancient in India. It occurs in significantly higher ($p < 0.05$) frequencies among tribals (28%), particularly among the AA tribals (32%), than among castes (8.8%). Furthermore, the coalescent time of M2 found in India was estimated to be greater than most east Asian and Papuan branches of HG-M (Forster et al. 2001), indicating that India was settled early after humankind came out of Africa (Kivisild et al. 1999b). These findings imply that the contemporary tribals are descendants of the initial settlers. HG-U is a complex mtDNA lineage, whose age was estimated from our data to be 45,000 ± 25,000 years, not significantly different from an

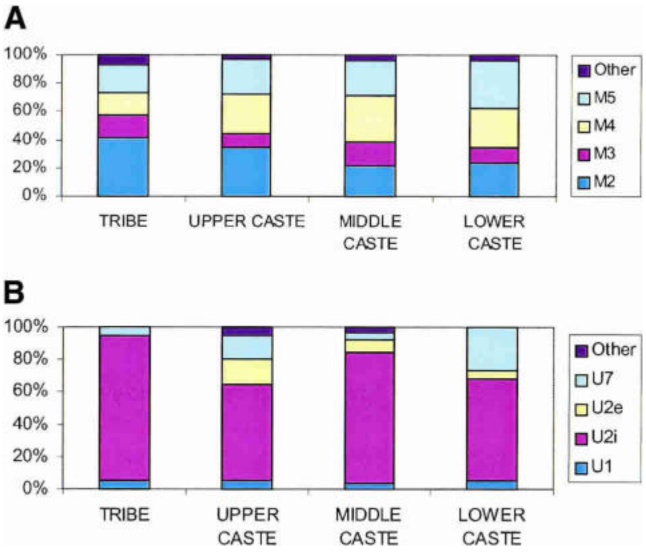


Figure 2 Frequencies (%) of subhaplogroups of (*A*) M and (*B*) U among tribal and ranked caste populations.

earlier estimate (Torroni et al. 1996). Its frequency is significantly ($p < 0.001$) higher among the IE-speaking caste groups, compared with other caste or tribal groups. Of particular interest are the frequencies of subHGs U2i (Indian-specific cluster of subHG U2 that predated the arrival of IE speakers from Central and West Asia into India; Kivisild et al. 1999a), U2e (Western-Eurasian cluster of U2), and U7 (an ancient Indian subHG). The frequencies of the subHGs of U are presented in Figure 2B. It is striking that the tribals do not possess U2e, and have the highest frequency of U2i. The gradations in frequencies (Figs. 1 and 2) of HG-M, particularly of subHG-M2, and also of HG-U, notably the absence of U2e among tribals, indicate that (1) tribals are more ancient than the castes, (2) there has been considerable admixture with Central and West Asians during the formation of the caste system, and (3) many new female lineages were introduced by the IE speakers.

Because U2i appears to be the indigenous subHG of U in India, we sought to find phylogenetic relationships among the distinct HVS1 sequences within this subHG and their observed frequencies in various linguistic groups. The phylogenetic network (Bandelt et al. 1999) is complex (Suppl. Fig. 1). There is no major starlike cluster to indicate sudden population expansions, nor is there any clear sociolinguistic clustering of related sequences.

The frequency differences among ethnic groups of the Y-HGs are dramatic (Fig. 3). The tribals, irrespective of linguistic group, possess significantly lower ($p < 0.0001$) frequencies of HG-P*, which probably arose in Central Asia (Zerjal et al. 2002), than castes (Table 2). The distribution of HG-BR* frequencies, which is the most ancestral lineage in Europe (Rosser et al. 2000), is rather inconsistent with IE admixture. The tribals possess higher frequencies than castes. However, there are significant differences ($p < 0.0001$) in frequencies among tribals belonging to the TB (8%), AA (27%), and DR (65%) tribals. These inconsistencies and differences may be due to the fact that HG-BR* probably contains a heterogeneous set of chromosomes that are not closely related (Zerjal et al. 2002). We could not find phylogenetic clades within this HG. Dramatically higher frequencies of HG-K* were observed among TB (72%) and AA (52%) tribals, compared with other tribal or caste groups (Table 2). The age of this HG (~20,000 yr) is 5000–15,000 yr higher than other HGs (Table 3). Contrary to Quintana-Murci et al.'s (2001) suggestion of a cline of HG-J frequencies from the Middle East (where this HG has high frequencies) into India, resulting from diffusion of people with agriculture, we do not find any clear cline with the addition of new data.

Anthropologists and historians have contended (Karve 1961; Kosambi 1991) that fission as a process has been very widespread in the formation of ethnic subgroups in India, resulting not only from pressures on natural resources but also because of social regulations. The common genetic consequences of fission are founder effect and drift, which are expected to result in high frequencies of haplotypes or nucleotide motifs in the daughter populations that are infrequent in the parental populations. Evidence of this is clearly seen from an evolutionary network, considerably simplified to highlight the principal features, of HVS1 sequences of individuals belonging to HG-M (Fig. 4). Many motifs occur exclusively or in high frequencies in some populations or groups or regions, which is consistent with fission and founder effect. Similarly, a motif GCGC at nps 16,051, 16,206, 16,230, and 16,311, respectively, was found in 16% of individuals belonging to subHG U2i. This motif occurs exclusively among tribal, middle-, and lower-caste populations, but not among the upper-caste populations. Evidence of fission and founder effect are also discernible from Y-STRP haplotype data: One 10-site haplotype is shared by several Kamar and Kota tribals, whereas another is shared by several Kurumbas. There are also large differences in the numbers of Y-STRP haplotypes and haplotype diversities among populations. (Detailed data are presented in Suppl. Tables 5 and 6.)

Austro-Asiatic-Speaking Tribals May Be the Earliest Inhabitants of India

Sociocultural and linguistic evidence indicates (Risley 1915; Thapar 1966; Pattanayak 1998) that the AA tribals are the original inhabitants of India. Some other scholars have, however, argued that tribal groups speaking DR and AA languages have evolved from an older original substrate of proto-Australoids (Keith 1936), whereas the TB tribals are later immigrants from Tibet and Myanmar (Guha 1935). Our findings strongly support the hypothesis that AA tribals are the earliest inhabitants of India. They possess the highest frequencies of the ancient east-Asian mtDNA HG-M and exhibit the highest HVS1 nucleotide diversity (Table 4). They also have the highest frequency of subHG M2 (19%), which had the highest HVS1 nucleotide diversity compared with other subHGs and therefore possibly the earliest settlers (the estimated coalescence time is 63,000 ± 6000 ybp; Kivisild et al. 1999a). Although all sociolinguistic groups

Table 2. Frequencies of Y Haplogroups in Ethnic Populations of India Belonging to Various Sociolinguistic Groups

Linguistic group	Social group	Sample size	Haplogroup (%)					
			P*	BR*	R1a	J	K*	L
Austro-Asiatic	Tribe	52	7.7	28.8	0.0	13.5	50.0	0.0
Dravidian	Tribe	84	4.8	67.9	3.6	9.5	6.0	8.2
	Caste	103	19.4	32.2	11.6	22.3	4.8	9.7
Tibeto-Burman	Tribe	87	4.7	10.3	0.0	14.9	70.1	0.0
Indo-European	Tribe	19	0.0	36.8	15.8	21.1	26.3	0.0
	Caste	122	22.1	32.8	23.8	12.3	6.6	2.4

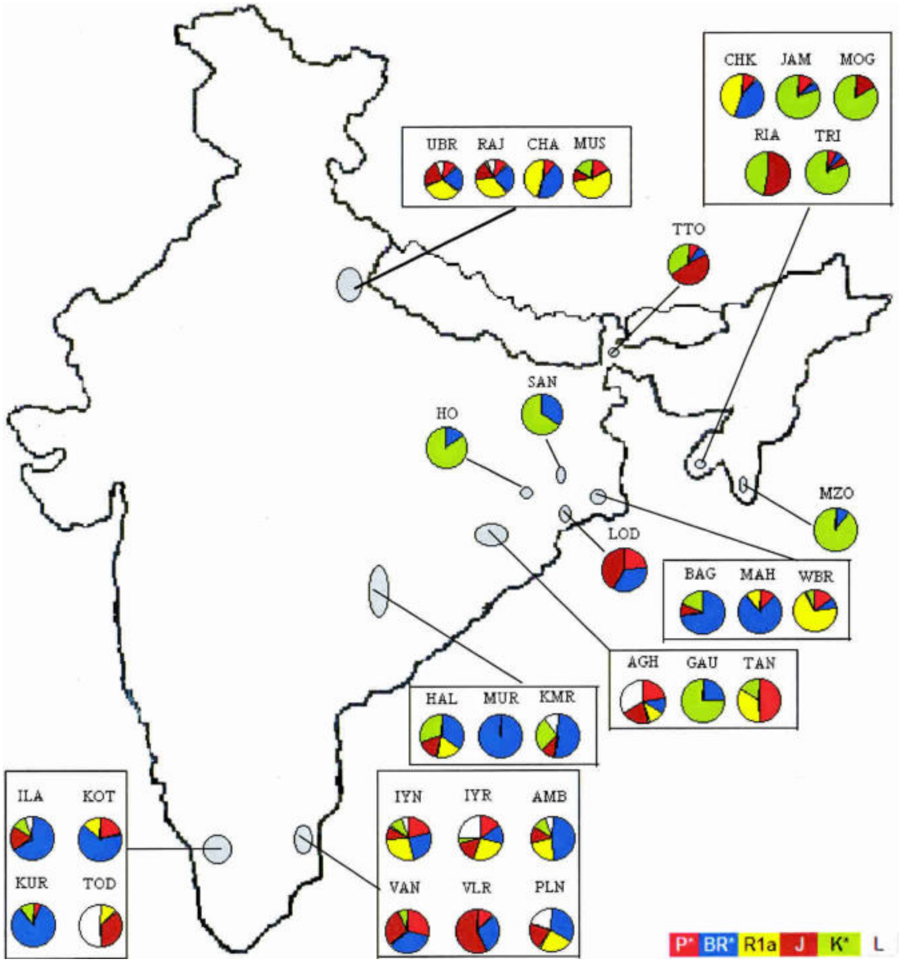


Figure 3 Frequencies (%) of Y-chromosomal haplogroups among ethnic populations. (Population codes are given in Table 1.)

seem to have undergone significant population expansions as evidenced by the unimodality of the HVS1 mismatch distributions (data not shown) and by the values of the relevant statistics (small values of the "raggedness" statistic and significantly large negative values of Fu's F_s statistic; Table 4), the AA tribals show the highest value of the estimated expansion time, ~55,000 years, which is ~15,000 years larger than the estimates for the other groups. Although we cannot be sure that this expansion took place in India, in conjunction with the other findings, it appears that this group of tribals may be the earliest inhabitants of India. A young subclade M4, with an estimated coalescence time of 32,000 ± 7500 ybp (Kivisild et al. 1999a), whose overall frequency is ~15% in India, is completely absent among them. It is, therefore, likely that M4 arose after the expansion of the AA tribals and their entry into India.

The Northeastern Corridor May Have Served as a Major Passage of Entry Into India

High frequencies of Y-HG K* (Fig. 3) are found among the TB populations, mainly confined to northeast India, and also among the Han Chinese (Su et al. 2000). The TB subfamily of the Sino-Tibetan language family has been subdivided (Grimes 1999) into four branches: Baric, Bodic, Burmese-Lolo, and Karen. Based on a study of Y-chromosomal haplotypes, Su et al. (2000) have contended that after the proto-Tibeto-Burman people left their homeland in the Yellow River basin, the Baric branch moved southward and peopled the northeastern Indian region after crossing the Himalayas. This branch did not possess the YAP insertion element, which has also not been found in any of the TB populations of India. Our findings, therefore, are consistent with Su et al.'s (2000) inference and indicate that the TB speakers entered India from the northeastern corridor. It is, however, surprising that the AA tribals, who primarily inhabit the eastern and central Indian regions, also possess high frequencies of Y-HG K* (Fig. 3). There is one major tribal group (Khasi) of northeastern India, who speak a dialect that belongs to the AA subfamily. Besides India, AA languages are spoken in south-east Asia. Thus, it is likely that a fraction of the AA tribals also entered India through the northeastern corridor. However, it does not seem that all of them have entered through this corri-

Table 3. Estimated Ages of Various Y-Chromosomal Haplogroups

Estimate (yr)	Haplogroup					
	P*	BR*	R1a	J	K*	L
Age	16,322	12,508	5416	15,300	20,049	14,290
Lower 95% CI	9477	7263	3145	8884	11,641	8297
Upper 95% CI	29,979	22,974	9947	28,103	36,824	26,246

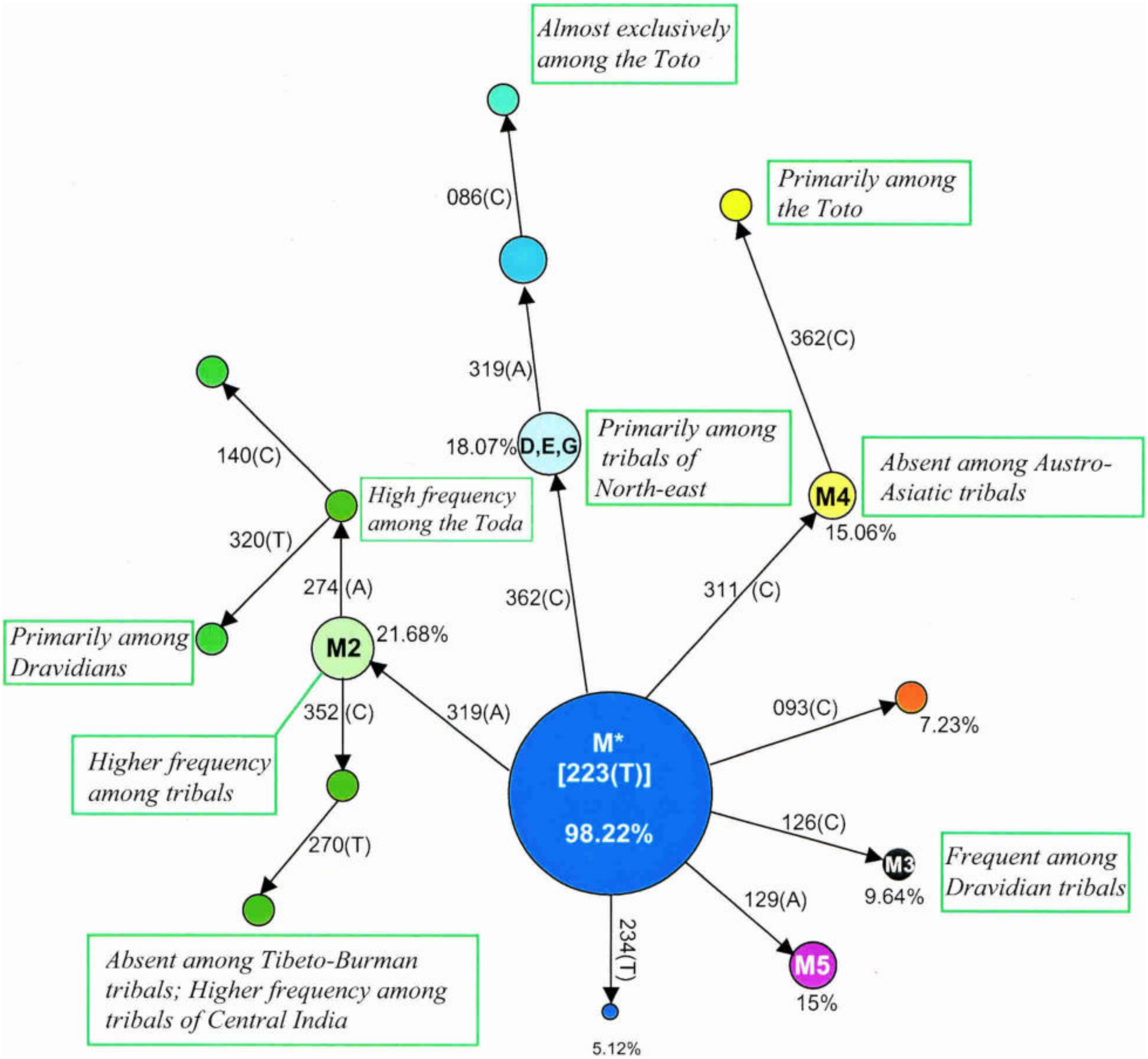

Almost exclusively among the Toto
086(C)
Primarily among the Toto
362(C)
319(A)
140(C)
18.07%
D,E,G
Primarily among tribals of North-east
High frequency among the Toda
M4
Absent among Austro-Asiatic tribals
15.06%
320(T)
274 (A)
362(C)
311 (C)
Primarily among Dravidians
M2
21.68%
352 (C)
319(A)
093(C)
7.23%
Higher frequency among tribals
M*
[223(T)]
98.22%
270(T)
126(C)
M3
9.64%
Frequent among Dravidian tribals
129(A)
234(T)
M5
15%
Absent among Tibeto-Burman tribals; Higher frequency among tribals of Central India
5.12%

Figure 4 Phylogenetic network of mtHVS1 sequences belonging to subhaplogroup M*, with frequency distributions of motifs in populations.

dor. The expansion time of AA tribals was estimated to be ~55,000 yr (Table 4), which is ~13,000 yr greater than that estimated for TB tribals. But the age of the Y-HG K* estimated from the pooled variance of repeat numbers at the STRP loci among AA tribals (8500 yr) is about half of that estimated for TB tribals (15,000 yr). The AA tribals can also be clearly distinguished (estimated probability of correct identification = 0.8) from the TB tribals on the basis of haplotype frequencies at three Y-STRP loci (Fig. 5). We, therefore, believe that the ancestors of the present-day AA tribals in India initially entered India either through the northwestern corridor or through a southern exit route from Africa, and then later through the northeastern corridor during the Austronesian diaspora from southern China through southeast Asia to Papua New Guinea (Diamond 1997). It is possible that the ancestors of the AA speakers entered India through the northwest from out-of-Africa as they moved south of the Himalayas, and another ancestral group moved north of the Himalayas, settled in southern China, and then entered later through the northeast.

Genetic Differentiation and Affinities

Estimation of F_{ST} statistics and AMOVA were carried out among populations grouped by well-defined criteria, separately for the different sets of markers, mt DNA (RSP and HVS1), Y-chromosomal (RSP and STRP), and autosomal. (Detailed allele and haplotype frequency data for autosomal loci are provided in Suppl. Tables 7 and 8.) The results for the different sets of markers are qualitatively similar. The F_{ST} values are highest for tribal populations inhabiting different geographical regions or belonging to different linguistic groups (Table 5). We note that there is some degree of confounding between geographical region of habitat and linguistic groups (see Methods). Genomic differentiation among the caste groups is substantially smaller than among the tribal groups. With respect to Y STRPs, the upper

Table 4. Numbers of Polymorphic Sites, Nucleotide Diversities, Mismatch Statistics, Estimated Expansion Times, and Other Statistics Pertaining to Population Expansion Based on HVS1 Sequence Data Classified by Social and Linguistic Affiliation

	Tribe			Caste	
	Linguistic group			Linguistic group	
	Austro-Asiatic	Dravidian	Tibeto-Burman	Dravidian	Indo-European
No. of sequences	46	94	95	60	195
No. of polymorphic sites	57	55	74	63	115
Nucleotide diversity (π) ± SD	0.0224 ± 0.0059	0.0170 ± 0.0045	0.0173 ± 0.0046	0.0154 ± 0.0042	0.0176 ± 0.0046
Mean no. of mismatches ± SD	8.00 ± 1.89	6.07 ± 1.46	6.16 ± 1.48	5.51 ± 1.35	6.28 ± 1.50
Expansion time (y) (range)	54,656 (40,243–69,024)	41,470 (30,488–52,439)	42,085 (30,976–53,415)	37,644 (27,439–47,683)	42,905 (31,707–54,146)
Raggedness, *r*	0.0199	0.0069	0.0200	0.0094	0.0097
Fu's *Fs* (*p* value)	−24.93 (0.0)	−25.26 (0.0)	−25.20 (0.0)	−25.45 (0.0)	−24.87 (0.0)

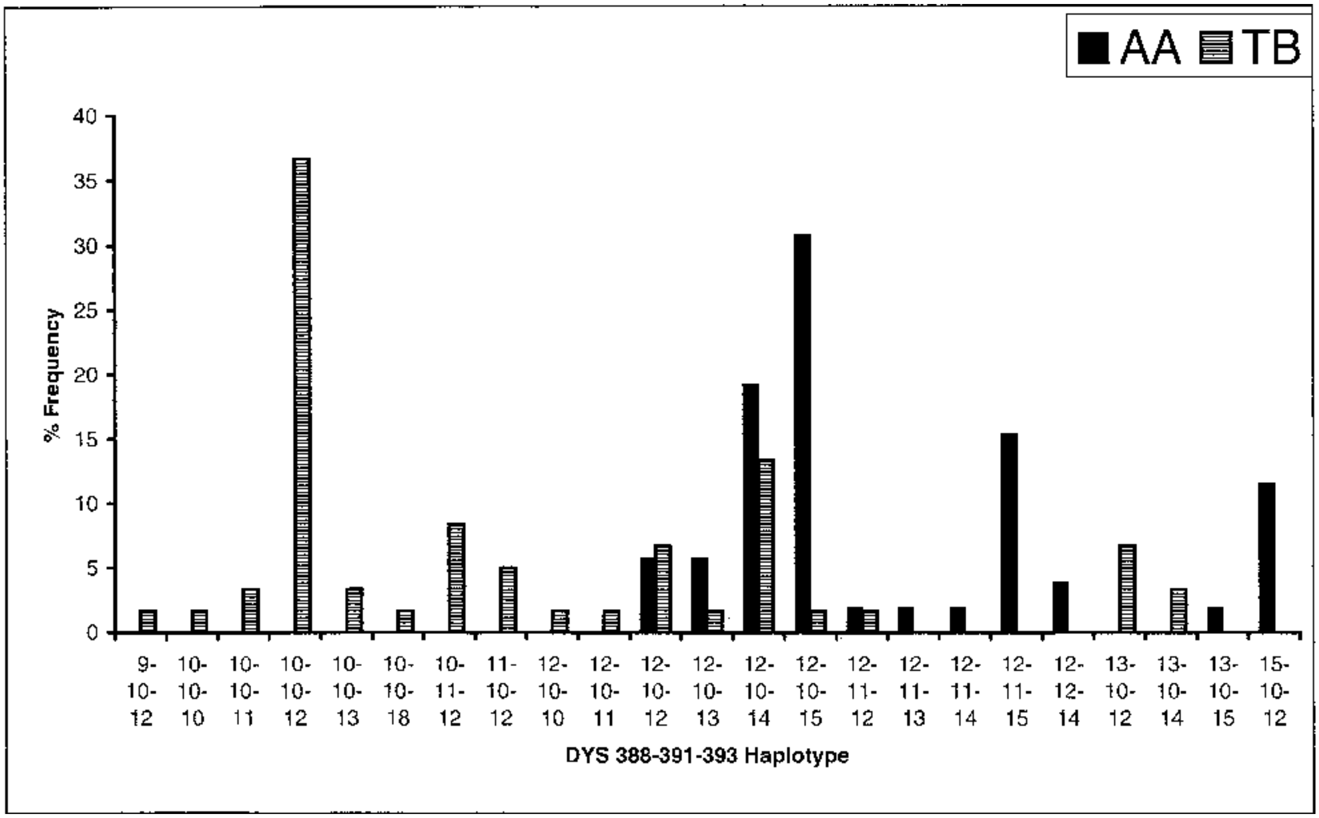


Figure 5 Frequency distributions of Y-chromosomal STR haplotypes that best discriminate between the Austro-Asiatic (AA) and Tibeto-Burman (TB) population groups.

castes of different geographical regions show stronger differentiation compared with middle or lower castes, possibly reflecting historical male gene flow. The F_{ST} values based on autosomal markers are less than those for mitochondrial and Y-chromosomal markers, which is expected because of stronger drift effects arising from the fact that the effective size of a population with respect to mt and Y markers is a quarter of that for autosomal markers. F_{ST} values for Y markers are higher than for mtDNA markers possibly because of social practices that enhance female mobility compared with male mobility (Bamshad et al. 1998; Bhattacharyya et al. 1999). The consistency of the results of our F_{ST} and AMOVA analyses is very reassuring. The AMOVA results (Table 5) indicate that the extent of variation is the highest among individuals within population groups. Genetic differentiation with respect to Y-chromosomal markers is generally high, particularly geographical differentiation with respect to Y-RSPs (11.29%), possibly because of the social practices mentioned above. The extent of variation in mt-RSP haplotype frequencies among upper castes of different geographical regions is also high (11.1%). This implies that there is stronger geographical substructuring of upper-caste populations, compared with populations of other ranks. Sociocultural effects on genomic substructuring seem minimal as the proportions of variance attributable to caste–tribal group differences or linguistic differences are quite low. However, there is high genetic differentiation among populations belonging to both caste and tribal groupings, particularly the tribal populations, implying that neither of these two broad groups is genetically homogeneous.

Genetic affinities were analyzed on the basis of data on different sets of markers (Fig. 6). No strong clustering of populations belonging to the same social, geographical, or linguistic group is observed, except that the TB speakers of northeast India form a separate cluster irrespective of the set of markers used. With respect to mtDNA-RSP haplotype frequencies (top of Fig. 6A), a small cluster comprising several populations of the north is also observed. This cluster reflects the high frequencies of haplotypes belonging to HG-U in these populations. Using the 323 distinct HVS1 sequences, a neighbor-joining (NJ) tree was also constructed (data not shown). There was again no clustering of the distinct sequences by geographical region, social rank, or linguistic group. However, most of the sequences belonging to HG-U formed a separate cluster. We have also formally tested whether there is any association between geographic and genetic distances by using the nonparametric Mantel test. No statistically significant association was found, irrespective of the set of marker loci used for computing the genetic distances.

To discover broader patterns of genetic variation, we have pooled populations into social × linguistic subgroups, and have carried out a phylogenetic analysis. The results are presented in Figure 7. The genetic affinities do not reveal any significant additional patterns, except that (1) the Muslims and Tibeto-Burman-speaking tribals are each differentiated from all other categories, irrespective of the genetic system; and (2) with respect to the Y-STRP markers, the Dravidian speakers, irrespective of their position in the social hierarchy, show close affinities (Fig. 7C). The close affinity of the Muslims of Uttar Pradesh with the Indo-European upper-caste groups with respect to Y-chromosomal and autosomal markers (Fig. 7B–D) is easily explained from historical evidence of conversions to Islam in north India (Thapar 1966).

Table 5. Estimates of F_{ST} and AMOVA Results Based on Mitochondrial, Y-Chromosomal, and Autosomal Polymorphisms for Different Groupings of the Populations Studied

						% variation attributable to[a]									
	F_{ST}					Between groups					Between populations within groups				
	mt		Y			mt		Y			mt		Y		
Grouping	RSP	HVS1	RSP	STRP	Auto	RSP	HVS1	RSP	STRP	Auto	RSP	HVS1	RSP	STRP	Auto
2 groups: caste and tribe	0.112	0.102	0.233	0.211	0.053	2.38	0.62	7.87	0.42	0.99	8.81	9.56	14.97	10.04	4.25
6 groups: geographical regions	0.108	0.099	0.241	0.229	0.052	4.32	1.01	11.29	3.16	1.88	6.51	8.99	12.49	8.1	3.32
4 groups: linguistic groups	0.109	0.101	0.243	0.232	0.055	2.64	1.29	8.51	1.86	1.89	8.29	8.90	15.30	8.97	3.46
3 groups: ranked castes—upper, middle and lower	0.056	0.033	0.144	0.119	0.020	0.37	≈0	5.39	≈0	0.11	5.23	3.83	8.38	4.67	1.78
Upper castes of different geographical regions	0.101	0.045	0.089	0.249	0.008	11.10	1.71	4.02	3.67	0.98	≈0	2.78	4.83	1.24	≈0
Middle castes of different geographical regions	0.053	0.047	0.050	0.115	0.022	2.46	0.66	≈0	0.65	≈0	2.88	4.05	10.51	3.88	2.29
Lower castes of four different geographical regions	0.010	0.016	0.192	0.121	0.023	0.84	≈0	≈0	≈0	≈0	0.16	1.83	32.24	9.17	4.77
4 groups: tribes of four different geographical regions (Northeast, East, South, and Central)	0.138	0.158	0.239	0.272	0.069	1.88	1.90	5.25	1.37	2.12	11.94	13.90	18.70	10.03	4.47
4 groups: tribes of four different linguistic groups	0.136	0.157	0.243	0.275	0.069	0.40	0.06	6.86	0.82	1.80	13.15	15.39	17.54	10.52	4.84

[a] The percentages of variation attributable to "between individuals within groups" are not shown. These are obtainable by subtracting from 100 the sum of the percentages of total variation attributable to the other two sources of variation that are shown.

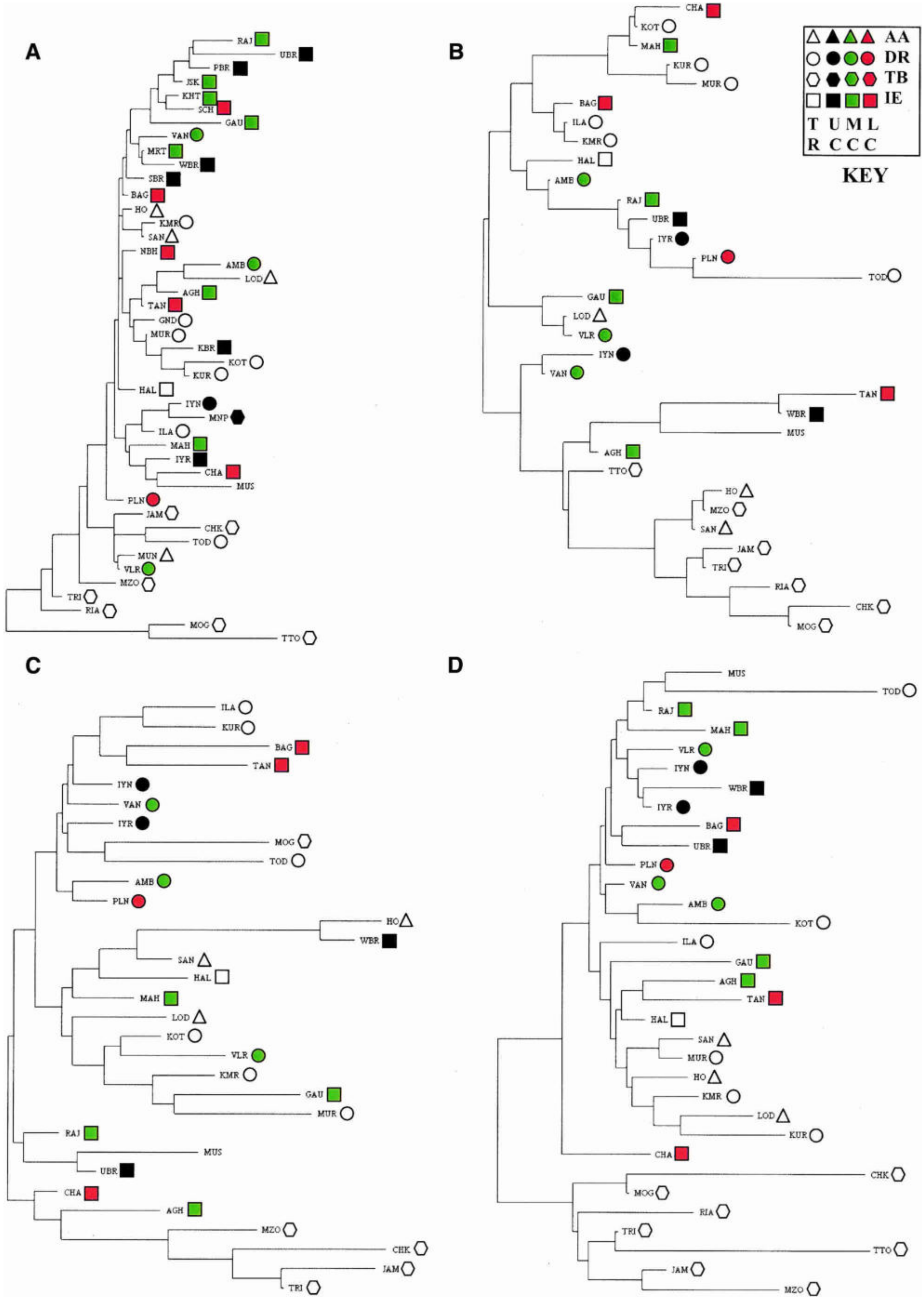


Figure 6 (Legend on next page)

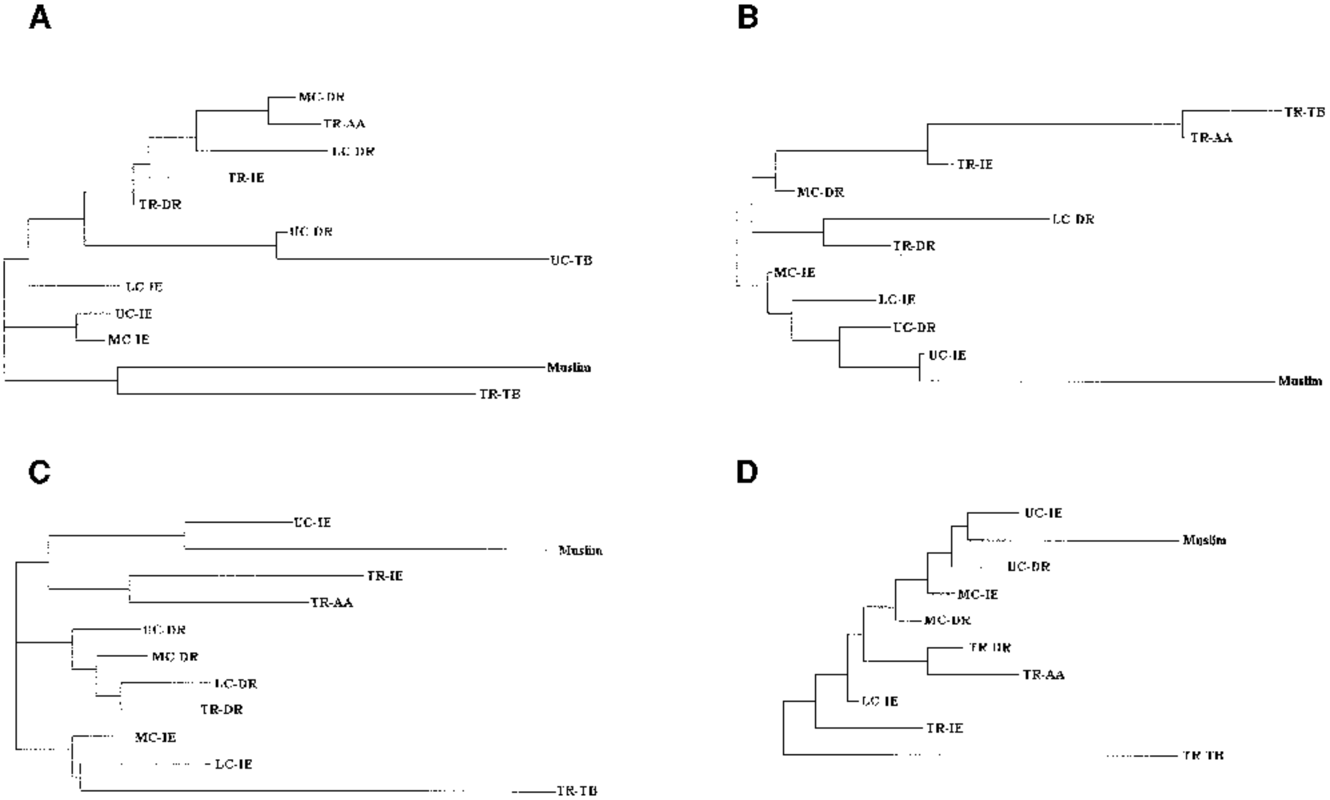


Figure 7 Neighbor-joining tree depicting genetic affinities among categories of populations cross-classified by social rank [(UC) upper caste, (MC) middle caste, (LC) lower caste, (TR) tribal] and linguistic group [(AA) Austro-Asiatic, (DR) Dravidian, (TB) Tibeto-Burman, (IE) Indo-European] based on (*A*) mitochondrial RSP haplotype frequencies, (*B*) Y-haplogroup frequencies, (*C*) Y-STRP frequencies, and (*D*) autosomal markers.

Dravidian Speakers, Now Confined to Southern India, May Have Earlier Been Widespread Throughout India: Genetic Signatures of "Elite Dominance"?

The IE and DR speakers share a significantly larger number of HVS1 sequences (Fig. 8A) compared with those between other groups. The number of individuals sharing these sequences is also the largest between IE and DR groups (Fig. 8B). These facts are striking, especially because the geographical regions presently inhabited by them are virtually disjoint. To explore in further detail the presence of cryptic population structure and the relationships among the various subgroups of populations, we have carried out a "population structure" analysis (Pritchard et al. 2000). In this analysis, an unknown number (*K*) of hypothetical ancestral populations is assumed to have contributed to the genetic profiles of contemporary populations. The number of hypothetical ancestral populations and their relative genetic contributions are statistically estimated from allele frequency data of contemporary populations. The results of population structure analysis based on our autosomal data also show (Fig. 9) that the DR and IE speakers are the most similar, in the sense that the proportional contributions of the five estimated hypothetical ancestral populations to these two groups are the most similar. These findings are consistent with the historical view that the DR speakers were possibly widespread throughout India (Thapar 2003). When the ranked caste system was formed after the arrival of the IE speakers ~3500 ybp, many indigenous people of India, who were possibly DR speakers, embraced (or were forced to embrace) the caste system, together with the IE language and admixture. In fact, Renfrew (1992) has suggested that the elite dominance model, which envisages the intrusion of a relatively small but well-organized group that takes over an existing system by the use of force, may be appropriate to explain the distribution of the IE languages in north India and Pakistan. As the IE speakers, who entered India primarily through the northwest corridor, advanced into the Indo-Gangetic plain, indigenous people, especially the DR speakers, may have retreated southward to avoid linguistic dominance, after an initial period of admixture and adoption of the caste system. As evidenced by their strong genetic similarities (data not shown), the IE-speaking Halba tribals were most probably a DR-speaking tribal group, which is consistent with IE dominance over DR tribals.

Central Asian Populations Have Contributed to the Genetic Profiles of Upper Castes, More in the North Than in the South

Central Asia is supposed to have been a major contributor to the Indian gene pool, particularly to the north Indian gene pool, and the migrants had supposedly moved to India through Afghanistan and Pakistan. From mtHVS1 data, we have estimated F_{ST}

Figure 6 Neighbor-joining tree depicting genetic affinities among Indian ethnic populations based on (*A*) mitochondrial RSP haplotype frequencies, (*B*) Y-haplogroup frequencies, (*C*) Y-STRP frequencies, and (*D*) autosomal markers. The social [(UC) upper caste, (MC) middle caste, (LC) lower caste, (TR) tribal] and linguistic [(AA) Austro-Asiatic, (DR) Dravidian, (TB) Tibeto-Burman, (IE) Indo-European] background of each population is color-coded; the key to the color codes is given on the top right-hand corner.

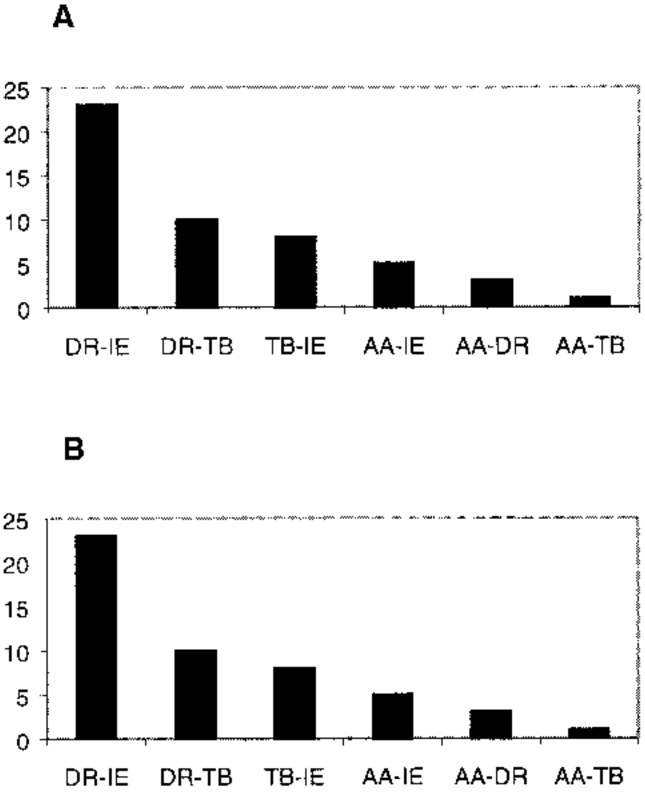


Figure 8 Frequency distributions of mtHVS1 sequences shared between groups of populations—Austro-Asiatic (AA), Dravidian (DR), Tibeto-Burman (TB), and Indo-European (IE). (*A*) Number of sequences shared and (*B*) number of individuals sharing sequences.

values between the populations of the Central Asia and Pakistan regions (data were collated from Calafell et al. 1996; Comas et al. 1998; Kivisild et al. 1999a) and those belonging to various geographical regions of India. Populations of Central Asia and Pakistan show the lowest (0.017) coefficient of genetic differentiation with the north Indian populations, higher (0.042) with the south Indian populations, and the highest (0.047) with the northeast Indian populations. The Central Asian populations are genetically closer to the upper-caste populations than to the middle- or lower-caste populations, which is in agreement with Bamshad et al.'s (2001) findings. Among the upper-caste populations, those of north India are, however, genetically much closer ($F_{ST} = 0.016$) than those of south India ($F_{ST} = 0.031$). Phylogenetic analysis of Y-HG data collated from various sources (Hammer et al. 2000; Nebel et al. 2000; Rosser et al. 2000; Qamar et al. 2002) and with those generated in the present study also showed a similar picture (data not shown). One explanation, consistent with those of the previous section, is that even after the DR speakers retreated to the south to avoid elite dominance, there has been admixture between Central and West Asians and northern Indian populations.

METHODS

Populations

Blood samples were drawn with informed consent from individuals, unrelated to the first cousin level, belonging to 44 populations, chosen to represent ethnic groups of all geographical regions and sociocultural and linguistic categories. A list of populations, with sample sizes and brief notes on their sociocultural backgrounds, is provided in Table 1. We note that (1) population groups of the north are IE speakers, and those of the south are DR speakers; (2) the AA speakers are all tribals and are primarily confined to the central, eastern, and northeastern regions; (3) the TB speakers are confined to the northeastern region; and (4) the number of IE-speaking tribal groups is very few. Thus, there is an extent of confounding of geography, culture, and language in the distribution of ethnic groups of India, which is reflected in the nature of our statistical analyses and inferences. Because of various reasons, including lack of adequate amounts of DNA, not all markers could be examined in all populations.

Loci and Protocols

Ten mtDNA restriction site polymorphisms (*Hae*III np 663, *Hpa*I np 3592, *Alu*I np 5176, *Alu*I np 7025, *Dde*I np 10,394, *Alu*I np 10,397, *Hin*fI np 12,308, *Hin*cII np 13,259, *Alu*I np 13,262, *Hae*III np 16517) and 1 Insertion/Deletion polymorphism (IDP; COII/tRNA^{Lys} intergenic 9-bp deletion) were screened using standard primers and protocols (Torroni et al. 1993, 1996). Sequencing of the mtDNA Hyperveriable Segment-1 (HVS1; np 16,024–16,380) was carried out on an ABI-3100 sequencer, after PCR amplification using standard primers in both directions (Vigilant et al. 1991). HVS1 sequencing was carried out in a subset (528 of 1490) of individuals, randomly selecting at least 10 individuals from each ethnic group.

We screened 22 Y-chromosomal markers; 12 were binary (*YAP*, *92r7*, *SRY 4064*, *sY81*, *SRY + 465*, *TAT*, *M9*, *M13*, *M17*, *M20*, *SRY10831*, and *p12f2*) and 10 were short tandem repeats (*STR*; *DYS19*, *DYS388*, *DYS 389I* and *II*, *DYS390*, *DYS391*, *DYS392*, *DYS393*, *DYS425*, *DYS426*), using primer sequences and protocols described before (Casanova et al. 1985; Thomas et al. 1999; Rosser et al. 2000; Mukherjee et al. 2001). For STRs, an ABI-3100 DNA sequencer and Genescan version 3.1 and Genotyper version 2.1 software packages were used. Haplogroup definitions as given in Y Chromosome Consortium (2002) were followed. However, for the lineages P*(×R1b8,R1a,Q3), BR*(×B2b,CE,F1,H,JK), and K*(×K1,LN,O2b,O3c,P), we have used the abbreviated HG symbols P*, BR*, and K*.

We screened 25 polymorphic autosomal loci. Eight were IDPs, and the remaining 17 were RSPs. The names and GDB (http://gdbwww.gdb.org) accession nos. or ALFRED UID (http://alfred.med.yale.edu) of the RSP loci are: *ESR1* (GDB: 185229); *NAT* (GDB: 187676); *CYP1A* (GDB: 9956062)-*Msp*I; *PSCR* (GDB: 182305); *T2* (GDB: 196856); *LPL* (GDB: 285016); *ALB* (GDB: 178648); *ALAD-Msp*I (GDB: 155925); *ALAD-Rsa*I (GDB: 155924); *HB* ψβ-*Hin*cII (GDB: 56084); *HB* 3′ψβ-*Hin*cII; *HB* 5′β-*Hin*fI; *HoxB4-Msp*I (UID: SI0001670); *DRD2* (UID: SI000191L) -*Taq*IB, -*Taq*ID, -*Taq*1A; and *ADH2-Rsa*I (UID: SI000002C). The names of the IDPs and primers and protocols used are as given in Majumder et al. (1999a) and Tishkoff et al. (1996). RSPs were screened using primers and protocols of Jorde et al. (1995), Majumder et al. (1999b), and K. Kidd (pers. comm.).

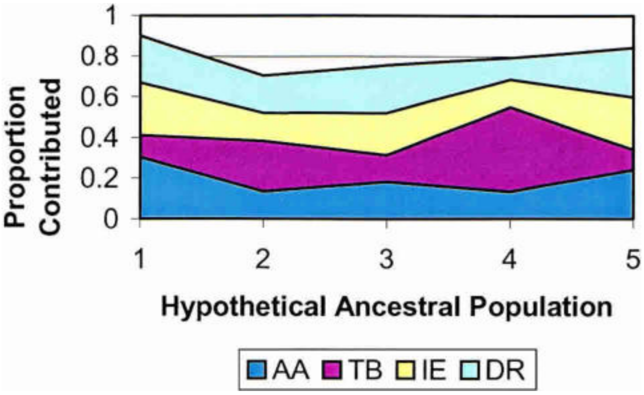


Figure 9 Results of STRUCTURE analysis: Proportional contributions of five hypothetical ancestral populations to Austro-Asiatic (AA), Dravidian (DR), Tibeto-Burman (TB), and Indo-European (IE) groups.

Statistical Methods

Maximum likelihood estimates of haplotype frequencies at linked autosomal loci were obtained via the EM algorithm using the HAPLOFREQ package (Majumdar and Majumder 2000). DNA sequences were aligned using CLUSTALW (http://www2.ebi.ac.uk/clustalw/). The Cambridge sequence was used as reference during alignment.

Tests of significance in sparse cross-classified tables were carried out by a bootstrap test procedure devised by us. Standard statistical analyses were carried out using SPSS. Population genetic analyses were performed using ARLEQUIN (Schneider et al. 2000; http://lgb.unige.ch/arlequin/) and DnaSP (Rozas and Rozas 1999; http://www.ub.es/dnasp/). Mantel tests were carried out using MANTEL, version 2.0 (http://www.sci.qut.edu.au/NRS/mantel.htm). Expansion times using HVS1 data were estimated (Slatkin and Hudson 1991) assuming a mutation rate of 20.5% per site per million years (Bonatto and Salzano 1997). For phylogenetic analyses using HVS1 data, DNA distances were calculated using NETWORK (http://www.fluxus-engineering.com/sharenet.htm) and the maximum likelihood method as implemented in PHYLIP version 3.5 (http://evolution.genetics.washington.edu/phylip.html) assuming a 30:1 transition:transversion ratio (Lundstrom et al. 1992). For phylogenetic reconstruction from data on autosomal markers, mtDNA-RSPs, and Y-chromosomal markers, Nei's (1987) D_A distance measure and the neighbor-joining method (Saitou and Nei 1987) were used, as implemented in DISPAN (http://oat.bio.indiana.edu:7580/) and PHYLIP version 3.5c packages. The age (A) of a Y-HG was estimated as $A = g \times s^2/\mu$, where g is the generation time (assumed to be 30 yr); s^2 is the variance of the STR repeat number among haplotypes belonging to the HG, averaged over all STR loci; and μ is the mutation rate per generation at an STR locus, taken to be 0.18% (Quintana-Murci et al. 2001). A Markov Chain Monte Carlo analysis of population structure (Pritchard et al. 2000) was carried out using the program STRUCTURE (http://pritch.bsd.uchicago.edu/software.html).

ACKNOWLEDGMENTS

We thank our collaborators C.S. Chakraborty, R. Lalthantluanga, Mitashree Mitra, A. Ramesh, N.K. Sengupta, Samir K. Sil, Jai Rup Singh, Chitra Mahadik Thakur, and M.V. Usha Rani for help in collecting samples and at various other stages of this work. We thank Sujit Maiti for help with data management and in preparing the figures. We also thank Lynn Jorde, Ken Kidd, Andy Merriwether, Antonio Torroni, and Chris Tyler-Smith for providing advice and for sharing laboratory protocols with us during the initial stages of this work. Hans Bandelt pointed out some errors in our sequence data; we are grateful to him. This work was partially supported by grants from the Department of Biotechnology, Government of India.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Bamshad, M.J., Watkins, W.S., Dixon, M.E., Jorde, L.B., Rao, B.B., Naidu, J.M., Prasad, B.V.R., Rasanayagam, A., and Hammer, M.F. 1998. Female gene flow stratifies Hindu castes. *Nature* **395:** 851–852.

Bamshad, M.J., Kivisild, T., Watkins, W.S., Dixon, M.P., Ricker, L.E., Rao, B.B., Naidu, M., Prasad, B.V.R., Reddy, P.G., Rasanayagam, A., et al. 2001. Genetic evidence on the origins of Indian caste populations. *Genome Res.* **11:** 994–1004.

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

Beteille, A. 1998. The Indian heritage—A sociological perspective. In *The Indian human heritage* (eds. D. Balasubramian and N. Appaji Rao), pp. 87–94. University Press, Hyderabad, India.

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

Bonatto, S.L. and Salzano, F.M. 1997. A single and early origin for the peopling of the Americas supported by mitochondrial DNA sequence data. *Proc. Natl. Acad. Sci.* **94:** 1866–1871.

Calafell, F., Underhill, P., Tolun, A., Angelicheva, D., and Kalaydjieva, L. 1996. From Asia to Europe: Mitochondrial DNA sequence variability in Bulgarians and Turks. *Ann. Hum. Genet.* **60:** 35–49.

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

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.

Comas, D., Calafell, F., Mateu, E., Perez-Lezaun, A., Bosch, E., Martinez-Arias, R., Clarimon, J., Facchini, F., Fiori, G., Luiselli, D., et al. 1998. Trading genes along the silk road: mtDNA sequences and the origin of Central Asian populations. *Am. J. Hum. Genet.* **63:** 1824–1838.

Cruciani, F., Santolamazza, P., Shen, P., Macaulay, V., Moral, P., Olckers, A., Modiano, D., Holmes, S., Destro-Bisol, G., Coia, V., et al. 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.

Diamond, J. 1997. *Guns, germs and steel: The fates of human societies*. Jonathan Cape, London.

Forster, P., Torroni, A., Renfrew, C., and Rohl, A. 2001. Phylogenetic star contraction applied to Asian and Papuan mtDNA evolution. *Mol. Biol. Evol.* **18:** 1864–1881.

Grimes, B.F. 1999. *The ethnologue: Languages of the world.* Summer Institute of Linguistics, California.

Guha, B.S. 1935. The racial affinities of the people of India. In *Census of India, 1931*, Part III—Ethnographical. Government of India Press, Simla, India.

Hammer, M.F., Redd, A.J., Wood, E.T., Bonner, M.R., Jarjanazi, H., Karafet, T., Santachiara-Benerecetti, S., Oppenheim, A., Jobling, M.A., Jenkins, T., et al. 2000. Jewish and middle eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. *Proc. Natl. Acad. Sci.* **97:** 6769–6774.

Jorde, L.B., Bamshad, M.J., Watkins, W.S., Zenger, R., Fraley, A.E., Krakowiak, P.A., Carpenter, K.D., Soodyall, H., Jenkins, T., and Rogers, A.R. 1995. Origins and affinities of modern humans: A comparison of mitochondrial and nuclear genetic data. *Am. J. Hum. Genet.* **57:** 523–538.

Karve, I. 1961. *Hindu society: An interepretation.* Deshmukh Prakashan, Poona, India.

Keith, A. 1936. Review of B.S. Guha's "Racial affinities of the peoples of India." *Man* **29:** 37.

Kivisild, T., Bamshad, M.J., Kaldma, K., Metspalu, M., Metspalu, E., Reidla, M., Laos, S., Parik, J., Watkins, W.S., Dixon, M.E., et al. 1999a. Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. *Curr. Biol.* **9:** 1331–1334.

Kivisild, T., Kaldma, K., Metspalu, M., Parik, J., Papiha, S., and Villems, R. 1999b. The place of the Indian mitochondrial DNA variants in the global network of the maternal lineages and the peopling of the old world. In *Genome diversity: Applications in human population genetics* (eds. S.S. Papiha et al.), pp. 135–152. Kluwer, New York.

Kosambi, D.D. 1991. *The culture and civilisation of ancient India in historical outline.* Vikas Publishing House, New Delhi, India.

Lundstrom, R.S., Tavare, S., and Ward, R.H. 1992. Estimating substitution rates from molecular data using the coalescent. *Proc. Natl. Acad. Sci.* **89:** 5961–5965.

Maca-Meyer, N., Gonzalez, A.M., Larruga, J.M., Flores, C., and Cabrera, V.M. 2001. Major genomic mitochondrial lineages delineate early human expansions. *BMC Genetics* **2:** 13–20.

Majumdar, P. and Majumder, P.P. 2000. *HAPLOFREQ: A computer program package for maximum-likelihood estimation of haplotype frequencies in a population via the EM algorithm.* Tech Rep AHGU-1/2000. Indian Statistical Institute, Calcutta, India.

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

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

Majumder, P.P., Roy, B., Balgir, R.S., and Dash, B.P. 1999b. Polymorphisms in the β-globin gene cluster in some ethnic populations of India and their implications on disease. In *Molecular intervention in disease* (eds. S. Gupta and O.P. Sood), pp. 75–83. Ranbaxy Science Foundation, New Delhi.

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

Misra, V.N. 1992. Stone age in India: An ecological perspective. *Man and Env.* **14:** 17–64.

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.

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. 1987. *Molecular evolutionary genetics*. Columbia University Press, New York.

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

Pritchard, J.K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155:** 945–959.

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

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

Ratnagar, S. 1995. Archaeological perspectives of early Indian societies. In *Recent perspectives of early Indian history* (ed. R. Thapar), pp. 1–52. Popular Prakashan, Bombay, India.

Ray, N. 1973. *Nationalism in India.* Aligarh Muslim University, Aligarh, India.

Renfrew, C. 1992. Archaeology, genetics and linguistic diversity. *Man* **27:** 445–478.

Risley, H.H. 1915. *The people of India.* Thacker Spink, Calcutta, India.

Rosser, Z.H., Zerjal, T., Hurles, M.E., Adojaan, M., Alavantic, D., Amorim, A., Amos, W., Armenteros, M., Arroyo, E., Barbujani, G., 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.

Roychoudhury, S., Roy, S., Basu, A., Banerjee, R., Vishwanathan, H., Usha Rani, M.V., Sil, S.K., Mitra, M., and Majumder, P.P. 2001. Genomic structures and population histories of linguistically distinct tribal groups of India. *Hum. Genet.* **109:** 339–350.

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

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

Sarkar, S.S. 1958. Race and race movements in India. In *The cultural heritage of India* (ed. S.K. Chatterjee), Vol. 1, pp. 17–32. The Ramakrishna Mission Institute of Culture, Calcutta, India.

Schneider, S., Kueffer, J.-M., Roessli, D., and Excoffier, L. 2000. *ARLEQUIN: A software for population genetic data analysis.* University of Geneva, Geneva, Switzerland.

Singh, K.S. 1992. *People of India: An introduction.* Anthropological Survey of India, Calcutta, India.

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

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

Thapar, R. 1966. *A history of India,* Volume 1. Penguin, Middlesex, UK.

———. 1995. The first millennium B.C. in northern India (up to the end of Mauryan period). In *Recent perspectives of early Indian history* (ed. R. Thapar), pp. 80–141. Popular Prakashan, Bombay, India.

———. 2003. *Early India: From the origins to AD 1300.* University of California Press, Berkeley, CA.

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.

Tishkoff, S.A., Ruano, G., Kidd, J.R., and Kidd, K.K. 1996. Distribution and frequency of a polymorphic *Alu* insertion at the plasminogen activator locus in humans. *Hum. Genet.* **97:** 759–764.

Torroni, A., Schurr, T.B., Cabell, M.F., Brown, M.D., Neel, J.V., Larsen, M., and Smith, D.G. 1993. Asian affinities and continental radiation of the four founding Native American mtDNAs. *Am. J. Hum. Genet.* **53:** 563–590.

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

Vigilant, L.A., Wilson, A.C., and Harpending, H. 1991. African populations and the evolution of human mitochondrial DNA. *Science* **253:** 1503–1507.

Y Chromosome Consortium. 2002. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res.* **12:** 339–348.

Zerjal, T., Wells, R.S., Yuldasheva, N., Ruzibakiev, R., and Tyler-Smith, C. 2002. A genetic landscape reshaped by recent events: Y-chromosomal insights into Central Asia. *Am. J. Hum. Genet.* **71:** 466–482.

WEB SITE REFERENCES

http://alfred.med.yale.edu; ALFRED.
http://evolution.genetics.washington.edu/phylip.html; PHYLIP.
http://gdbwww.gdb.org; GDB.
http://lgb.unige.ch/arlequin/; ARLEQUIN.
http://oat.bio.indiana.edu:7580/; DISPAN.
http://pritch.bsd.uchicago.edu/software.html; STRUCTURE.
http://www.fluxus-engineering.com/sharenet.htm; NETWORK.
http://www2.ebi.ac.uk/clustalw/; CLUSTALW.
http://www.sci.qut.edu.au/NRS/mantel.htm; MANTEL.
http://www.ub.es/dnasp/; DnaSP.

Received April 16, 2003; accepted in revised form August 5, 2003.

Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure

Analabha Basu[a,1], Neeta Sarkar-Roy[a], and Partha P. Majumder[a,b,1]

[a]National Institute of BioMedical Genomics, NetajiSubhas Sanatorium (Tuberculosis Hospital), Kalyani 741251, West Bengal, India; and [b]Human Genetics Unit, Indian Statistical Institute, Kolkata 700108, West Bengal, India

Edited by Masatoshi Nei, Pennsylvania State University, University Park, PA, and approved December 17, 2015 (received for review July 5, 2015)

India, occupying the center stage of Paleolithic and Neolithic migrations, has been underrepresented in genome-wide studies of variation. Systematic analysis of genome-wide data, using multiple robust statistical methods, on (*i*) 367 unrelated individuals drawn from 18 mainland and 2 island (Andaman and Nicobar Islands) populations selected to represent geographic, linguistic, and ethnic diversities, and (*ii*) individuals from populations represented in the Human Genome Diversity Panel (HGDP), reveal four major ancestries in mainland India. This contrasts with an earlier inference of two ancestries based on limited population sampling. A distinct ancestry of the populations of Andaman archipelago was identified and found to be coancestral to Oceanic populations. Analysis of ancestral haplotype blocks revealed that extant mainland populations (*i*) admixed widely irrespective of ancestry, although admixtures between populations was not always symmetric, and (*ii*) this practice was rapidly replaced by endogamy about 70 generations ago, among upper castes and Indo-European speakers predominantly. This estimated time coincides with the historical period of formulation and adoption of sociocultural norms restricting intermarriage in large social strata. A similar replacement observed among tribal populations was temporally less uniform.

ancestry | admixture | haplotype blocks | endogamy | social stratification

India has served as a major corridor for both Paleolithic and Neolithic migrations of anatomically modern humans (1). An early dispersal of modern humans from Africa into India through the southern coastal route (2–4) and migration from West and Central Asia through the northwest corridor (5–8) inferred by past genetic studies have been supported by archaeological evidence, admittedly scattered (2). This evidence fits with Reich et al.'s (9) proposed model that most extant populations of India are a result of admixture between two ancestral populations—Ancestral North Indian (ANI) and Ancestral South Indian (ASI) (9, 10). Anthropologists believe that some of Negrito hunter-gatherer tribes of the Andaman and Nicobar archipelago (A&N) in the Indian Ocean (such as the Jarawa and Onge included in this study) may hold the key to understand the peopling of eastern and southern Asia after anatomically modern humans came out to Africa. Reich et al. (9) also found a distinct component of ancestry among the tribals of A&N, and noted that these tribals are "unique in being ASI-related groups without ANI ancestry" (9). The process by which this archipelago was peopled is unknown but possibly holds the key to our understanding of peopling of South Asia, Pacific Islands, and Australia. Furthermore, multiple lines of evidence, including popularity of rice cultivation in East and Northeast India (11, 12), abundance of the Tibeto-Burman (TB) and Austro-Asiatic (AA) language speakers (13, 14), findings from past archeological and anthropometric (15) as well as genetic studies (6, 16), indicate major waves of migration through India's northeast corridor.

Reich et al.'s (9) model that all populations of mainland India arose from admixture between two ancestral populations relied strongly on the finding of a north-to-south clinal arrangement of individuals drawn from various populations on a plot of the first two principal components (PCs). A decreasing proportion of "Middle Easterners, Central Asians, and Europeans-like" ancestry from north to south was noted (9). However, TB- and AA-speaking individuals, who were "off-cline" in the PC plot and excluded from further analysis (9, 10), represent additional ancestral components in the Indian population. By analyzing more representative population samples using robust statistical methods, here we provide a fine-grained reconstruction of India's population history.

Contemporary populations of India are linguistically, geographically, and socially stratified (6, 16), and are largely endogamous with variable degrees of porosity. We analyzed high-quality genotype data, generated using a DNA microarray (*Methods*) at 803,570 autosomal SNPs on 367 individuals drawn from 20 ethnic populations of India (Table 1 and *SI Appendix*, Fig. S1), to provide evidence that the ancestry of the hunter-gatherers of A&N is distinct from mainland Indian populations, but is coancestral to contemporary Pacific Islanders (PI). Our analysis reveals that the genomic structure of mainland Indian populations is best explained by contributions from four ancestral components. In addition to the ANI and ASI, we identified two ancestral components in mainland India that are major for the AA-speaking tribals and the TB speakers, which we respectively denote as AAA (for "Ancestral Austro-Asiatic") and ATB (for "Ancestral Tibeto-Burman"). Extant populations have experienced extensive multicomponent admixtures. Our results indicate that the census sizes of AA and TB speakers in contemporary India are gross underestimates of the extent of the AAA and the ATB components in extant populations. We have

Significance

India, harboring more than one-sixth of the world population, has been underrepresented in genome-wide studies of variation. Our analysis reveals that there are four dominant ancestries in mainland populations of India, contrary to two ancestries inferred earlier. We also show that (*i*) there is a distinctive ancestry of the Andaman and Nicobar Islands populations that is likely ancestral also to Oceanic populations, and (*ii*) the extant mainland populations admixed widely irrespective of ancestry, which was rapidly replaced by endogamy, particularly among Indo-European–speaking upper castes, about 70 generations ago. This coincides with the historical period of formulation and adoption of some relevant sociocultural norms.

Author contributions: A.B. and P.P.M. designed research; A.B. and N.S.-R. performed research; A.B. analyzed data; and A.B. and P.P.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. Email: ppm1@nibmg.ac.in or ab1@nibmg.ac.in.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1513197113/-/DCSupplemental.

Table 1. Sociocultural and linguistic characteristics of 20 population groups sampled from different geographical locations of India, with sample sizes

Population name	Social hierarchy	Geography	Linguistic group	Primary occupation*	Sample size
Khatri (KSH)	Upper caste	North	Indo-European	Traditionally warrior*	19
Gujarati Brahmin (GBR)	Upper caste	Northwest	Indo-European	Traditionally priest*	20
West Bengal Brahmin (WBR)	Upper caste	East	Indo-European	Traditionally priest*	18
Maratha (MRT)	Upper caste	West	Indo-European	Traditionally warriors*	7
Iyer (IYR)	Upper caste	South	Dravidian	Traditionally priest*	20
Pallan (PLN)	Lower-middle caste	South	Dravidian	Agriculturist*	20
Kadar (KDR)	Tribe	South	Dravidian	Hunter-gatherer	20
Irula (IRL)	Tribe	South	Dravidian	Hunter-gatherer	20
Paniya (PNY)	Tribe	South	Dravidian	Hunter-gatherer	18
Gond (GND)	Tribe	Central	Dravidian/Austro-Asiatic	Agriculturist Hunter-gatherer	20
Ho (HO)	Tribe	Central and East	Austro-Asiatic	Agriculturist Hunter-gatherer	18
Santal (SAN)	Tribe	Central and East	Austro-Asiatic	Agriculturist Hunter-gatherer	20
Korwa (KOR)	Tribe	Central	Austro-Asiatic	Hunter-gatherer	18
Birhor (BIR)	Tribe	Central	Austro-Asiatic	Hunter-gatherer	16
Manipuri Brahmin (MPB)	Upper caste	Northeast	Tibeto-Burman	Traditionally warrior*	20
Tharu (THR)	Tribe	North	Indo-European	Agriculturist	20
Tripuri (TRI)	Tribe	Northeast	Tibeto-Burman	Agriculturist	19
Jamatia (JAM)	Tribe	Northeast	Tibeto-Burman	Agriculturist	18
Jarawa (JRW)	Tribe	Andaman and Nicobar	Ongan	Hunter-gatherer	19
Onge (ONG)	Tribe	Andaman and Nicobar	Ongan	Hunter-gatherer	17

*With the formation of the caste system, which is a system of social stratification, endogamous caste groups were traditionally attributed occupations that were to be hereditary. All of the caste groups in contemporary India are large populations and are engaged in a variety of occupations. The "Primary occupation" column describes the traditional occupation.

inferred that the practice of endogamy was established almost simultaneously, possibly by decree of the rulers, in upper-caste populations of all geographical regions, about 70 generations before present, probably during the reign (319–550 CE) of the ardent Hindu Gupta rulers. The time of establishment of endogamy among tribal populations was less uniform.

Islanders and Mainlanders: Exclusive Ancestries

We determined the axes of human genomic variation using principal-components analysis (PCA), as implemented in EIGENSTRAT (17). Using a dynamic programming-driven unsupervised clustering algorithm, ADMIXTURE (18), we determined the genomic admixture at the individual level, by partitioning the genome of an individual into K components contributed by hypothetical ancestors and then estimating their relative contributions. The first principal component (PC-1) explained a high fraction (over 13%) of genomic variation and differentiated the populations of A&N Islands—JRW and ONG—from the mainland populations (Fig. 1), indicating long separation and negligible gene flow. This inference was strongly supported by ADMIXTURE analysis considering two ancestral populations ($K = 2$) that were found to have contributed disjointedly to the gene pools of the islanders and mainlanders (Fig. 1 and *SI Appendix*, section 1).

Mainlanders: Four Ancestral Components

The analysis of genome-wide SNP data on 331 individuals from 18 mainland populations (excluding the 19 ONG and 17 JRW individuals), revealed four ancestral components that formed distinct clusters and clines, in contrast to two components inferred earlier (9) (Fig. 2*A*). The TB speakers formed a distinct clinal cluster along PC-1, representing descendants of ATB. Along PC-2, the dominant cline was the north-to-south ANI–ASI (9) cline. The AA speakers were also distributed along PC-2 but formed a separate cline indicative of a large contribution from a separate ancestral source (AAA). Central Indian tribal populations, such as Gond and Ho, occupying "central" positions in the PC-1 vs. PC-2 plot, have been noted to be extremely heterogeneous (19) and reported to be quite admixed. These features were also recapitulated by ADMIXTURE analysis (Fig. 2*B* and Table 2). Multiple runs of ADMIXTURE established the model with four ancestral components ($K = 4$) as the best-fitting model (*SI Appendix*, section 2). Model validation by optimum choice of the number of ancestral components (K) was achieved for each dataset by minimizing the cross-validation error (CVE) (18) considering different cutoff values for linkage disequilibrium (LD) and the proportion of data masked for CVE estimation (*SI Appendix*, section 2, Figs. S2 and S3 *A*–*F*). Detailed results for multiple runs of ADMIXTURE (provided in *SI Appendix*, section 2) show that the convergence is robust.

The proportions of inferred ancestral components for each population estimated by ADMIXTURE (Table 2) were compared with maximum-likelihood estimates obtained using *frappe* (20); both sets of estimates were nearly identical (Table 2 compared with *SI Appendix*, section 2, Table S4). This concordant finding was further investigated using fineSTRUCTURE (21), which is robust to existing LD and is capable of identifying subtle population subdivisions. fineSTRUCTURE identified 69 subpopulations (*SI Appendix*, Fig. S4 *A* and *B* and section 2) from the data on 331 individuals drawn from 18 ethnic groups. These subpopulations were largely nonoverlapping and belonged to four major clades whose compositions were nearly identical to the four ancestral components identified by ADMIXTURE analysis (*SI Appendix*, Fig. S4*A*: depicting the close concordance between the coancestry matrix estimated by fineSTRUCTURE with proportions of ancestral components derived from ADMIXTURE analysis). Sixty (87%) of the 69 subpopulations identified by fineSTRUCTURE comprised individuals drawn from 1 of the 18 original ethnic groups. Viewed differently, only nine subpopulations contained individuals drawn from more than a single ethnic group; even in these rare instances, the

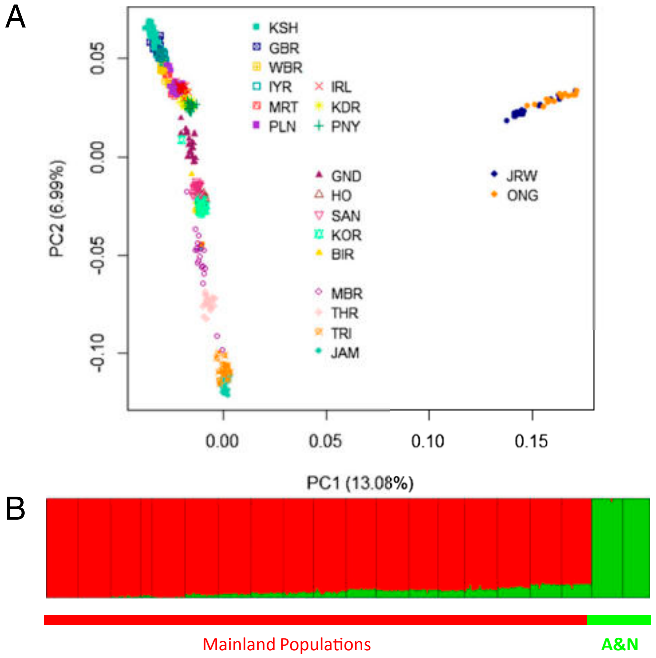


Fig. 1. (*A*) Scatterplot of the 367 individuals sampled from 20 Indian populations by the first two PCs extracted from genome-wide genotype data. The Andamanese populations (JRW and ONG) cluster together and are widely separated from mainland populations. (*B*) Ancestries of individuals estimated using ADMIXTURE with two ancestral components. The 367 individuals are clustered into two distinct groups: the mainlanders (red) and Andamanese islanders (green). (Ancestries of individuals estimated using ADMIXTURE for $K = 2$, 3, and 4 and related results are in *SI Appendix*.)

individuals were always from closely related ethnic groups (for example, two AA-speaking tribal populations residing in the geographical region) (*SI Appendix*, section 2). Thus, the numerically larger ethnic groups are well differentiated, even though genomic subdivisions are discernible within them.

Compared with autosomes, the X chromosome has a smaller effective population size (hence more strongly affected by random genetic drift), lower mutation and recombination rates, and greater selective pressure in males. However, the X chromosome is the most informative source to evaluate sex bias in admixture. Sex bias in ancestry contribution was explored using the 107 females, identified unambiguously from genotype data (*Methods*), belonging to 15 mainland populations. ADMIXTURE analysis using data on these 107 females, with $K = 4$, separately for the X chromosome and autosomes (*SI Appendix*, Fig. S5*A*), reveals a sex bias in all ancestries, except AAA. Although the ATB shows clear excess of X-chromosomal component compared with autosomes, a reverse trend was observed for the ANI component (*SI Appendix*, Fig. S5*B*). We also combined the X-chromosome haplotypes of males with the inferred haplotypes (*SI Appendix*, section 6, *Methods*) of females and used them to construct a phylogenetic tree (*SI Appendix*, Fig. S6 *A* and *B*). The phylogenetic tree shows distinct clustering (*SI Appendix*, Fig. S6*B*) of the haplotypes in clades that belong to genetically closely related populations as inferred from the autosomal data.

More Robust Identification of the Ancestral Components

To more robustly identify and characterize the ancestral components, we combined our data on mainland populations of India with Europe (Eur), Middle Easterners (ME), Central-South Asians (CS-Asian), East Asians (E-Asian) included in Human Genome Diversity Panel (HGDP) (22, 23). The resultant dataset comprised a common set of 630,918 markers. Reich et al. (9) have characterized the ANI ancestry as "genetically close to Middle Easterners, Central Asians, and Europeans." Similar to Li et al. (22), our PCA plot shows the Eur and ME cluster distinctly, despite being genetically close to the CS-Asians and populations that have high proportion of ANI ancestry (*SI Appendix*, Fig. S7).

In Fig. 3, PC-1 represents the systematic variation broadly separating the CS-Asian ancestry from E-Asian ancestry, whereas PC-2 represents the systematic variation broadly between the combined AAA plus ASI ancestry and others. The separation of the CS-Asians and E-Asians broadly recapitulated the findings of Li et al. (22). The populations of India with a

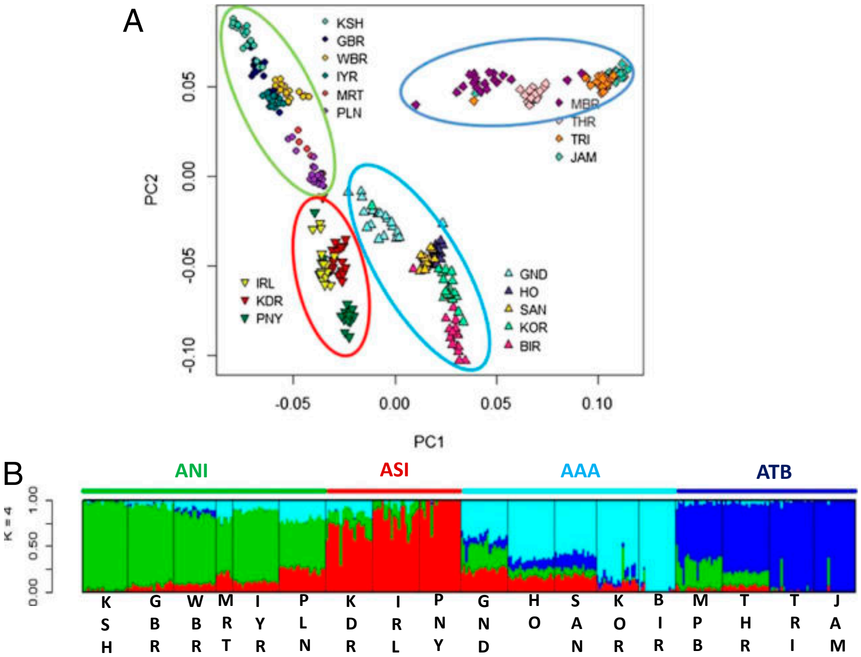


Fig. 2. (*A*) Scatterplot of 331 individuals from 18 mainland Indian populations by the first two PCs extracted from genome-wide genotype data. Four distinct clines and clusters were noted; these are encircled using four colors. (*B*) Estimates of ancestral components of 331 individuals from 18 mainland Indian populations. A model with four ancestral components ($K = 4$) was the most parsimonious to explain the variation and similarities of the genome-wide genotype data on the 331 individuals. Each individual is represented by a vertical line partitioned into colored segments whose lengths are proportional to the contributions of the ancestral components to the genome of the individual. Population labels were added only after each individual's ancestry had been estimated. We have used green and red to represent ANI and ASI ancestries; and cyan and blue with the inferred AAA and ATB ancestries. These colors correspond to the colors used to encircle clusters of individuals in *A*. (Also see *SI Appendix*, Figs. S2 and S3.)

Table 2. Estimates of ancestry proportions of 18 mainland Indian populations under the best-fitting ADMIXTURE model with (K = 4) four ancestral components

Population	ANI	ASI	AAA	ATB
KSH	**0.9793**	**0.0149**	**0.0045**	**0.0013**
GBR	0.8823	0.0759	0.0412	0.0006
WBR	0.7663	0.0994	0.101	0.0332
MRT	0.5751	0.2141	0.2105	0.0003
IYR	0.8046	0.111	0.0837	0.0007
PLN	0.4902	0.2761	0.2331	0.0006
KAD	0.0895	0.7681	0.1414	0.0011
IRL	0.0532	0.9255	0.0213	0
PNY	**0.0252**	**0.9696**	**0.0052**	**0**
GND	0.3697	0.193	0.3756	0.0617
HO	0.0475	0.1705	0.7116	0.0704
SAN	0.0347	0.1933	0.6398	0.1321
KOR	0.0181	0.0471	0.9091	0.0257
BIR	**0.0082**	**0.0054**	**0.9864**	**0**
MPB	0.2635	0.0512	0.0351	0.6502
THR	0.0935	0.0951	0.0447	0.7667
TRI	0.0156	0.0084	0.0117	0.9643
JAM	**0.0149**	**0.0044**	**0.0031**	**0.9776**

Names of the four populations in bold are identified with the four distinct ancestries.

large proportion of ANI component; particularly the KSH with $\sim$97% ANI ancestry is inseparable from the CS-Asian, particularly Burusho, Pathan, and Sindhi. The hypothesis that the root of ANI is in Central Asia is further bolstered by the recent evidence derived from analysis of ancient DNA samples (24) and linguistic studies (25). Similarly, the JAM and TRI who have more than 95% ATB ancestry are inseparable from E-Asian populations, e.g., Dai, Lahu, and Cambodian, who live in or near southwestern China and have the lowest "northern" Chinese ancestry (22). Fig. 3 reveals concordance of geographical residence and genetic axes of variation between populations (*SI Appendix*, section 3).

The Indian dataset, including the JRW and ONG data (A&N), when combined with the HGDP populations of CS-Asia, E-Asia, and Oceania, reveal discernable components of genetic variation that distinguish the CS-Asians from E-Asians, and the Oceanic from other populations (*SI Appendix*, Fig. S8*A*). The A&N populations also appear to share a common ancestry with the Oceanic PIs, particularly the Papuans (*SI Appendix*, Fig. S8*A*). Owing probably to geographical separation and random genetic drift due to isolation of the island populations, they also separate along the third PC (*SI Appendix*, Fig. S8 *B* and *C*).

Admixture to Endogamy

The extent of borrowed Dravidian and AA linguistic elements (26, 27) in the Rigveda, the earliest of the Vedic texts (dated between 1500 and 1000 BCE), has prompted historians and linguists to argue in favor of a "fair degree" of mixing of the populations (15, 25, 27). Earlier genetic studies have also argued that India was a "relatively" pan-mixing society that embraced endogamy between 1,900 and 4,200 y (9, 10). We reinvestigated the extent of ancient admixture, using a model where individuals could derive their ancestries, at varying degrees, from four genetically distinct components (ANI, ASI, AAA, ATB), instead of three (ANI, ASI, AAA) as the linguists have proposed (26, 27) or two (ANI, ASI) as inferred from previous genetic studies (9, 10).

At homologous genomic regions, distinct ancestral populations are expected to possess distinctive DNA sequences. In other words, different ancestral populations possess a large number of distinguishable haplotype blocks. Meiotic recombination results in exchange of homologous segments between the chromosomes of individuals. Therefore, for an individual with multiple ancestral contributions, distinctive haplotype blocks corresponding to the ancestral populations get fragmented with each event of recombination. When a recipient population (P2) receives, in each generation, a small proportion of haplotypes from a donor ancestral population (P1), the haplotypes of P2 will contain a mixture of fragmented haplotypes and intact haplotypes from P1. If the influx of genetic material from P1 to P2 suddenly ceases, in each subsequent generation, intact haplotypes of P1 in P2 will get fragmented due to recombination. Recombination events, on an average, occur at a rate of one per morgan per generation, and can be appropriately modeled as a Poisson process. Therefore, in the recipient population P2, the distribution of the lengths of haplotype (chromosomal) segments of the donor population P1 will follow an exponential distribution with mean $1/(1-\alpha)T$ (28, 29), where α (small) is the proportion of admixture per generation of genes from P1 to P2 and T is the number of generations before present (GBP) when this admixture stopped. It is to be noted here that α,

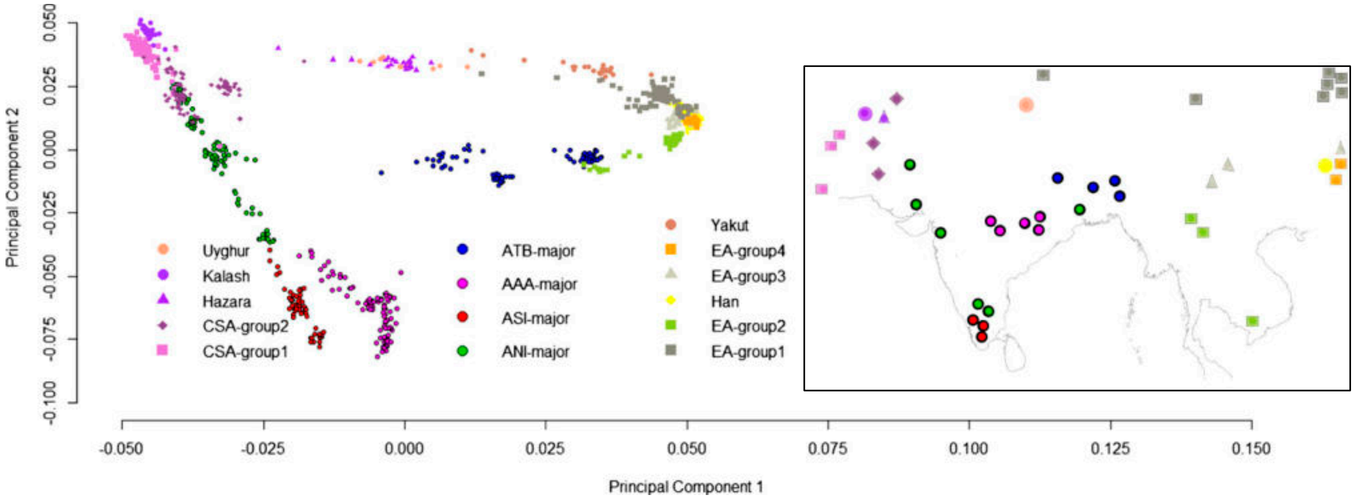


Fig. 3. Approximate "mirroring" of genes and geography. Genomic variation of individuals, represented by the first two PCs, sampled from 18 mainland Indians combined with the CS-Asians) and E-Asians from HGDP, compared with the map of the Indian subcontinent showing the approximate locations from which the individuals and populations were sampled.

if large, that is, if the major portions of the haplotypes are from a particular ancestry, will imply that even if haplotypes break down by recombination into smaller blocks these will not be identifiable because of their similarities with background haplotypes (NA in Table 3). Thus, the time and extent of admixture can be estimated from the distribution of the length of haplotype tracts identified with distinct ancestries in admixed genomes.

We inferred local ancestries and reconstructed each individual's genome as a potential mosaic of the four components. Individual haplotypes were inferred using Shapeit2 (30, 31) and ancestry of each block was identified using PCAdmix (32) (*Methods*). Owing to their near nonadmixed status, KSH (98% ANI), PNY (97% ASI), BIR (99% AAA), and JAM (98% ATB) were chosen as best representatives of the ANI, ASI, AAA, and ATB populations.

In each population, the distribution of the ancestral block lengths (ABLs) thus identified, fitted well with the exponential distribution expected under the assumption of sudden cessation of admixture (*SI Appendix*, section 5). For each population, the times, in generations before present, at cessation of admixture with distinct ancestries were estimated by the method of moments (Table 3).

We estimated that all upper-caste populations, except MPB from Northeast India, started to practice endogamy about 70 generations ago (Table 3). The length distributions of the AAA blocks and the ASI blocks within any one of these populations (GBR, WBR, IYR) were very similar (*SI Appendix*, section 5). The most parsimonious explanation of this is that the practice of gene flow between ancestries in India came to an abrupt end about 1,575 y ago (assuming 22.5 y to a generation). This time estimate belongs to the latter half of the period when the Gupta emperors ruled large tracts of India (Gupta Empire, 319–550 CE).

Except WBR, with whom the northeast populations are geographically proximal, we found that there is significant ATB ancestry only among AA speakers. Even though the AA speakers presently occupy fragmented geographical regions in India, their presence in Northeast India (Khasis inhabiting Assam and Riang inhabiting Tripura) may indicate a more shared habitat with TB speakers in earlier times. Consistent with an earlier estimate (33), we estimated that the extant TB speakers freely admixed until more recently, 1,500–1,000 y ago (Table 3). Our results indicate that tribal populations may have practiced admixture until more recent times compared to upper-caste populations.

Table 3. Estimates of time (in GBP) of contribution of each of the ancestral components to the populations considered

Population	ANI	AAA	ASI	ATB
GBR	NA*	69.3833	69.3265	†
WBR	NA	69.5409	68.3778	63.3518
MRT	NA	48.7989	48.92	†
IYR	NA	69.1751	71.699	†
PLN	NA	74.3893	76.1979	†
KAD	47.5509	60.7911	NA	†
IRL	39.4951	49.8475	NA	†
GND	77.6637	91.9575	70.509	58.1287
HO	54.0405	NA	67.8753	52.9333
SAN	54.8661	NA	71.5929	61.5647
KOW	46.5407	NA	55.7532	46.6478
MPB	69.7002	67.6769	70.4008	NA
THR	62.7826	65.2317	72.9749	NA
TRI	65.1124	69.6447	70.5565	NA

This table pertains to the 14 populations that are considered as admixed and excludes the four populations (KSH, PNY, BIR, JAM) that are considered as representatives of the ancestral components (ANI, ASI, AAA, ATB, respectively).

*See text for an explanation of "NA" (not applicable).

†The contribution of the ancestral component is too low for reliable estimation of time depth.

An asymmetry of admixture was also revealed; ABLs attributable to ANI among AA speakers, Dravidian tribes, and TB speakers are longer than those attributable to other ancestries (Table 3), indicating that the ancestral North Indian population continued to provide genomic inputs into these populations (Table 3) well after inputs from other ancestries had ceased.

Discussion

By sampling populations, especially the autochthonous tribal populations, which represent the geographical, ethnic, and linguistic diversity of India, we have inferred that at least four distinct ancestral components—not two, as estimated earlier (9, 10)—have contributed to the gene pools of extant populations of mainland India. The Andaman archipelago was peopled by members of a distinct, fifth ancestry.

The absence of significant resemblance with any of the neighboring populations is indicative of the ASI and the AAA being early settlers in India, possibly arriving on the "southern exit" wave out of Africa. Differentiation between the ASI and the AAA possibly took place after their arrival in India (ADMIXTURE analysis with $K = 3$ shows ASI plus AAA to be a single population in *SI Appendix*, Fig. S2). The ANI and the ATB can clearly be rooted to the CS-Asians and E-Asians (Fig. 3 and *SI Appendix*, Fig. S7*B*), respectively; they likely entered India through the northwest and northeast corridors, respectively. Ancestral populations seem to have occupied geographically separated habitats. However, there was some degree of early admixture among the ancestral populations (ref. 9 and this study) as evidenced by extant populations possessing multiancestral components and some geographical displacements as well (6).

We have provided evidence that gene flow ended abruptly with the defining imposition of some social values and norms. The reign of the ardent Hindu Gupta rulers, known as the age of Vedic Brahminism, was marked by strictures laid down in Dharmaśāstra—the ancient compendium of moral laws and principles for religious duty and righteous conduct to be followed by a Hindu—and enforced through the powerful state machinery of a developing political economy (15). These strictures and enforcements resulted in a shift to endogamy. The evidence of more recent admixture among the Maratha (MRT) is in agreement with the known history of the post-Gupta Chalukya (543–753 CE) and the Rashtrakuta empires (753–982 CE) of western India, which established a clan of warriors (Kshatriyas) drawn from the local peasantry (15). In eastern and northeastern India, populations such as the West Bengal Brahmins (WBR) and the TB populations continued to admix until the emergence of the Buddhist Pala dynasty during the 8th to 12th centuries CE. The asymmetry of admixture, with ANI populations providing genomic inputs to tribal populations (AA, Dravidian tribe, and TB) but not vice versa, is consistent with elite dominance and patriarchy. Males from dominant populations, possibly upper castes, with high ANI component, mated outside of their caste, but their offspring were not allowed to be inducted into the caste. This phenomenon has been previously observed as asymmetry in homogeneity of mtDNA and heterogeneity of Y-chromosomal haplotypes in tribal populations of India (6) as well as the African Americans in United States (34). In this study, we noted that, although there are subtle sex-specific differences in admixture proportions, there are no major differences in inferences about population relationships and peopling whether X-chromosomal or autosomal data are used. We have also found our inferences to become more robust when our data are jointly analyzed with HGDP data.

We surmise that the number of ancestral components in the populations of India may have been underestimated by Reich et al. (9) because of (*i*) lack of inclusion of tribal populations, who are considered by anthropologists to be the autochthones of India, and (*ii*) inadequate representation of the geocultural diversity of India in the set of sampled populations, and (*iii*) selective removal of some populations based on deviance of their genomic profiles. Our study has corrected this deficiency and has provided a more robust explanation of the genomic diversities and affinities among extant populations of the Indian subcontinent, elucidating in finer detail the peopling of the region.

Methods

Ethical Approval and Informed Consent. DNA samples were collected with informed consent and after obtaining approvals of institutional ethics committees of the Indian Statistical Institute and the National Institute of BioMedical Genomics.

DNA Isolation, Assessment of Quality and Quantity. DNA was isolated by the salting-out method (35). Quantity and quality of isolated DNA were assessed using NanoDrop 8000 spectrophotometer.

DNA Microarray Analysis and Data Curation. Genotyping of each DNA sample was done using Illumina Omni 1-Quad, version 1.0, DNA analysis bead chip on IlluminaiScan, using the manufacturer's protocol as described in Infinium HD Assay Super Protocol Guide, catalog WG-901–4002. Genotype calling was done using Illumina Genome Studio following Genotyping Module, version 1.0, part 11319113. Quality metric, Gen Call score threshold was set to 0.25 to determine higher stringency in genotype calling. Markers with genotype calls for >90% individuals were included only (details in *SI Appendix*, section 6).

Because there was no information available about the sex of the individuals sampled, we inferred sex from the X-chromosome genotype. If the inbreeding (homozygosity) estimate (F) was more than 0.8, the individual was inferred to be a male; she was inferred to be a female if F was less than 0.2 (36) (*SI Appendix*, section 6).

Population Structure. An unsupervised clustering algorithm, ADMIXTURE (18), was run on our high-density dataset to explore global patterns of population structure varying the number of ancestral clusters ($K = 2$ through 6) and were successively tested. As LD can adversely affect the inferences of ADMIXTURE (18), the program was run on multiple datasets after pruning SNPs at LD (*SI Appendix*, sections 1 and 2). Cross-validation errors for each K are available in *SI Appendix*, sections 1 and 2. PCA was applied to both datasets using EIGENSOFT 4.2 (17) and plots were generated using R 2.12.2 (https://www.r-project.org/). fineSTRUCTURE (21) and *frappe* (20) were run using the default parameters.

Phasing. Haplotype estimation both for the autosomes and X chromosome from genome-wide data of unrelated individuals was separately done using segmented haplotype estimation and imputation tool (Shapeit2) (30, 31). Shapeit2 uses a modified hidden Markov model. The algorithm was run only on genotypes with no missing data. Both the model parameters and the number of iterations were set as the default options in Shapeit2.

ABL Estimation. Local ancestry assignment was performed using PCAdmix (https://sites.google.com/site/pcadmix/) (32) with $K = 4$ ancestral groups. This approach relies on phased data from reference panels and the admixed individuals. The populations Khatri (KSH), Paniya (PNY), Birhor (BIR), and Jamatia (JAM) with more than 97% ancestry from the ANI, ASI, AAA, and ATB, respectively, were used as the reference panel. Each chromosome is analyzed independently, and local ancestry assignment is based on loadings from PCA of the four putative ancestral population panels. PCAdmix partitions the genomic data into nonoverlapping windows, and for each of these windows the distribution of individual scores within a population is modeled by fitting a multivariate normal distribution (32). Given an admixed chromosome, these distributions are used to compute likelihoods of belonging to each panel. We only considered local ancestry assignments using a greater than 0.85 posterior probability threshold for each window (*SI Appendix*, section 6).

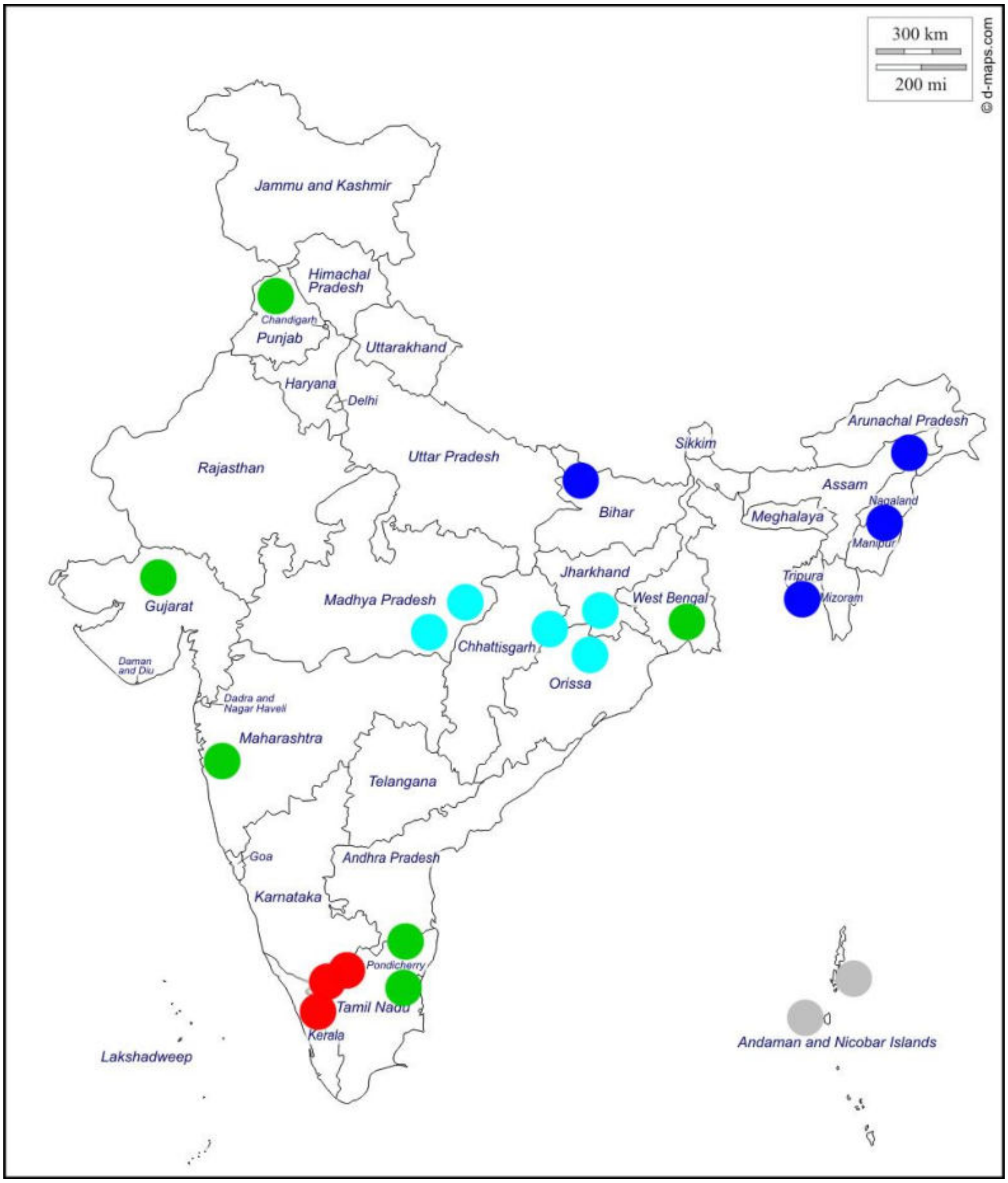
Data curation, statistical analysis, and graphical representations were done using PLINK (36), version 1.07 (pngu.mgh.harvard.edu/~purcell/plink/download.shtml), and R, version 2.12.2 (https://www.r-project.org/).

ACKNOWLEDGMENTS. We thank all of the individuals who volunteered to donate their DNA for the analysis. In addition to some of the authors of this study, sample collection with informed consent was done by C. S. Chakraborty, R. Lalthantluanga, M. Mitra, A. Ramesh, N. K. Sengupta, S. K. Sil, J. R. Singh, C. M. Thakur, and M. V. Usha Rani. We thank B. Dey and B. Bairagya for the sample curation; I. Bagchi and R. Dhar for assistance in generating DNA microarray data; and S. Bhattacharjee, A. Mukherjee, N. K. Biswas, D. Tagore, and S. Chakraborty for assistance in preparing figures. The sample collection and some DNA analyses were partially supported by agencies of the Government of India, including Department of Biotechnology, Department of Science and Technology, and the Indian Council of Medical Research (primarily to P.P.M.).

1. Cann RL (2001) Genetic clues to dispersal in human populations: Retracing the past from the present. *Science* 291(5509):1742–1748.
2. Mellars P (2006) Going east: New genetic and archaeological perspectives on the modern human colonization of Eurasia. *Science* 313(5788):796–800.
3. Macaulay V, et al. (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* 308(5724):1034–1036.
4. Quintana-Murci L, et al. (1999) Genetic evidence of an early exit of *Homo sapiens* sapiens from Africa through eastern Africa. *Nat Genet* 23(4):437–441.
5. Sengupta S, et al. (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(2):202–221.
6. Basu A, et al. (2003) Ethnic India: A genomic view, with special reference to peopling and structure. *Genome Res* 13(10):2277–2290.
7. Cordaux R, et al. (2004) Independent origins of Indian caste and tribal paternal lineages. *Curr Biol* 14(3):231–235.
8. Bamshad M, et al. (2001) Genetic evidence on the origins of Indian caste populations. *Genome Res* 11(6):994–1004.
9. Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. *Nature* 461(7263):489–494.
10. Moorjani P, et al. (2013) Genetic evidence for recent population mixture in India. *Am J Hum Genet* 93(3):422–438.
11. Diamond J, Bellwood P (2003) Farmers and their languages: The first expansions. *Science* 300(5619):597–603.
12. Kumar V, et al. (2007) Y-chromosome evidence suggests a common paternal heritage of Austro-Asiatic populations. *BMC Evol Biol* 7:47.
13. Chaubey G, et al. (2011) Population genetic structure in Indian Austroasiatic speakers: The role of landscape barriers and sex-specific admixture. *Mol Biol Evol* 28(2):1013–1024.
14. Chaubey G, Metspalu M, Kivisild T, Villems R (2007) Peopling of South Asia: Investigating the caste-tribe continuum in India. *BioEssays* 29(1):91–100.
15. Thapar R (2004) *Early India: From the Origins to AD 1300* (Univ of California Press, Berkeley, CA).
16. Abdulla MA, et al.; HUGO Pan-Asian SNP Consortium; Indian Genome Variation Consortium (2009) Mapping human genetic diversity in Asia. *Science* 326(5959):1541–1545.
17. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904–909.
18. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655–1664.
19. Russell RV (1916) *The Tribes and Castes of the Entral Provinces of India* (Macmillan, London), Vol 1.
20. Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: Analytical and study design considerations. *Genet Epidemiol* 28(4):289–301.
21. Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. *PLoS Genet* 8(1):e1002453.
22. Li JZ, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319(5866):1100–1104.
23. Rosenberg NA, et al. (2002) Genetic structure of human populations. *Science* 298(5602):2381–2385.
24. Haak W, et al. (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522(7555):207–211.
25. Chang W, Chundra C, Hall D, Garrett A (2015) Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1):194–244.
26. Kuiper PBJ, ed (1991) *Aryans in the Rigveda*. Leiden Studies in Indo-European I (RODOPI, Amsterdam).
27. Witzel M (1999) Substrate languages in Old Indo-Aryan (Rigvedic, Middle and Later Vedic). *Electron J Vedic Stud* 5:1–97.
28. Pool JE, Nielsen R (2009) Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* 181(2):711–719.
29. Jin W, Li R, Zhou Y, Xu S (2014) Distribution of ancestral chromosomal segments in admixed genomes and its implications for inferring population history and admixture mapping. *Eur J Hum Genet* 22(7):930–937.
30. Delaneau O, Marchini J, Zagury JF (2012) A linear complexity phasing method for thousands of genomes. *Nat Methods* 9(2):179–181.
31. Delaneau O, Zagury JF, Marchini J (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 10(1):5–6.
32. Brisbin A, et al. (2012) PCAdmix: Principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum Biol* 84(4):343–364.
33. Karlsson EK, et al. (2013) Natural selection in a Bangladeshi population from the cholera-endemic Ganges River delta. *Sci Transl Med* 5(192):192ra86.
34. Lind JM, et al. (2007) Elevated male European and female African contributions to the genomes of African American individuals. *Hum Genet* 120(5):713–722.
35. Miller SA, Dykes DD, Polesky HF (1988) A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 16(3):1215.
36. Purcell S, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–575.

Supplementary Information

Fig S1: Map of India showing approximate locations of sampling of the populations included in this study. Populations shown in ‘grey’ are populations from the Andaman and Nicober archipelago. Populations shown in ‘red’ are Dravidian speaking tribal populations from the Nilgiri Hills in Southern India. Populations shown in ‘cyan’ are Austro-Asiatic speaking tribal populations from the East and Central India. Populations shown in ‘green’ are caste populations primarily speaking the Indo-European language. Populations shown in ‘blue’ are Tibeto-Burman speaking populations of North-East India and are predominantly tribes except the Manipuri Brahmins. (More description in Table 1)



Supplementary Information 1 (SI-1)

Detailed results of ADMIXTURE analysis with ***all*** 20 populations

The conclusions that we wish to highlight in **SI-1** are:

1) The cross-validation (CV) error is minimized when K=5, irrespective of whether the entire dataset or the LD pruned dataset is used, or whether CV is taken to be 5% and 10%

2) At K=2, a small proportion (mean=0.06) of Jarwa and Onge ancestries are noted to be present in individuals drawn from mainland populations, (first panel in Fig. Supplement). However, this proportion decreases as we increase K. At K=3 it stands at 0.02 and at K=4, 5 it further reduces to 0.004). Therefore, it appears that these estimates are of statistical noise, rather than real admixture estimates.

Contents for this section

Fig. Supplement(i) Individual ancestry inferred with ADMIXTURE with K = 2, 3 and 4 are plotted. Each individual is represented by a vertical line partitioned into colored segments whose heights correspond to his/her ancestry coefficients for up to four inferred ancestral groups. Population labels were added only after each individual's ancestry had been estimated; the labels were used to order the samples in plotting. *(SNPs were pruned only to include those for which the pairwise Linkage-Disequilibrium was less than 0.5)*

Fig. Supplement (ii) Individual ancestry inferred with ADMIXTURE with K = 5 for which the CV error is minimized. *(SNPs were pruned only to include those for which the pairwise Linkage-Disequilibrium was less than 0.5)*

Table S1. (A) (B) (C) and (D) Cross-Validation error for different choices of K,

(A) CV error calculated at 5% with all SNPs
(B) CV error calculated at 5% after pruning SNPs in LD *(pairwise LD <0.5)*
(C) CV error calculated at 10% with all SNPs
(D) CV error calculated at 10% after pruning SNPs in LD *(pairwise LD <0.5)*

Fig. Supplement(iii).(A), (B), (C), (D): These graphs correspond to Table S1 (A) (B)(C) and (D) respectively

Table S1 E: The ADMIXTURE Estimates pertaining to K=5 for 20 populations*(SNPs were pruned only to include those for which the pairwise Linkage-Disequilibrium was less than 0.5)*

Fig. Supplement(i): Individual ancestry, ***(all 367 individuals from 20 populations)*** inferred with ADMIXTURE with K = 2, 3 and 4. Each individual is represented by a vertical line partitioned into colored segments whose heights correspond to his/her ancestry coefficients in up to four inferred ancestral groups. Population labels were added only after each individual's ancestry had been estimated; they were used to order the samples in plotting.

With K=2 The mainland Indian populations separate from the Jarawa (JRW) and Onge(ONG) or the hunter-gatherer tribal populations of Andaman and Nicober Islands

With K=3 In addition to the Island and mainland separation, the Tibeto-Burman speaking populations from NE-India separate from the other mainland populations.

With K=4 The caste populations in India, primarily Indo-European speakers, separate from the tribal populations (i.e the Austro-Asiatic speaking tribes of East and Central India and the Dravidian speaking tribes of Nilgiri Hills)

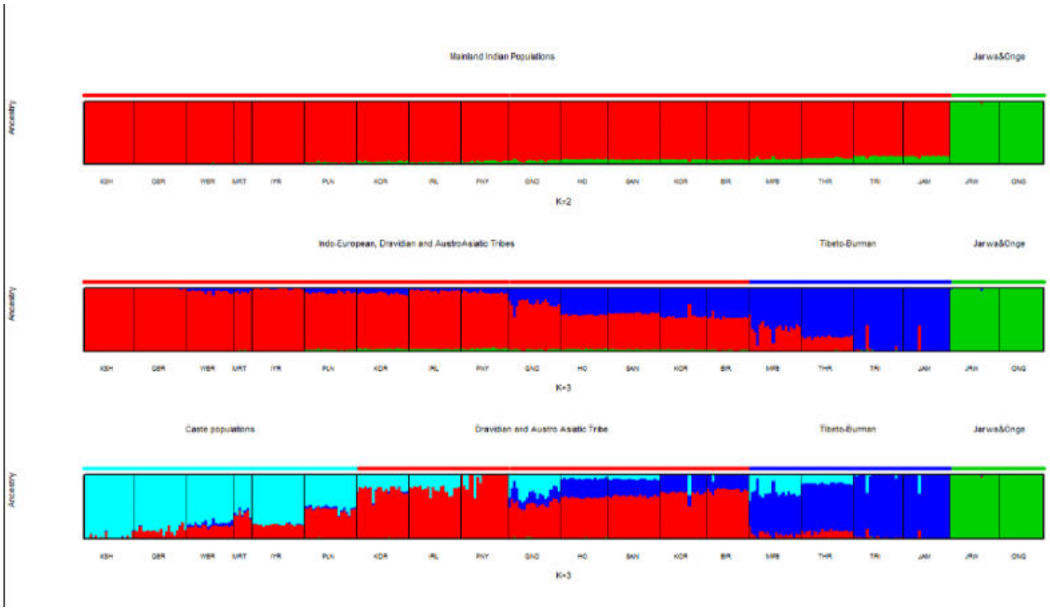


Fig. Supplement (ii): Individual ancestry ***(all 367 individuals from 20 populations)*** inferred with ADMIXTURE with K = 5 for which the CV error is minimized. The Austro-Asiatic speaking tribes of East and Central India separate from the Dravidian speaking tribes of Nilgiri Hills. However we see substantial evidence of admixture in the populations.

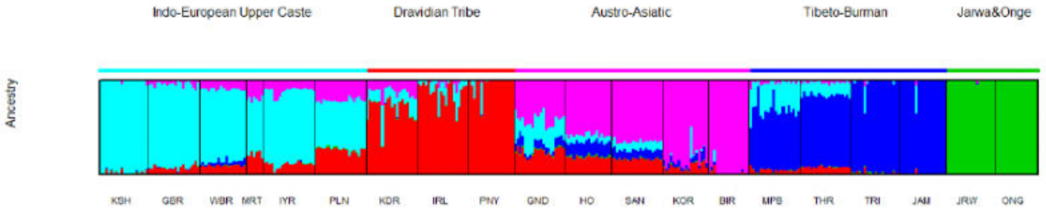


Table S1:

A. CV 5% all SNPs

K	CV Error
2	0.4416
3	0.43605
4	0.4327
5	0.4316
6	0.43255

B. CV 5% LD pruned SNPs

K	CV Error
2	0.51511
3	0.50892
4	0.5053
5	0.5039
6	0.50571

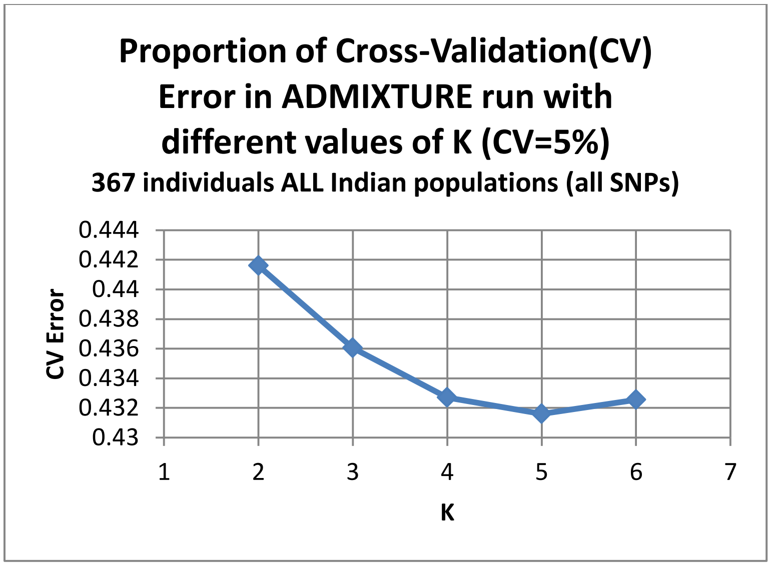
C. CV 10% all SNPs

K	CV Error
2	0.441
3	0.43522
4	0.43168
5	0.43022
6	0.43078

D. CV 10% LD pruned SNPs

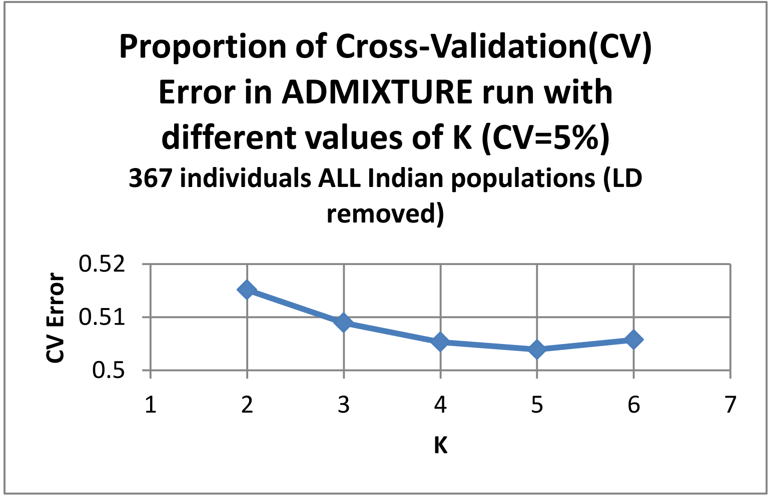
K	CV Error
2	0.441
3	0.43522
4	0.43168
5	0.43022
6	0.43078

Fig. Supplement (iii) (A), (B), (C) and (D)



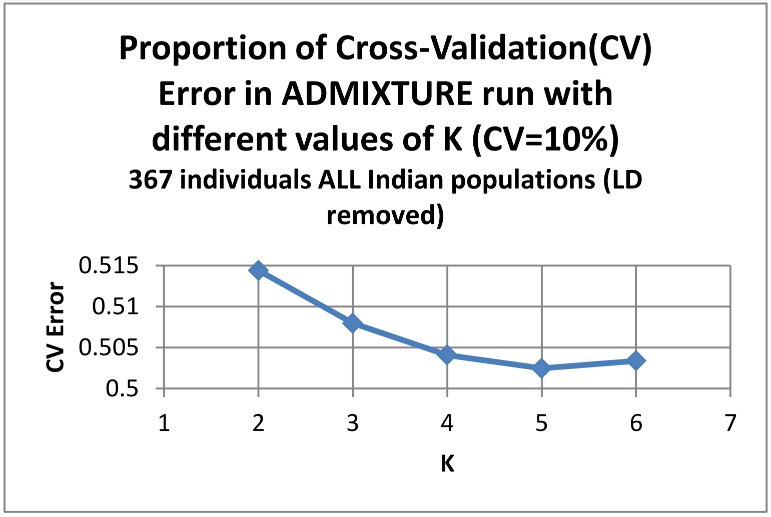
Proportion of Cross-Validation(CV) Error in ADMIXTURE run with different values of K (CV=5%)

367 individuals ALL Indian populations (all SNPs)



Proportion of Cross-Validation(CV) Error in ADMIXTURE run with different values of K (CV=5%)

367 individuals ALL Indian populations (LD removed)



Proportion of Cross-Validation(CV) Error in ADMIXTURE run with different values of K (CV=10%)

367 individuals ALL Indian populations (LD removed)

Table S1E: ADMIXTURE Estimates with K=5 for 20 populations

Population Name	ANI	ASI	AAA	ATB	Ancestral Andaman
KSH	**0.97**	**0.014**	**0.007**	**0.007**	**0.002**
GBR	0.875	0.07	0.045	0.006	0.003
WBR	0.769	0.086	0.097	0.042	0.005
MRT	0.59	0.199	0.207	0.001	0.002
IYR	0.785	0.106	0.105	0.001	0.003
PLN	0.523	0.249	0.219	0	0.009
KAD	0.171	0.682	0.136	0.003	0.008
IRL	0.132	0.835	0.032	0	0.002
PNY	**0.037**	**0.959**	**0.002**	**0.002**	**0**
GND	0.249	0.241	0.417	0.082	0.011
HO	0.072	0.174	0.591	0.15	0.013
SAN	0.087	0.157	0.656	0.093	0.007
KOR	0.027	0.079	0.823	0.066	0.005
BIR	**0.008**	**0.013**	**0.972**	**0.006**	**0**
MPB	0.292	0.038	0.032	0.634	0.005
THR	0.139	0.074	0.037	0.747	0.003
TRI	0.024	0.012	0.013	0.943	0.009
JAM	**0.016**	**0.004**	**0.003**	**0.975**	**0.002**
JRW	**0**	**0**	**0.002**	**0.001**	**0.996**
ONG	**0**	**0**	**0**	**0**	**1**

Supplementary Information 2 (SI-2)

In SI-2 we detail the results of our ADMIXTURE run with ***the 18mainland Indian*** populations

The observations of interest that we have emphasized in this **SI-2** are:

1) The cross-validation (CV) error is minimized when K=4, irrespective of whether we have used the entire dataset or the LD pruned dataset, or whether we have used CV at 5 fold or 10 fold. [Table S2 (A)−(F); Fig. S3 (A)−(F)]

2) ADMIXTURE, which explored a very high-dimensional likelihood space, was robust in detecting population structure and the inferences are stable in multiple runs of the program with a random initialization (Random seed as starting point). [Table S3 (A) – (D)]

3) Multiple programs which estimate ancestry and admixture proportions from genotype-data converge to similar inference about population structure and admixture in Indian populations. [Table S4, S5; Fig. S4]. Also elaborating on some findings using fineSTRUCTURE.

4) We explore the sex-bias in admixture proportions. [Fig. S5, S6]

Contents for this section

Fig. S2: Individual ancestry inferred with ADMIXTURE with K = 2 and 3. Each individual is represented by a vertical line partitioned into colored segments whose heights correspond to his/her ancestry coefficients in up to four inferred ancestral groups. Population labels were added only after each individual's ancestry had been estimated; they were used to order the samples in plotting.

Table S2 (A)−(F): Cross-Validation error for different choice of K

(A) CVE calculated at 5-fold when all the SNPs are included (no LD pruning)
(B) CVE calculated at 10-fold when all the SNPs are included (no LD pruning)
(C) CVE calculated at 5-fold when SNPs with pairwise LD <0.5 were included
(D) CVE calculated at 10-fold when SNPs with pairwise LD <0.5 were included
(E) CVE calculated at 5-fold when SNPs with pairwise LD <0.1 were included
(F) CVE calculated at 10-fold when SNPs with pairwise LD <0.1 were included

Fig. S3 (A)−(F): The graphs corresponding to Table S2 (A)−(F) respectively

Table S3: Summary Table of cross-validation error (CVE) generated from multiple runs (10) of ADMIXTURE, using
(A) CVE at 5 fold all SNPs with pairwise LD <0.5
(B) CVE at 10 fold all SNPs with pairwise LD <0.5
(C) CVE at 5 fold all SNPs with pairwise LD <0.1
(D) CVE at 10 fold all SNPs with pairwise LD <0.1

Table S4: The *frappe* estimates with K=4 for 18 populations

Table S5: Details of the 69 populations as identified by fineSTRUCTURE

Fig. S4A and B: Relationships among the 69 populations identified by fineSTRUCTURE (Table S5)

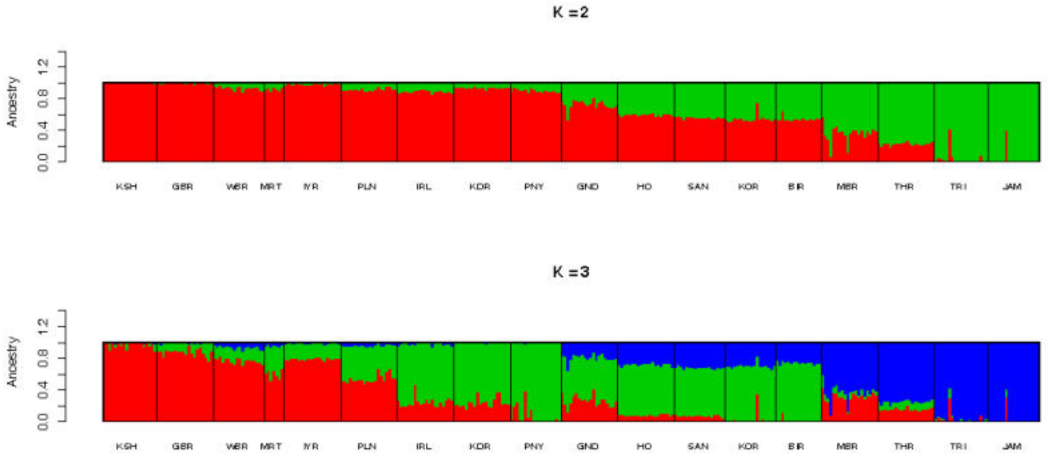
Fig. S5A: ADMIXTURE with K=4 on 107 females from 15 populations shows more ATB component and reduced ANI component in the X-Chromosome of individuals from KSH, GBR, MRT, IYR, PLN as well as GND, HO, SAN, KOR populations.

Fig. S5B: Q-Q Plot of the 107 females.

Fig. S6A: Dendrogram of the X-chromosome haplotypes show separate clades belonging to different populations (This is a large format figure and can be viewed clearly when magnified)

Fig. S6A: Dendrogram of the X-chromosome haplotypes show separate clades belonging to different populations (This is a large format figure and can be viewed clearly when magnified). The color codes are consistent with the colors used in previous figures.

Fig. S2: Individual ancestry inferred with ADMIXTURE with K = 2 and 3. Each individual is represented by a vertical line partitioned into colored segments whose heights correspond to his/her ancestry coefficients in up to four inferred ancestral groups. Population labels were added only after each individual's ancestry had been estimated; they were used to order the samples in plotting.



With K=2, like before **(SI-1)**, the close to 100% GREEN are the TB speakers from NE India and close to 100% RED are caste populations, primarily IE speakers of North India. We have identified these GREEN as the ATB component.

With K=3, like before **(SI-1)**, the close to 100% BLUE are the TB speakers from NE India. The other population (RED with K=2) is split into RED and GREEN. We have identified the 'RED' component as the ANI ancestry. The GREEN is the combined (ASI+AAA), which separate at K=4.

Table S2: Cross-Validation error (CVE) for different choices of K clearly shows CVE to be minimum when K=4

(A) CVE calculated at 5-fold when all the SNPs are included (no LD pruning)
(B) CVE calculated at 10-fold when all the SNPs are included (no LD pruning)
(C) CVE calculated at 5-fold when SNPs with pairwise LD <0.5 were included
(D) CVE calculated at 10-fold when SNPs with pairwise LD <0.5 were included
(E) CVE calculated at 5-fold when SNPs with pairwise LD <0.1 were included
(F) CVE calculated at 10-fold when SNPs with pairwise LD <0.1 were included

Table S2A:

K	CV Error
2	0.54127
3	0.53672
4	0.53508
5	0.53678

Table S2B:

K	CV Error
2	0.54069
3	0.53585
4	0.53388
5	0.5347

Table S2C:

K	CV Error
2	0.50353
3	0.4998
4	0.4989
5	0.50011

Table S2D:

K	CV Error
2	0.50283
3	0.49872
4	0.49738
5	0.49859

Table S2E:

K	CV Error
2	0.50331
3	0.49983
4	0.49853
5	0.49953

Table S2F:

K	CV Error
2	0.50265
3	0.49882
4	0.49737
5	0.49766

Fig. S3 (A)–(F): The graphs corresponding to Table S2 (A)–(F) respectively

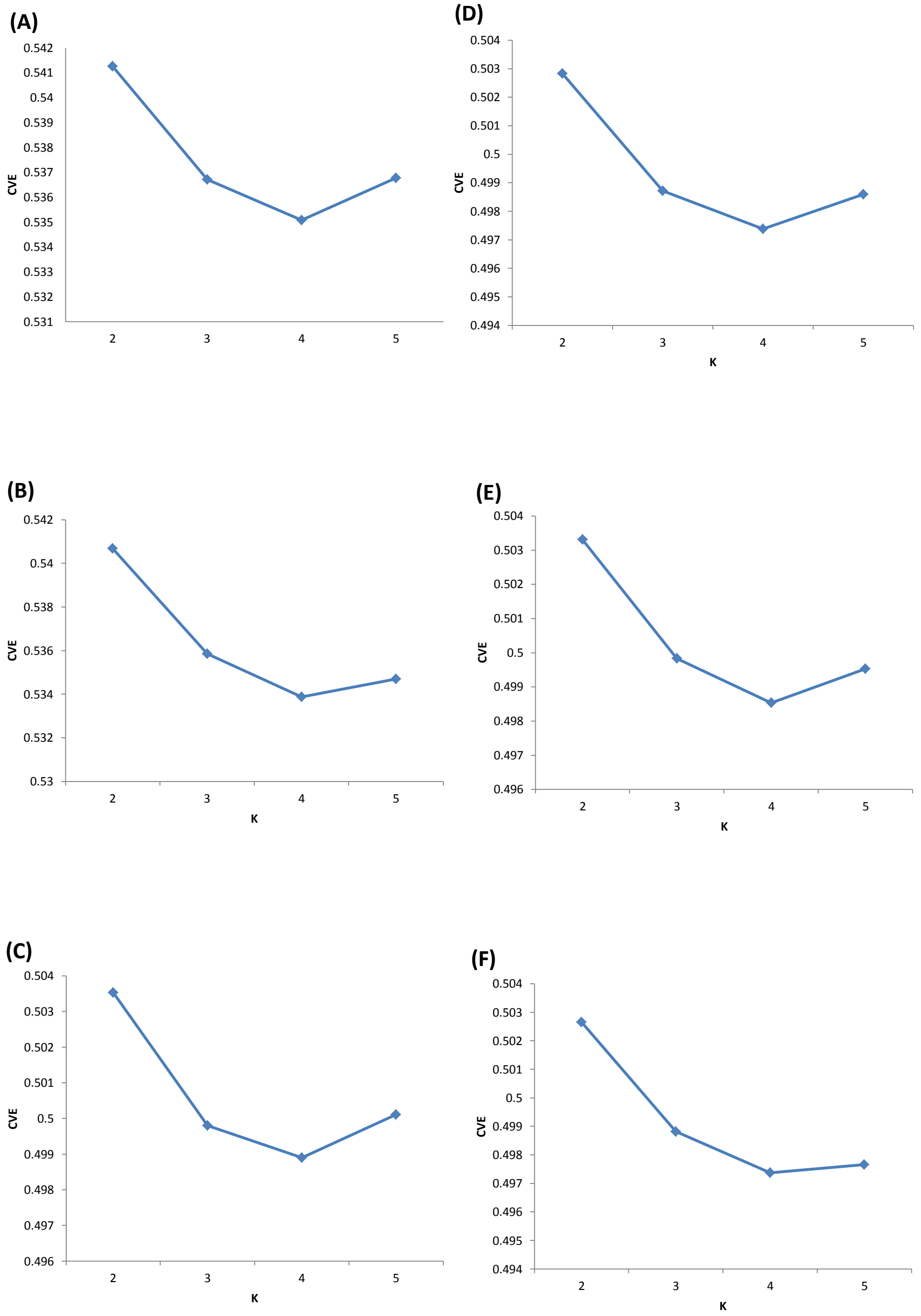


Table S3: Summary Table of cross-validation error (CVE) generated from multiple runs (10) of ADMIXTURE, using

(A) CVE at 5 fold all SNPs with pairwise LD <0.5

(B) CVE at 10 fold all SNPs with pairwise LD <0.5
(C) CVE at 5 fold all SNPs with pairwise LD <0.1
(D) CVE at 10 fold all SNPs with pairwise LD <0.1

Table S3A

	K=2	K=3	K=4	K=5
Mean	0.5189967	0.5150022	0.5136544	0.5155000
Standard Deviation	3.08×10^{-05}	3.52×10^{-05}	6.48×10^{-05}	2.23×10^{-04}
Minimum	0.51896	0.51495	0.51358	0.51533
Maximum	0.51904	0.51508	0.51378	0.51608

Table S3B

	K=2	K=3	K=4	K=5
Mean	0.5183767	0.5140633	0.5123056	0.5133944
Standard Deviation	1.22×10^{-05}	2.34×10^{-05}	2.40×10^{-05}	2.59×10^{-04}
Minimum	0.51835	0.51402	0.51227	0.51317
Maximum	0.51839	0.51409	0.51234	0.51387

Table S3C

	K=2	K=3	K=4	K=5
Mean	0.5034133	0.4997322	0.4986922	0.4996167
Standard Deviation	9.92×10^{-05}	7.72×10^{-05}	1.88×10^{-04}	2.65×10^{-04}
Minimum	0.50331	0.49962	0.49845	0.49905
Maximum	0.50356	0.49983	0.49904	0.50001

Table S3D

	K=2	K=3	K=4	K=5
Mean	0.5026544	0.4987611	0.4972900	0.4974733
Standard Deviation	2.24×10^{-05}	4.28×10^{-05}	1.17×10^{-04}	2.08×10^{-04}
Minimum	0.50262	0.49866	0.49713	0.49741
Maximum	0.50268	0.49882	0.49741	0.49767

Table S4: Ancestry proportions of 18 mainland Indian populations as estimated by the best fit (K=4) model in *frappe*

Population Name	ANI ancestry	ASI ancestry	AAA ancestry	ATB ancestry
KSH	**0.9793**	**0.0149**	**0.0045**	**0.0013**
GBR	0.8823	0.0759	0.0412	6.00E-04
WBR	0.7663	0.0994	0.101	0.0332
MRT	0.5751	0.2141	0.2105	3.00E-04
IYR	0.8046	0.111	0.0837	7.00E-04
PLN	0.4902	0.2761	0.2331	6.00E-04
KAD	0.0895	0.7681	0.1414	0.0011
IRL	0.0532	0.9255	0.0213	0
PNY	**0.0252**	**0.9696**	**0.0052**	**0**
GND	0.3697	0.193	0.3756	0.0617
HO	0.0475	0.1705	0.7116	0.0704
SAN	0.0347	0.1933	0.6398	0.1321
KOR	0.0181	0.0471	0.9091	0.0257
BIR	**0.0082**	**0.0054**	**0.9864**	**0**
MPB	0.2635	0.0512	0.0351	0.6502
THR	0.0935	0.0951	0.0447	0.7667
TRI	0.0156	0.0084	0.0117	0.9643
JAM	**0.0149**	**0.0044**	**0.0031**	**0.9776**

Detailed result of fineSTRUCTURE analysis:

Table S5: The 69 subpopulations identified by fineSTRUCTURE

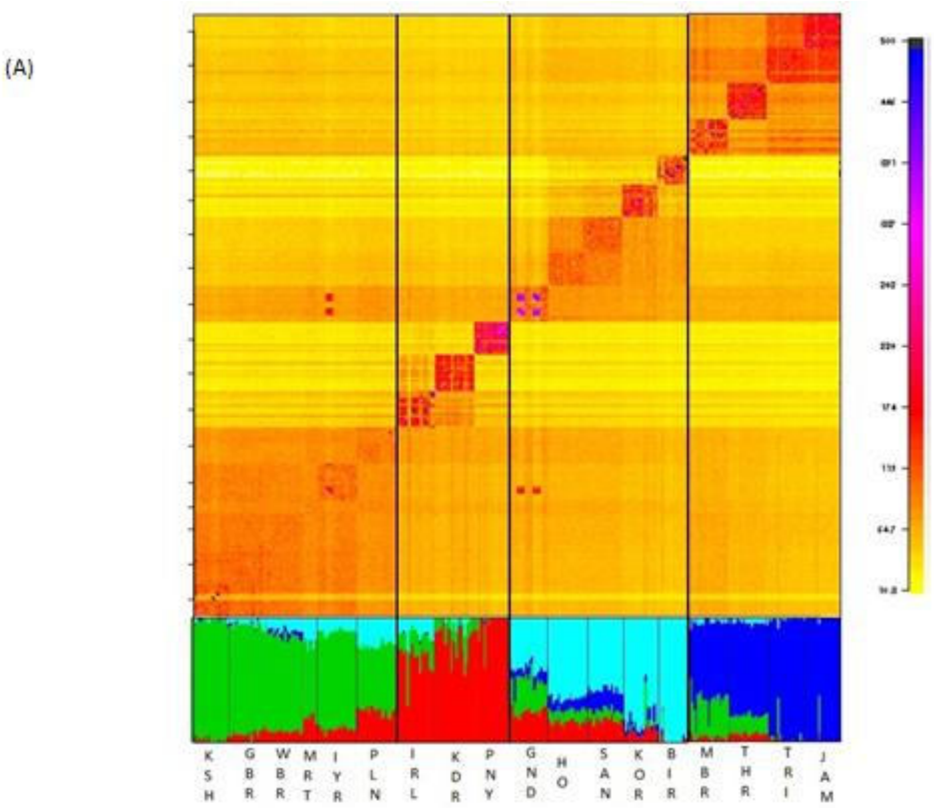
Sub-Population	Number of Individuals and Original Population Label
1	2IYR
2	14IYR
3	18WBR;4GBR;1IYR;1MRT
4	2KSH
5	6KSH;1GBR
6	2KSH
7	9KSH;15GBR
8	1PNY
9	2PNY
10	1PNY
11	1PNY
12	1PNY
13	12PNY
14	1KDR
15	4KDR
16	3IYR
17	8GND

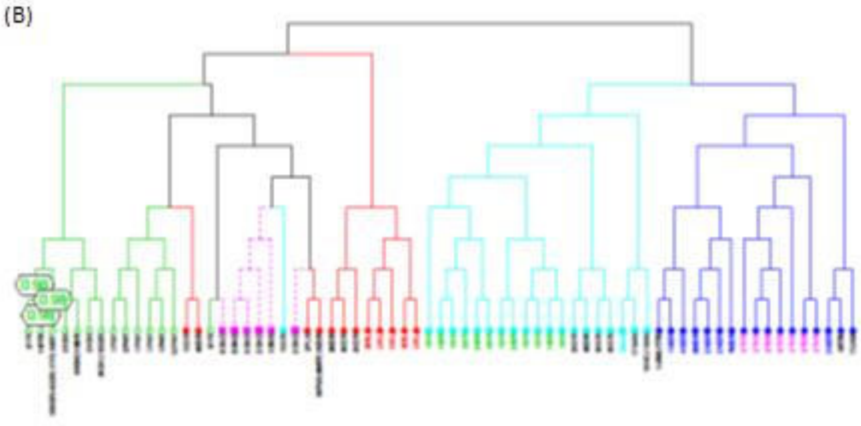
18	5GND
19	2GND
20	2GND
21	1GND
22	1KOR
23	2GND
24	2PLN
25	18PLN;6MRT;1KDR
26	3KDR
27	9KDR
28	2KDR
29	8IRL
30	2IRL
31	5IRL
32	3IRL
33	2IRL
34	1BIR
35	1BIR
36	1BIR
37	2BIR
38	2BIR
39	2BIR
40	1BIR
41	2BIR
42	1BIR
43	1BIR
44	1BIR
45	1BIR
46	2KOR
47	4KOR
48	5KOR
49	6KOR
50	2SAN
51	17SAN
52	18HO;1SAN
53	1JAM;1TRI
54	1MBR
55	4MBR
56	2MBR
57	2MBR
58	2MBR
59	7MBR
60	2THR
61	2THR
62	2THR
63	8THR
64	2THR
65	2THR
66	2THR

67	2MBR
68	18TRI
69	17JAM

Fig. S4A: Heat map of the ‘Coancestry Matrix’ of 331 individuals from 18 mainland Indian populations. The co-ancestry matrix broadly conforms to the inferences of the 4- ancestral components identified by ADMIXTURE.

Fig. S4B: Relationship between the 69 populations identified by fineSTRUCTURE (Table S5)





Supplementary text: Findings using fineSTRUCTURE

ADMIXTURE analysis indicated that the Gond (GND) is an extremely heterogeneous and admixed tribal population (Fig. 3B and Table 2). Both ADMIXTURE and fineSTRUCTURE have revealed that the upper caste Iyers (IYR), in spite of being Dravidian speakers and residing in south India, possess a high fraction of the ANI component (Fig. 3B, Table 2 and Fig. 3C). fineSTRUCTURE has also revealed the co-ancestry of the ANI component of IYR and GND, but no striking similarity of the ANI component with the other AA speaking Ho tribals living in the same geographical region (Fig. 3C). fineSTRUCTURE analysis has thus reestablished that some of the hunter-gatherer tribals of mainland India (Table 1) irrespective of their linguistic affiliation, have remained very isolated and demographically small after evolving from an ancestral population; these features have resulted in decreasing genomic similarities among them by genetic drift (Fig. 3C).

Sex-Bias in Admixture:

Fig. S5A: ADMIXTURE with K=4 on 107 females from 15 populations shows more ATB component and reduced ANI component in the X-Chromosome of individuals from KSH, GBR, MRT, IYR, PLN as well as GND, HO, SAN, KOR populations.

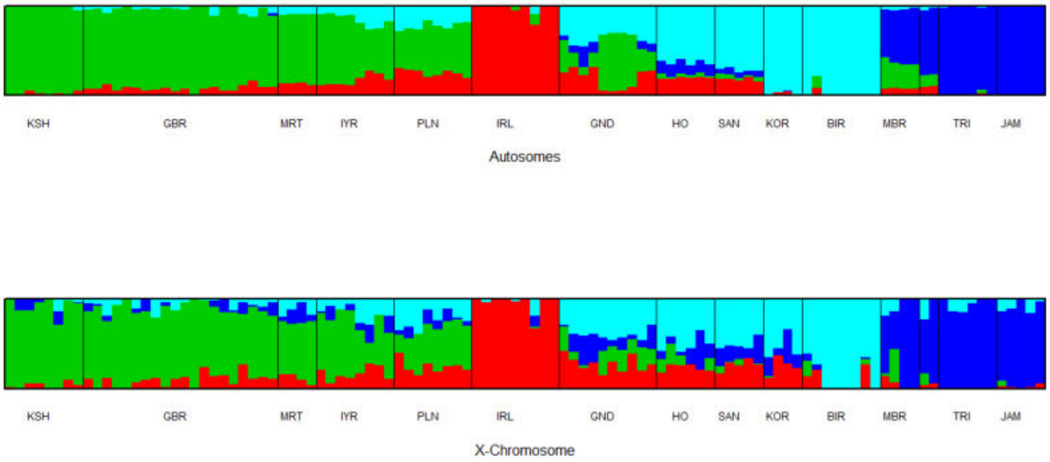


Fig. S5B: Q-Q Plot of the 107 females.

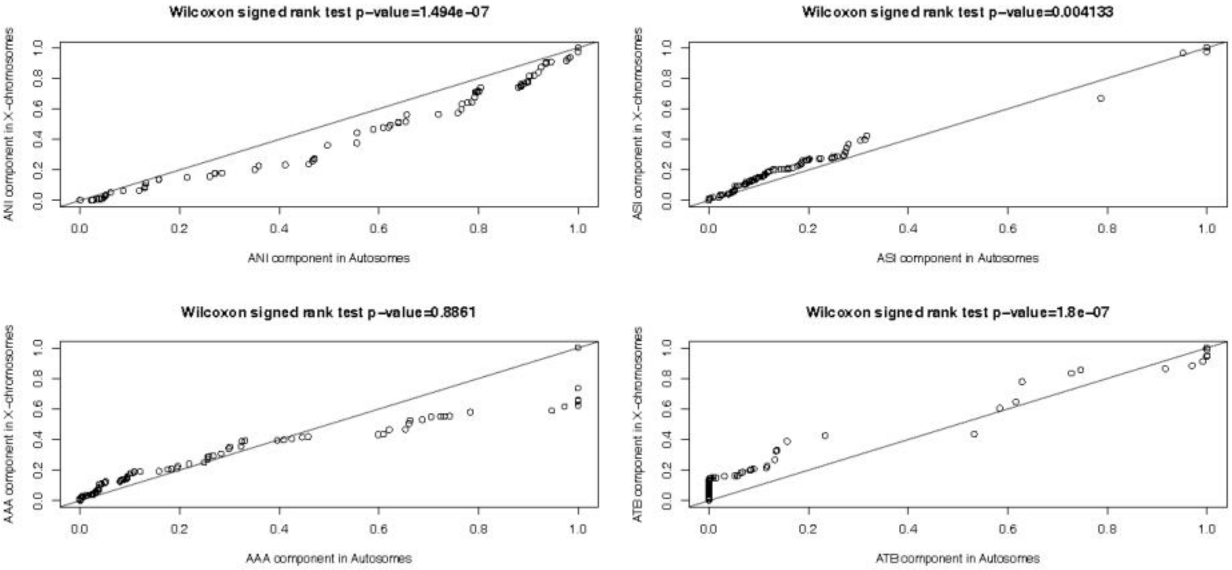


Fig. S6A: Dendrogram of the X-chromosome haplotypes show separate clades belonging to different populations (This is a large format figure and can be viewed clearly when magnified)

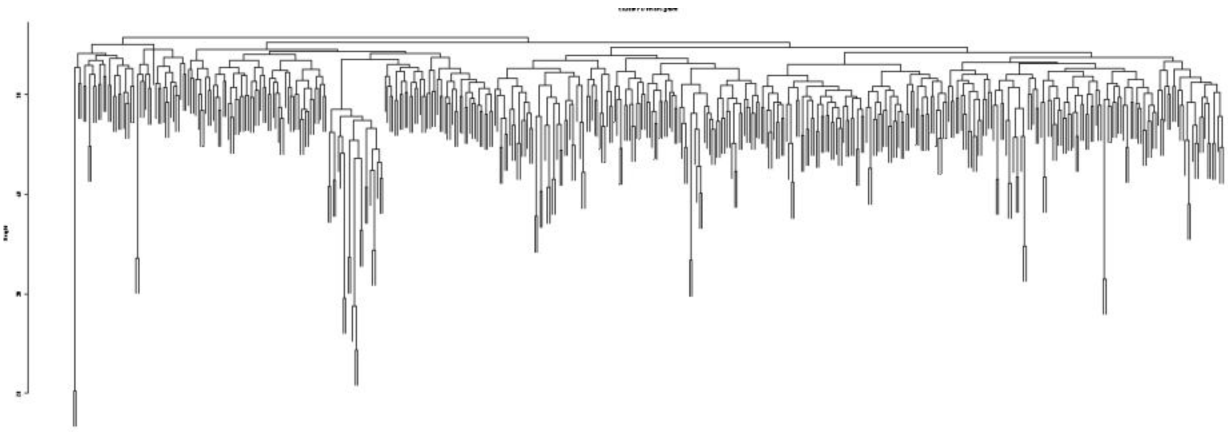


Fig. S6B: Dendrogram of the X-chromosome haplotypes show separate clades belonging to different populations (This is a large format figure and can be viewed clearly when magnified). The color codes are consistent with the colors used in previous figures.

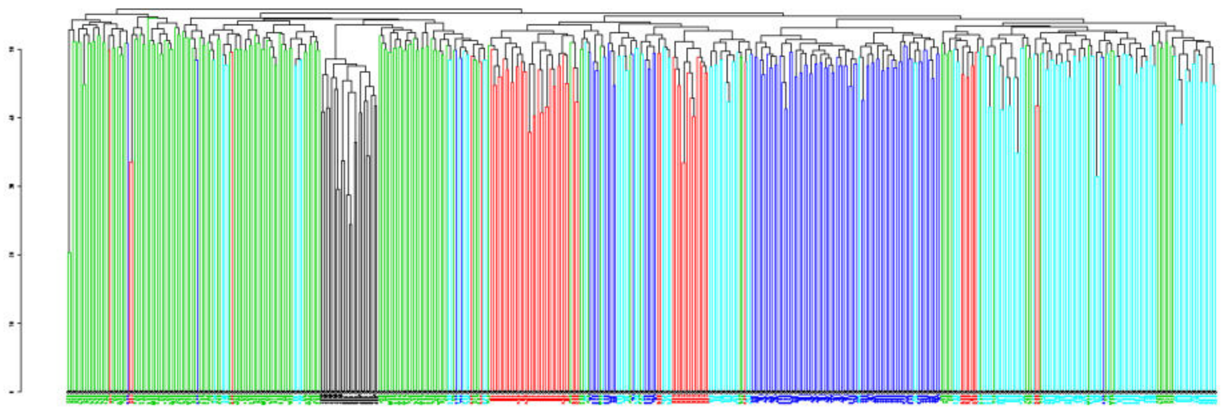
GREEN is used for haplotypes of individuals with a major ANI ancestry (i.e. KSH, GBR, IYR, MRT, PLN)

RED is used for haplotypes of individuals with a major ASI ancestry (IRL, KDR, PNY)

CYAN is used for haplotypes of individuals with a major AAA ancestry (GND, HO, SAN, KOR, BIR)

BLUE is used for haplotypes of individuals with a major ATB ancestry (MPB, THR, TRI, JAM)

BLACK is used for haplotypes of individuals from the JRW and ONG populations.



Supplementary Information 3 (SI 3)

The detail results exploring:

1) The genetic relationship of the ancestries present in mainland India with neighbouring populations. (page 26)

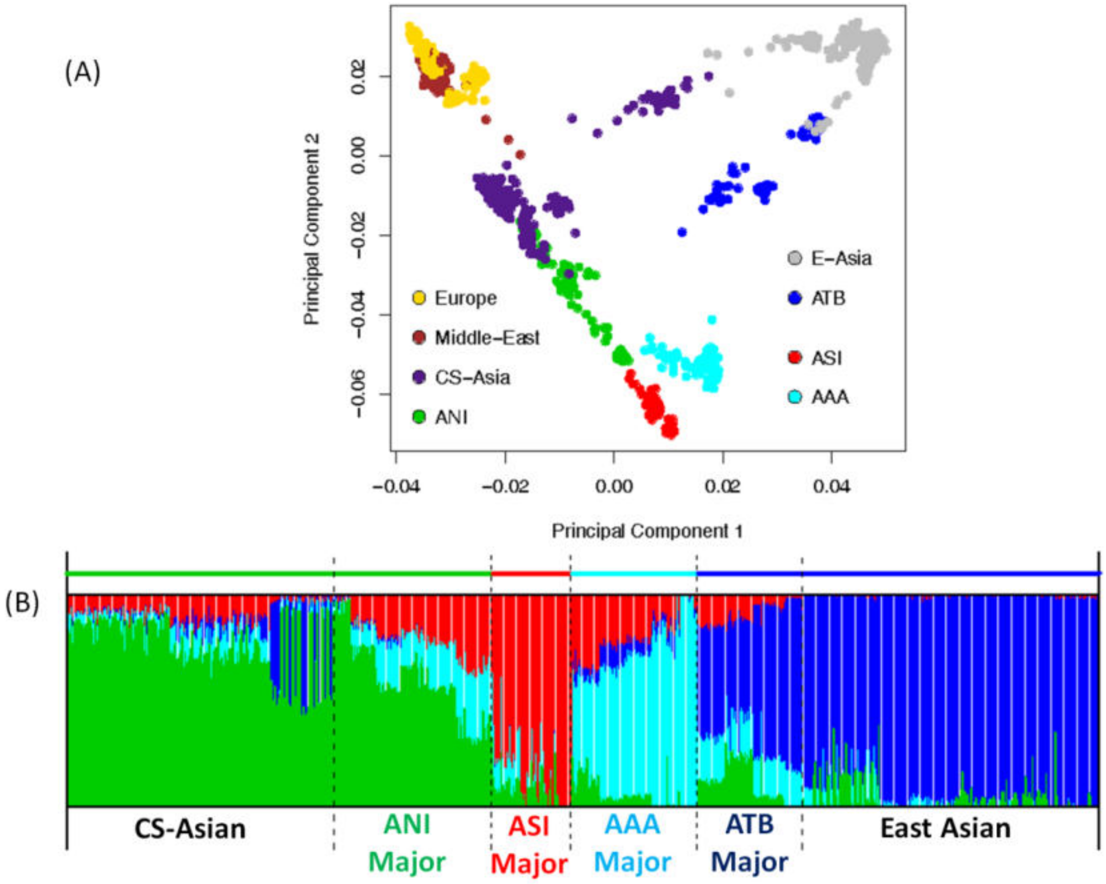


Fig. S7A: The PCA plot with Europeans, Middle-Easterners, Central-South Asians (CS-Asian), East-Asians (E-Asian) included in Human Genome Diversity Panel (HGDP) shows that the Europeans and Middle-Easterners cluster distinctly, in spite of being genetically close to the C-S Asians and populations which have high proportion of ANI ancestry

Fig. S7B: **Estimates of ancestral components of 331 individuals from 18 mainland Indian populations along with 207 CS Asian and 235 E-Asian individuals of HGDP**. A model with four ancestral components (K=4) was the most parsimonious to explain the variation and similarities of the genome-wide genotype data. The CS-Asians are similar to ANI-major populations and E-Asians are similar to ATB major population indicating common ancestry for the respective populations before subdividing into the population identities that we see today. It also clearly shows that the AAA and the ASI cannot be readily identified with any of these global population groups. Population labels were added only after each individual's ancestry had

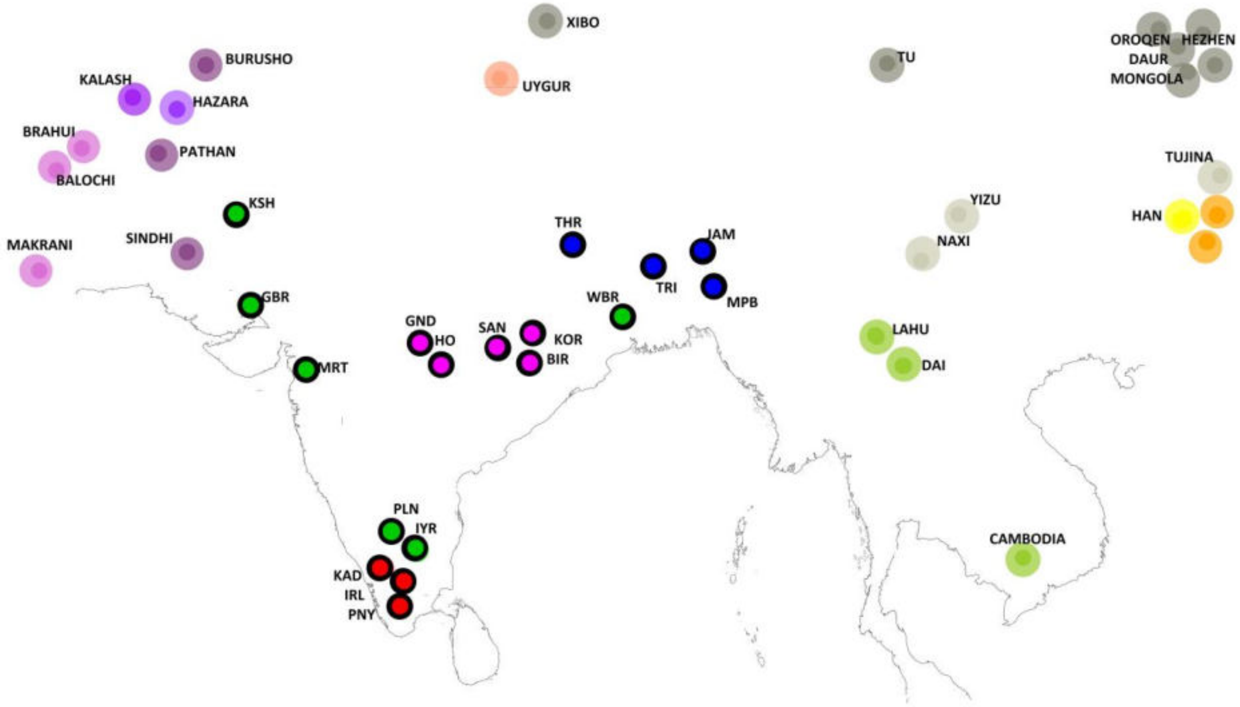
been estimated. The colors correspond to the colors used to encircle clusters of individuals in Fig. 2A. [also mainland Indians in Fig. S2, S3 and S7A]

Supplementary Text:

Analysis of 18 Mainland Indian Populations combined with the Central-South Asian and East-Asian samples of HGDP.

We combined our data set of 18 mainland populations with the Central-South Asians (CS-Asians) and East Asians (E-Asians) from HGDP. The CS-Asians populations included are: Brahui, Balochi, Hazara, Makrani, Sindhi, Pathan, Kalash and Burusho. While the E-Asians populations included are: Han, Tujia, Yizu(Yi), Miaozu, Oroqen, Daur, Mongola, Hezhen, Xibo, Dai, Lahu, She, Naxi, Tu, Yakut, Japanese, Combodian. Uygur who are admixed between CS-Asians and E-Asians are also included

Fig. Supplement : Approximate sampling location and population names



Definition of the Groups:

The Li et al paper(22) (Supplementary material 2.2 and figure S2 B), has subdivided the CS-Asians into 2 major groups: Group-1consisting Brahui, Makrani, Balochi and Group-2 consisting Burusho, Pathan and Sindhi. They identified Hazara and Kalash as outlier populations.

The detailed analysis of the HGDP E-Asians (Supplementary material 2.2 and figure S2 B) shows that populations in E-Asia also have multiple subgroups. We have defined:

E-Asian-group 1: Populations with high 'northern' ancestry include Mongola, Oroqen, Hezhen, Daur, Tu, Xibo, and Japanese. These groups reside in high latitude areas and speak languages of the Altaic family.

E-Asian-group 2: In contrast to E-Asian-group-1, populations like Dai, Lahu and Cambodian, who live in or near southwestern China have the lowest northern ancestry.

The Han and northern Han Chinese can be distinguished, although the former is most likely a mixture of southern and more central individuals.

E-Asian group 3: The Naxi and Yi are from the Yunnan Province in Southwest China, also have high northern ancestry, possibly due to their shared ancestry with the nomadic Qiang, an ethnic group from the Tibetan plateau.

E-Asian group 4: Other southern populations to the east (She and Miao)

Yakut as a separate group because it is an admixed population.

Tracing the ANI and the ATB ancestries

We have followed the above definition. In Fig. 3 (main text) PC-1 represents the systematic variation broadly separating the CS-Asian ancestry from E-Asian ancestry (Fig Supplement above shows the approximate positions on the map from where the populations were sampled), whereas the PC-2 represents the systematic variation broadly between the AAA + ASI ancestry and others.

In Fig. 3 we have broadly recapitulated the findings of Li et al(22). The Hazara and the Kalash are isolated clustered populations in the scatter of PC-1 versus PC-2. There is a thin line of separation between CSA-group-1 and group-2, with group-2 slightly closer to E-Asians. The ANI-major populations of India, particularly KSH which has ~97% ANI ancestry is inseparable from the CS-Asian group-2. Similarly, the JAM and TRI who have more than 95% ATB ancestry are inseparable from E-Asian-group 2. This identifies the origin of the ANI and the ATB ancestries with other major ancestries of the world, thus emphasizing the possible migration corridors through NE and NW India.

The proportion of variation explained by PC-1 (5.16%) and PC-2 (3.62%) in Figure 3, are both large compared to population data with 630918 markers. This indicates that the systematic variation separating the (AAA + ASI) ancestry from others is large, and thus the origin of these ancestries remain not well understood.

Our inferences inform that (i) four ancestral populations arrived in India with the ANI major populations probably using the NW corridor and the ATB major populations using the NE corridor (ii) after their arrival there was considerable admixture among them (iii) endogamy was abruptly established about 1600 years ago, and (iv) the practice of endogamy has been strictly followed resulting in strong ethnic sub-structuring that is evident even to this day.

Supplementary Information 4 (SI-4)

We show the genetic similarity of the Andaman Island populations (Jarawa and Onge) with Papuan and Melanesian populations of HGDP.

Pages 29, 30

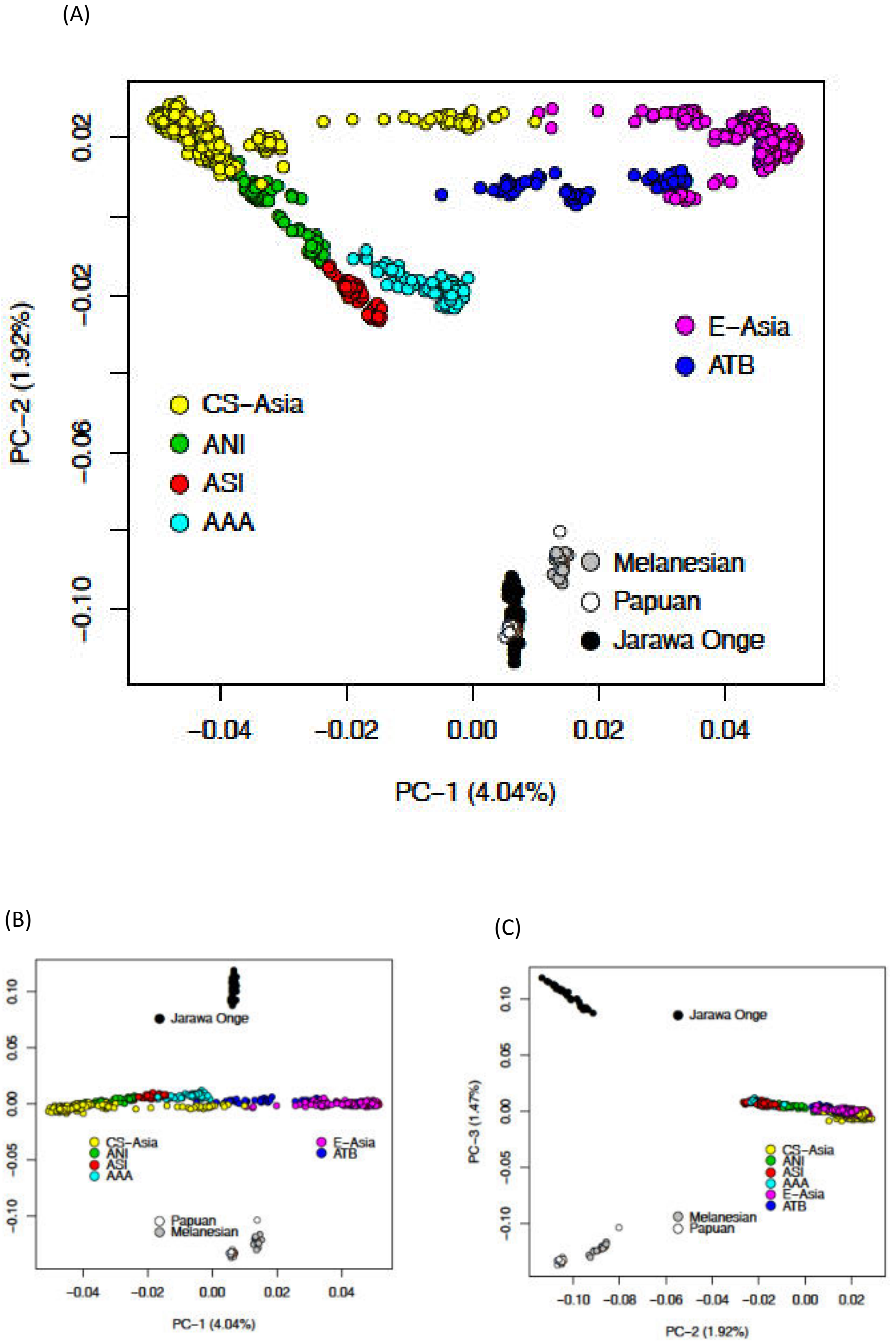
The joint analysis of the 20 Indian populations (18 mainland Indian + 2 Island population (JWA and ONG) along with the CS-Asians, E-Asians and Oceania population of HGDP reveal that the Island ancestry of JWA and ONG, which is clearly distinct from all ancestries found in mainland India is indeed also different from CS-Asians and E-Asians but is very similar to the Oceania ancestry.

The PC-1 versus PC-2 scatter plot reveals that the Oceania populations of HGDP, especially the Papuans are close to the JWA and ONG (Supplementary Figure 4.1). The variation explained by PC-1 and PC-2 are both high. However, the JWA and ONG separate from the Oceania population along PC-3 (Supplementary Figure 4.2 and Supplementary Figure 4.3). This indicates that the genetic difference between island populations + Oceania population is large compared to the 4 mainland Indian population clusters as well as the CS-Asia and E-Asians **also** it establishes that the genetic difference between the Island populations and Oceania (archipelago ancestry) is small compared to that between the other ancestries and this archipelago.

Fig. S8 A: The PCA plot (PC-1 versus PC-2) of JWA and ONG along with mainland Indians and CS-Asians, E-Asians and Oceania populations of HGDP. It shows clustering of the JWA and ONG populations with Oceania population of HGDP.

Fig. S8 B: The PC-1 versus PC-3 plot of JWA and ONG along with mainland Indians and CS-Asians, E-Asians and Oceania populations of HGDP. It shows separation of the JWA and ONG populations.

Fig. S8 C: The PC-2 versus PC-3 plot of JWA and ONG along with mainland Indians and CS-Asians, E-Asians and Oceania populations of HGDP. It shows separation of the JWA and ONG populations with Oceania population of HGDP


(A)
PC-2 (1.92%)
0.02
-0.02
-0.06
-0.10
-0.04
-0.02
0.00
0.02
0.04
PC-1 (4.04%)
E-Asia
ATB
CS-Asia
ANI
ASI
AAA
Melanesian
Papuan
Jarawa Onge
(B)
PC-3 (1.47%)
0.10
0.05
0.00
-0.05
-0.10
Jarawa Onge
CS-Asia
ANI
ASI
AAA
E-Asia
ATB
Papuan
Melanesian
-0.04
-0.02
0.00
0.02
0.04
PC-1 (4.04%)
(C)
PC-3 (1.47%)
0.10
0.05
0.00
-0.05
-0.10
Jarawa Onge
CS-Asia
ANI
ASI
AAA
E-Asia
ATB
Melanesian
Papuan
-0.10
-0.08
-0.06
-0.04
-0.02
0.00
0.02
PC-2 (1.92%)

Supplementary Information 5 (SI-5)

In **SI-5** we show

(1) The Ancestral Chromosomal Block Length (ACBL) distribution fits to the theoretical exponential distribution.

Fig. Supplementary A: The distribution of ACSL pertaining to ASI, AAA and ATB, and the fitted exponential distribution among GBR, WBR and IYR population. (The Kolmogorov-Smirnov Test was performed to check the equality of the distribution of ACSL and the fitted exponential)

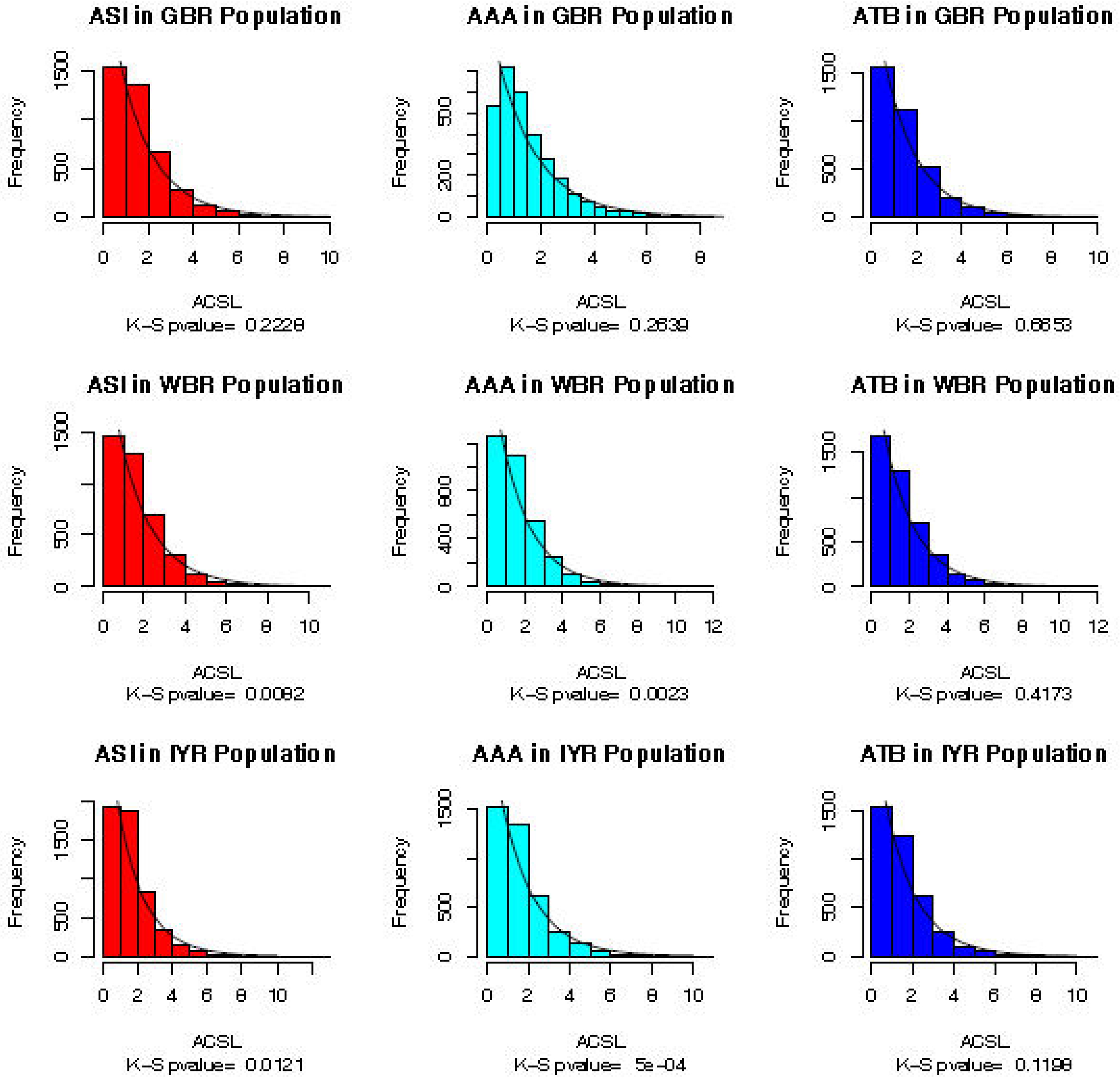


Fig. Supplementary B: The distribution of ACSL pertaining to ASI, AAA and ATB, and the fitted exponential distribution among MRT and PLN population. (The Kolmogorov-Smirnov Test was performed to check the equality of the distribution of ACSL and the fitted exponential)

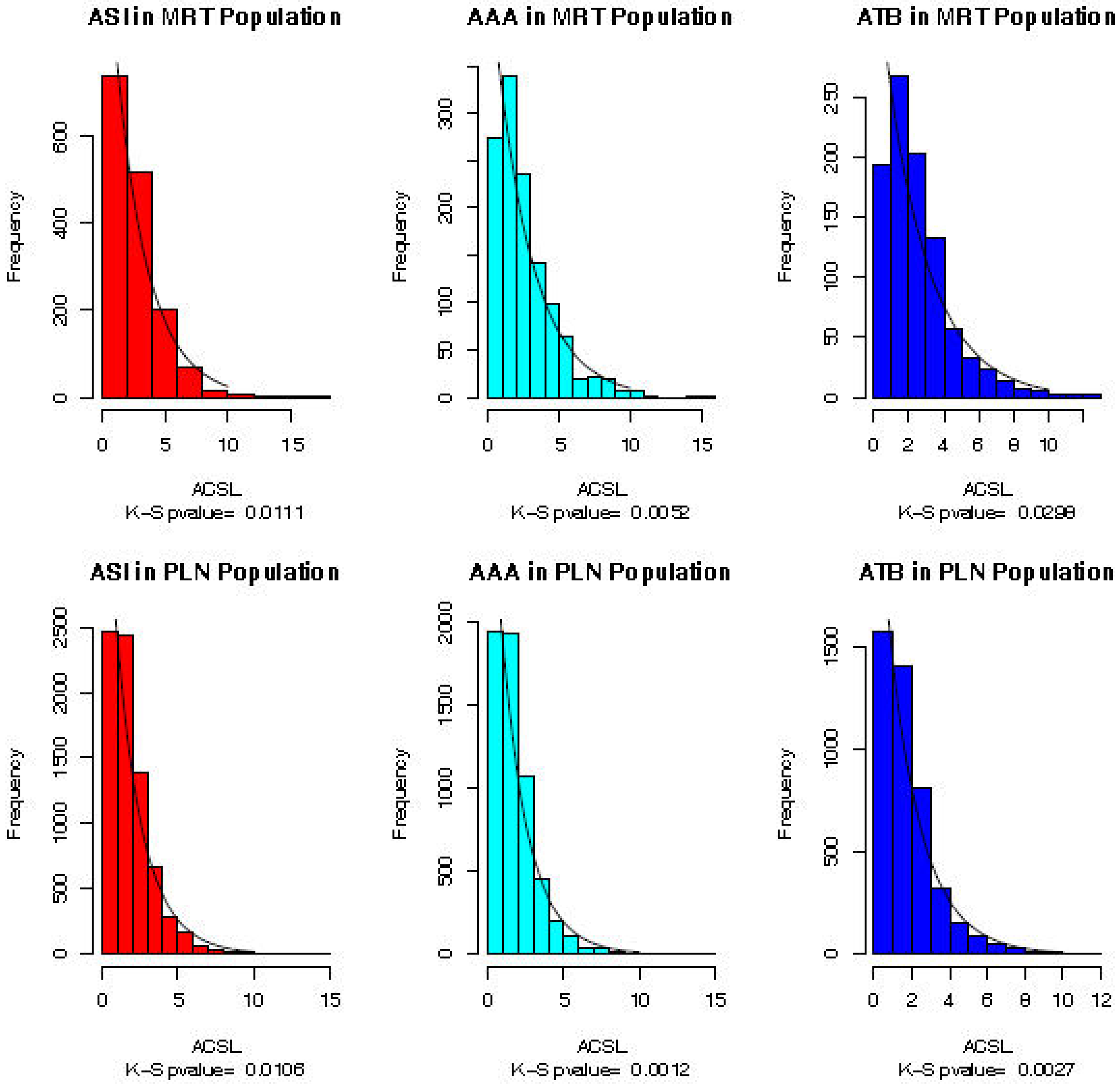


Fig. Supplementary C: The distribution of ACSL pertaining to ANI, AAA and ATB, and the fitted exponential distribution among KDR and IRL population. (The Kolmogorov-Smirnov Test was performed to check the equality of the distribution of ACSL and the fitted exponential)

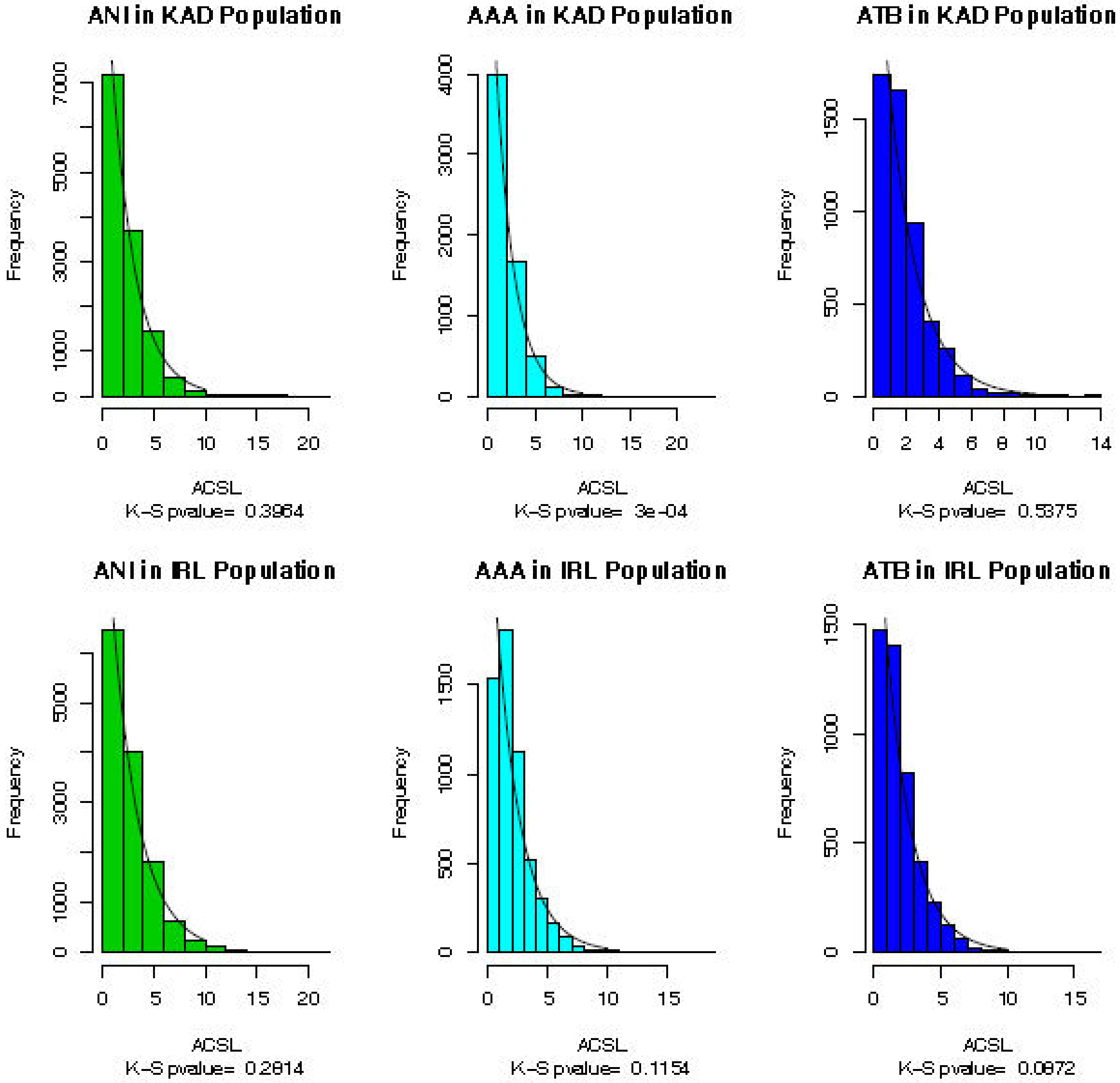


Fig. Supplementary D: The distribution of ACSL pertaining to ANI, ASI and ATB, and the fitted exponential distribution among GND, HO, SAN and KOR population. (The Kolmogorov-

Smirnov Test was performed to check the equality of the distribution of ACSL and the fitted exponential)

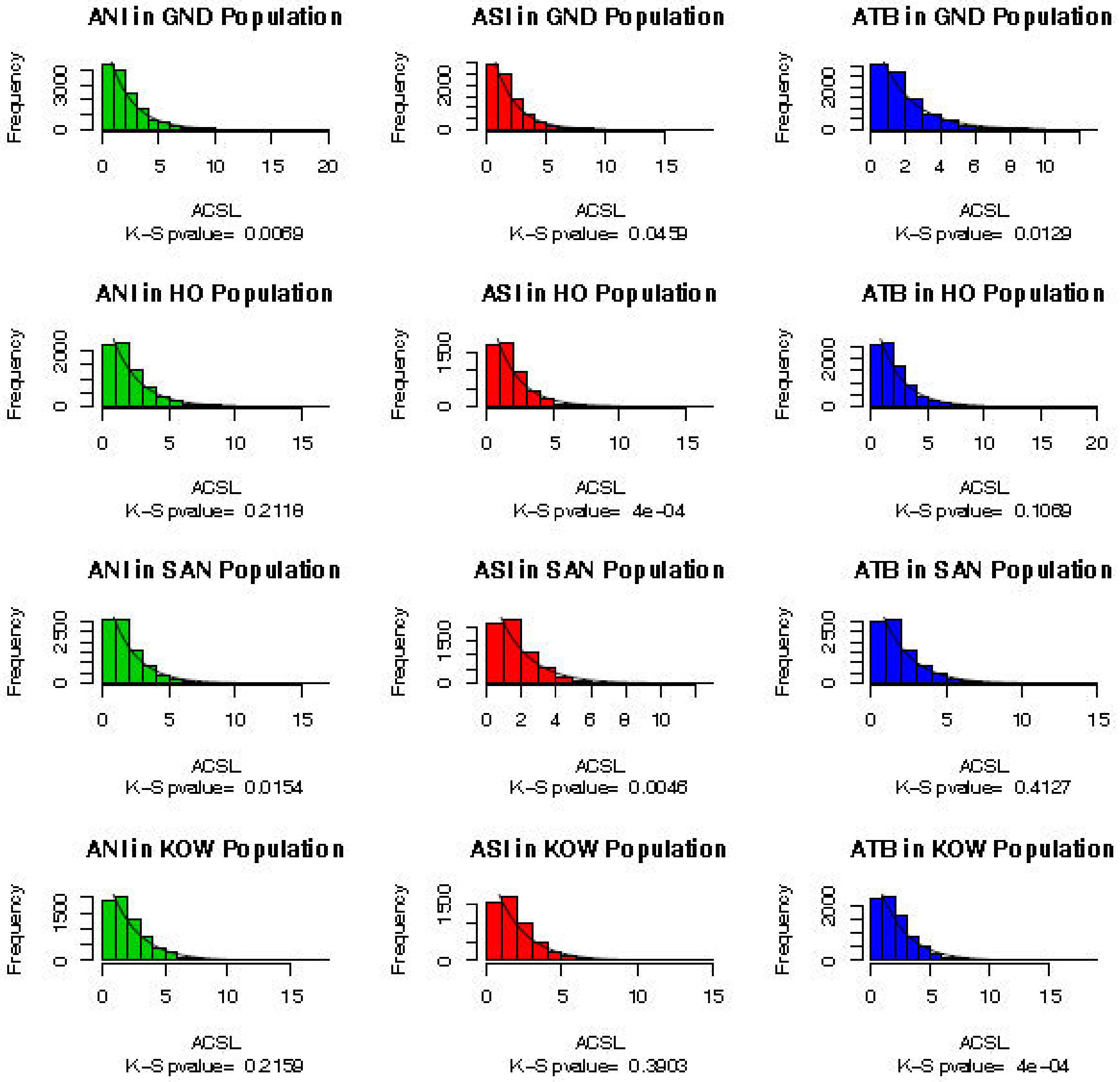
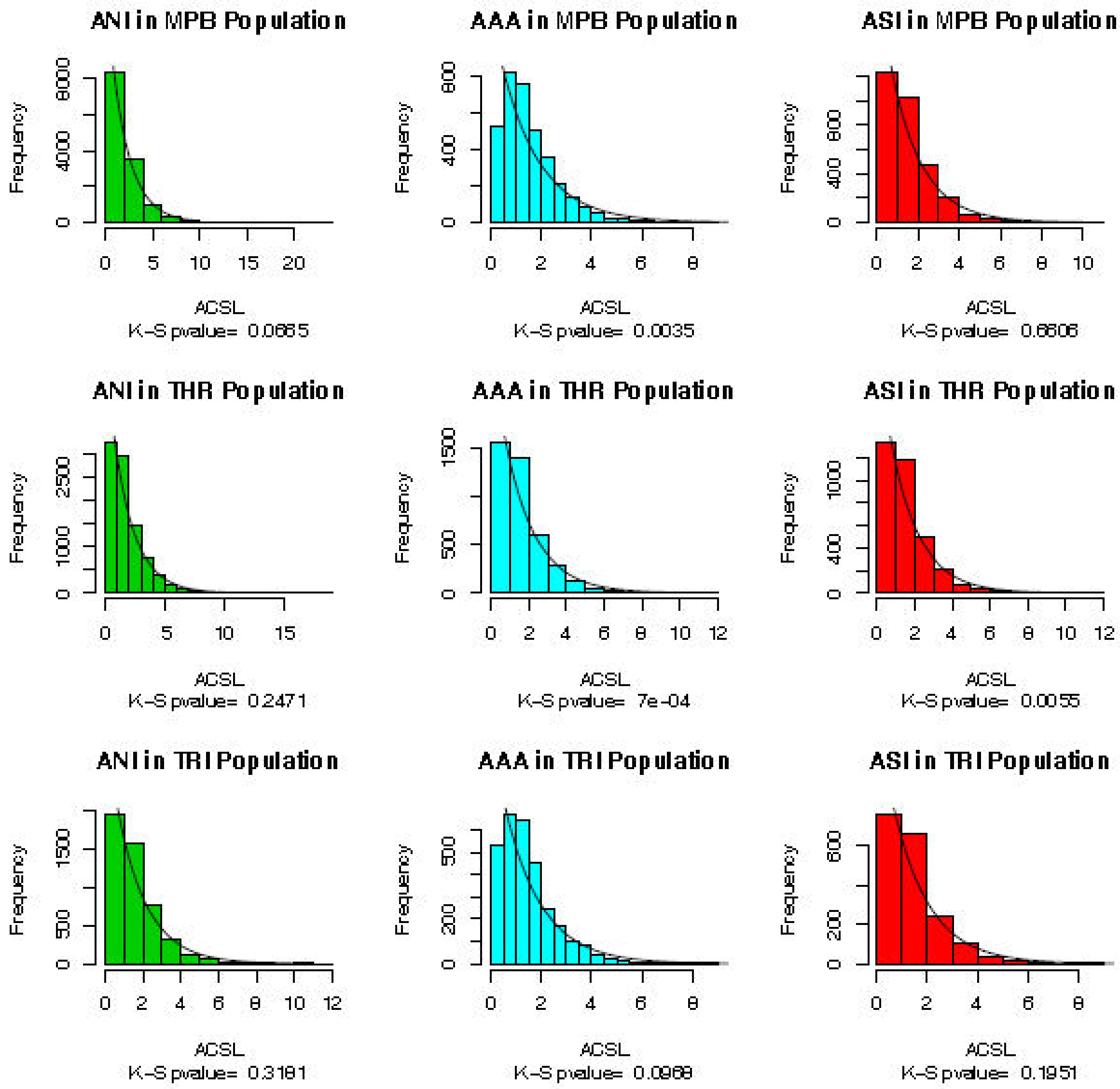


Fig. Supplementary E: The distribution of ACSL pertaining to ANI, ASI and AAA, and the fitted exponential distribution among MPB, THR and TRI population. (The Kolmogorov-Smirnov Test was performed to check the equality of the distribution of ACSL and the fitted exponential)



Supplementary Information 6 (SI-6)
SI-6 is the detailed methods section:

DNA microarray analysis and data curation:

The study was originally planned with 20 individuals from 20 populations. Individuals with genotype calls at <90% of markers were eliminated. Two individual from the Birhor (BIR) and one individual each from the populations Korwa (KOR), Onge (ONG) and Ho were excluded because of relatedness closer to second cousin, inferred by high IBD. While choosing between a relative pair thus identified, we have retained the individual with higher proportion of genotype calls. Markers with minor allele frequency <5% in one or more populations or those that deviated from HWE (p<0.001) were excluded. The final data set comprised data on 367 individuals and 803570 markers.

X-Chromosome haplotyping:

As mentioned in the main text, females in the samples were identified using X-chromosome data. In order to infer the X-chromosome haplotypes for each female individual we used Shapeit2 (30,31).

Sex-Bias in admixture:

Sex bias in ancestry contributions was evaluated by selecting only females (to ensure we compare a diploid X chromosome to diploid autosomes), and running ADMIXTURE with K= 4 on the X chromosome and autosomes separately. The Wilcoxon signed rank test, a non-parametric version of the paired Student's t-test that does not require the normality assumption, was applied to assess the significance of the difference in X and autosomal ancestry proportions.

The distance matrix for the phylogenic tree on the phased X-chromosome haplotype was generated using the ‘complete linkage’ method in hierarchical cluster analysis. The clustering and the dendrogram plotting was done using R 2.12.2 (https://www.**r**-project.org/).

Distribution of ancestral block lengths (ABL)

For all the 16 admixed mainland populations (except the Khatri (KSH), Paniya (PNY), Birhor (BIR) and Jamatia (JAM) which were used as reference, ancestral block segments were inferred for each individual haplotype. We calculated the mean and variance from the distribution of the observed ABLs belonging to each of the 3 ancestral components, except the major one, within a population. That was then compared with an exponential distribution with the same mean. We used the non-parametric Kolmogrov-Smirnov test to compare the distributions.