Journal of Human Genetics (2009) 54, 47–55

www.nature.com/jhg

ORIGINAL ARTICLE

The Indian origin of paternal haplogroup R1a1* substantiates the autochthonous origin of Brahmins and the caste system

Swarkar Sharma[1,2,4], Ekta Rai[1,2,4], Prithviraj Sharma[1,3], Mamata Jena[1], Shweta Singh[1], Katayoon Darvishi[1], Audesh K Bhat[1], AJS Bhanwer[2], Pramod Kumar Tiwari[3] and Rameshwar NK Bamezai[1]

Many major rival models of the origin of the Hindu caste system co-exist despite extensive studies, each with associated genetic evidences. One of the major factors that has still kept the origin of the Indian caste system obscure is the unresolved question of the origin of Y-haplogroup R1a1*, at times associated with a male-mediated major genetic influx from Central Asia or Eurasia, which has contributed to the higher castes in India. Y-haplogroup R1a1* has a widespread distribution and high frequency across Eurasia, Central Asia and the Indian subcontinent, with scanty reports of its ancestral (R*, R1* and R1a*) and derived lineages (R1a1a, R1a1b and R1a1c). To resolve these issues, we screened 621 Y-chromosomes (of Brahmins occupying the upper-most caste position and schedule castes/tribals occupying the lower-most positions) with 55 Y-chromosomal binary markers and seven Y-microsatellite markers and compiled an extensive dataset of 2809 Y-chromosomes (681 Brahmins, and 2128 tribals and schedule castes) for conclusions. A peculiar observation of the highest frequency (up to 72.22%) of Y-haplogroup R1a1* in Brahmins hinted at its presence as a founder lineage for this caste group. Further, observation of R1a1* in different tribal population groups, existence of Y-haplogroup R1a* in ancestors and extended phylogenetic analyses of the pooled dataset of 530 Indians, 224 Pakistanis and 276 Central Asians and Eurasians bearing the R1a1* haplogroup supported the autochthonous origin of R1a1 lineage in India and a tribal link to Indian Brahmins. However, it is important to discover novel Y-chromosomal binary marker(s) for a higher resolution of R1a1* and confirm the present conclusions.
Journal of Human Genetics (2009) **54,** 47–55; doi:10.1038/jhg.2008.2; published online 9 January 2009

Keywords: caste system; Indian population; R1a1; Y-haplogroup

INTRODUCTION

India comprises one of the largest ethnic populations with enormous cultural, morphological and genetic diversity,[1–4] which linguistically belong to Austro-Asiatic (AA), Dravidian (DR), Tibeto-Burman (TB) and Indo-European (IE) families.[5] Indian populations are culturally stratified as tribals and non-tribals.[6] Tribals constitute 8.08% of the total population[7,8] and the majority of them speak languages belonging to AA, DR and TB families;[8] also, most of them are believed to be autochthones of India.[9] On the contrary, most of the contemporary non-tribal populations belong to Hindu religion and speak languages of IE and DR descent. In addition, there are several other religious communities contributing a fraction to the total Indian population structure.[6] The Hindu caste system has played a major role in the social and economic organization of India[10] and is constituted by four major classes (varna)—namely, Brahmin (priestly class), Kshatriya (warrior class), Vyasa (business class) and Shudra (menial labor class).[6] The fifth class, 'Panchama' (standing for tribals), was added at a later date, giving them the lowest rank.[11] The co-existence of IE tribes and DR castes indicates a complex historical interaction and suggests no 'one to one correlation' between language and this social organization.[12] In spite of the consensus on the relatively uniform maternal gene pool of Indian populations and the large efforts through many philological,[13,14] archaeological[15,16] and recent molecular genetic approaches to elucidate rival models,[6,7,9,11,12,17–21] the history and concepts of the origin of the caste system are still controversial and unclear. The competing main models (the first of them based on shared IE languages) suggest that contemporary Hindu Indians are descendants of primarily West Eurasians who migrated from the Near east, Antolia and the Caucasus 3000–8000 years ago,[13,14] which has been supported by the demic diffusion model[1,22] and validated by molecular genetic data.[7,11] The second model, based on molecular genetic data, mainly the Y-chromosomal M17 marker

[1]National Centre of Applied Human Genetics, School of Life Sciences, Jawaharlal Nehru University, New Delhi, India; [2]Department of Human Genetics, Guru Nanak Dev University, Amritsar, India and [3]Centre for Genomics, School of Studies in Zoology, Jiwaji University, Gwalior, India
Correspondence: Professor RNK Bamezai, National Centre of Applied Human Genetics, School of Life Sciences, Jawaharlal Nehru University, New Delhi 110067, India.
E-mail: bamezai@hotmail.com and bamezai@ncahg.org
[4]These authors contributed equally to this work.
Received 17 August 2008; revised 30 October 2008; accepted 6 November 2008; published online 9 January 2009

(R1a1 haplogroup), suggests the migration of IE people from Central Asia to India.[23] Another model suggests that later on 'not alone but a package' of Y-haplogroups migrated from Central Asia, introducing the caste system to India.[7] Yet another model suggests the late Pleistocene heritage of tribal and caste populations, with limited recent gene flow between them[24] and the largely South Asian origin of Indian caste communities, indicating no major genetic influx either with the development of agriculture or with the spread of the Indo-Aryan (IE) language family.[12] It has also been suggested that there was a minor influence from Central Asia and the pre-Holocene and Holocene era, not Indo-European expansions, which shaped the pre-existing South Asian gene pool.[9] Alternatively, another recent study[21] has suggested that distinct paternal distribution patterns exist among caste and tribal populations and tribals have contributed to the lower caste groups (schedule castes) as well as expansions and establishment of Indo-European populations as upper castes. All these proposed hypotheses make the question of the origin of the caste system and the relationship among these hazy and obscure.

This study was designed to evaluate these competing hypotheses of the origin of the caste system by taking into account the information available in the literature about the cultural practices of the Hindu caste system being endogamous.[2,11,25–27] To test the concepts of Central Asian introduction of the Indian caste system[7] by Indo-Aryans, who plausibly and predominantly appointed themselves to castes of higher rank to legitimize and maintain their power on land, labor and resources;[14] and to test the rank-related West Eurasian admixture,[11,21] we chose the Brahmin class, occupying similar socio-economic upper-most caste positions, and the schedule castes and tribals, occupying the lower-most positions in the Indian caste hierarchy, irrespective of linguistic and geographic affiliations within India, for an ideal comparative study.

Here, we report our analyses based on Y-chromosomal data of 621 Brahmins and schedule caste/tribal samples and its extension with the compiled data of Brahmin, scheduled caste/tribal populations from published sources. The total dataset represents 2809 Y-chromosomes (767 Brahmins, 2042 scheduled caste/tribal samples) constituting an extensive dataset of Brahmins and tribals (including 621 samples from this study). We also attempted to assess the affinities among Brahmins from different regions speaking different languages, and evaluated the hypothesis of large migration of IE people and introduction of the caste system to India with the purpose of elucidating their genetic relationship with other Indian and worldwide populations, using the data available in the literature. Further, taking into consideration the recent study[28] that found a high level of male genetic substructure as a result of the founder effect and social stratification among the Brahmins and Kshatriyas of Jaunpur district, we also explored the probability of any such phenomenon or other genetic patterns in other regions of India.

MATERIALS AND METHODS

Samples analyzed

A total of 621 Y-chromosomes (367 Brahmin samples and 254 scheduled caste/tribal samples (the term 'schedule castes' is used instead of 'lower caste')) were screened (Table 1). To substantiate our data and compare our results, we further compiled the data available in the literature for Indian populations (400 Brahmin Y-chromosomes as well as 1788 scheduled caste/tribal Y-chromosomes).[7,9,12,21,23,24,29] as well as for other population groups, mainly from Central Asia, Eurasia and Southeast Asia.[9,23,30–34]

Sample collection and DNA isolation

About 2 ml of venous blood was collected from each individual, with their informed written consent and the approval of Jawaharlal Nehru University ethical committee. DNA was isolated from the samples by the standard protocol used in the laboratory.

Screening of Y-chromosomal polymorphisms

Fifty-five binary markers known to dissect out paternal haplogroups (HGs) at high resolution were selected (Supplementary Figure 1) following the hierarchy of the Y-chromosome phylogeny[35] (www.isogg.org) and genotyped, using either direct Sequencing (ABI 3100 Sequencer, USA) or RFLP analysis. Most of the binary markers have been described in the literature.[9,35,36] A novel informative polymorphism (ss4bp, rs41352448) defining a newly found Y-haplogroup Q sublineage (Q5), presently restricted to India,[37] was also typed. These samples were also genotyped for seven Y-microsatellite markers (Y-STRs): DYS 389I, DYS 389II, DYS390, DYS391, DYS392, DYS19/394 and DYS448, to estimate the haplotype variation within the HG defined by binary markers (Y-STR haplotypes data are available through personal communication). Although seven Y-STR markers were analyzed, to keep uniformity in evaluation with the reported literature the data of only six overlapping Y-STR markers (DYS 389I, DYS 389II, DYS390, DYS391, DYS392, DYS19/394) were used for the analyses.

Statistical and phylogenetic analyses

Y-chromosomal haplogroup frequencies were calculated by the simple gene count method and various frequency charts were prepared in Microsoft Excel. Tests of significance for comparisons of HGs among different population groups were done using contingency χ^2 tests, with Yates's correction when required.

Both pairwise genetic distances for different population groups using Y-chromosome data (computed as a linearization of FST/(1−FST))[38] and analysis of molecular variance (AMOVA) were computed using ARLEQUIN 3.01 software.[39] All the values were tested for significance with 10 000 permutations. Linearized distance values were used for the formation of MDS plots based on Y-STR data using SPSS 10.0.5 (Chicago, IL, USA) for the different population groups. Median joining (MJ) networks[40] for Y-STR haplotypes within specific Y-haplogroups were constructed by the software NETWORK 4.1.0.9.

Admixture proportions were calculated by using ADMIX2 software.[41] ADMIX2 uses an algorithm to estimate the admixture proportions by dividing the whole genetic composition of a hybrid between given parental populations. In case source populations are genetically identical (that is, of same allelic composition), the software logs error messages. The statistical significance of the observations could be increased by bootstrap measures (refer ADMIX2 manual). Nested cladistic analysis[42] was performed using the program GEODIS 2.4[43] and inferences were made using the latest keys provided. A nested clade phylogeographic analysis attempts to sort out the roles of recurrent forces (such as gene flow) from historical events (such as fragmentation or range expansion events) by overlaying the geographical distributions of haplotypes upon the evolutionary tree of the haplotypes. Methodologically, it involves repeated connection of haplotypes, different only by one mutational step, into a single category spanning the entire original haplotype network. (Please refer GEODIS documentation for an example.)

Proportions of shared alleles (PSA) between different regions were calculated using PsaCALC, software and 1−PSA as a measure of distance.[44] The spatial frequency/diversity maps were generated by the Kriging procedure[45] using SURFER version 8.0 (Golden Software Inc., CO, USA). Spearman's rank correlation coefficient, which has values +1 and −1 for perfect positive and negative correlations, respectively,[46] was calculated for the latitude and longitude with haplogroup frequency and diversity by applying the formula in Microsoft Excel. To estimate the age of various haplogroups in the Indian subcontinent, we used the mean variance of the microsatellite repeats in dating.[47] The Y-chromosome microsatellite diversity and variance in repeat number were calculated using the software package MICROSAT version 1.5. The latest acknowledged Y-STR phylogenetic mutation rate (μ) of 6.9×10^{-4} (0.069%) per generation and 95% CI (9.5×10^{-4}–4.3×10^{-4})[48] were taken into account, with the frequently used human generation time (g) of 25 years. We also applied g=32 in calculating the age estimates, keeping in mind the recent compelling evidences that the age of the male generation in modern times is closer to 32 or 35 years.[49,50]

Table 1 Y-haplogroups percentage distribution in studied regional population groups of India

			Haplogroups																	
Population	*Status*	N	*C5*	*E*	*F**	*G*	*H**	*H1*	*J2*	*K*/K2*	*L*	*N*	*O*	*P**	*Q (XQ5)*	*Q5*	*R**	*R1**	*R1a**	*R1a1*	*R2*
North:																					
J&K Kashmiri Pandits	Br	51	1.96		3.92	1.96		9.80	9.80	9.80	5.88				5.88		1.96	11.76	3.92	19.61	13.73
J&K Kashmir Gujars	Tr	49	2.04		4.08			10.20	6.12	8.16	16.33				2.0			2.04		40.86	8.16
Uttar Pradesh Brahmin	Br	31						16.13	3.23		3.21				3.23	3.23				67.74	3.23
Punjab Brahmin	Br	49	3.58		3.57	3.57			21.43		7.14									35.71	25.00
Himachal Brahmin	Br	30	5.26		15.79			10.53	5.26		5.26							5.26		47.37	5.26
Central:																					
Uttar Pradesh (South) Kols	Tr	30						11.11	33.34				40.74							14.81	
Uttar Pradesh (South) gonds	Tr	38						59.46	18.92	10.81		2.7	8.11								
Madhya Pradesh Brahmins	Br	42			2.38			7.14	23.81		7.14	2.38		2.38	2.38	2.38				38.10	11.90
Madhya Pradesh Gonds	Tr	31					6.25	56.25		6.25			6.25		6.25					18.75	
Madhya Pradesh Saharia	Tr	57			5.08		10.29	23.4			3.24				1.75				22.8	28.07	5.37
East:																					
Bihar Brahmins	Br	38	2.63						2.63	5.26	13.16				2.63	2.63		5.26		60.53	5.26
Bihar Paswan	SC	27			3.70	11.11	11.11	3.70	3.70		7.41		3.70				3.70	11.11		40.74	
West Bengal Brahmins	Br	30						5.56												72.22	22.22
West:																					
Maharashtra Brahmins	Br	32	3.33			3.33	3.33	6.67	16.67	3.33	10.0	3.33	3.33					0.0		43.33	3.33
Gujarat Bhils	Tr	22	9.09				18.18	9.09	18.18		18.18									9.09	18.18
Gujarat Brahmins	Br	64	3.33	3.33		10.94		1.56	15.63	3.13	7.81	3.13						9.38		32.81	9.38
Total		621																			

Abbreviations: Br, Brahmins; Tr, tribe; SC, Scheduled caste.

RESULTS AND DISCUSSION

Genetic structure of the studied regional population groups

We observed a total of 19 Y-haplogroups in our analyzed dataset of 621 Y-chromosomes (Table 1) defined by 31 informative polymorphisms out of 55 genotyped binary polymorphisms (Supplementary Figure 1). It has been argued in the literature that the Indian higher caste groups show relatively small genetic distances when compared with the West Eurasians,[11] linking this to hypothetical migrations by Indo-Aryan speakers. Further, M17-R1a (presently designated as R1a1) was suggested as a potential marker with decreasing frequencies from Central Asia towards South India.[23] On similar lines, it was suggested that a package of Y-HGs (J2, R1a, R2 and L) was associated with the migration of Indo-European people from Central Asia.[7] Although our study observed a high frequency of Y-HGs, R1a1, J*/J2, R2 and L, it was not exclusively restricted to any region or population (Table 1). Moreover, most of the population groups from the studied regions showed a less frequency of the highly frequent haplogroups of Central Asia: C3, DE, I, G, J*, N and O, except for some population-specific distributions. Y-haplogroup G was observed to be present at high frequency in Gujarat Brahmins and Bihar Paswans, whereas Y-haplogroup O was more frequent in Uttar Pradesh Kols and Gonds (Table 1). In case of a recent gene flow (associated with the migration of Indo-European people), we expected the more frequent Central Asian Y-haplogroups (C3, DE, I, G, J*, N and O)[12] to be present at least at similar frequencies, as observed for R1a1* in Northwest India, which, however, was not the case in this study.

Comparison of Brahmins and scheduled castes/tribals

To explore further, we analyzed the dataset (consisting of 510 Y-chromosomes), which could be classified as Brahmins (n=256) and scheduled castes/tribes (n=254) from the studied six regions of India (Jammu and Kashmir, Uttar Pradesh, Bihar, Madhya Pradesh, Maharastra and Gujarat) to evaluate regional distribution patterns between these two extreme end population groups of the Hindu caste hierarchy (Supplementary Table 1), where intermixing due to marriages has been absent because of social unacceptability. AMOVA showed no variation between different geographical regions (−2.3%), some variation between populations within regions (12.67%) and most of the variation within populations (89.63%). The percentage distribution of haplogroups (Supplementary Table 1) in Brahmins (n=256) showed a total of six most frequent (percentage >5%) haplogroups: R1a1* (40.63%), J2 (12.5%), R2 (8.59%), L (7.81%), H1 (6.25%) and R1* (5.47%), contributing to 81.25% of the total distribution in Brahmins. Tribals and scheduled castes (n=254) also showed six haplogroups: H1 (31.10%), R1a1* (20.47%), J2 (10.24%), L (7.87%), H* (7.87%) and O (6.69%), contributing in total to 84.25%. Interestingly, four of the haplogroups were overlapping in percentage (>5%) distribution with Brahmins. The haplogroup diversity and s.d. in each population are also given in Supplementary Table 1.

Study of the compiled dataset

The pooled percentage distribution of Y-haplogroups in the overall dataset of 2809 Y-chromosomes (767 Brahmins, 674 schedule castes and 1368 tribals) is summarized in Supplementary Table 2. All together (Brahmins, schedule castes and tribals), 22 Y-haplogroups were observed. The percentages of seven of these haplogroups (with percentage >5%) accounted for 85.5% of the total number of Y-chromosomes (n=2809). The haplogroups with their percentages in descending order were: R1a1* (21.1%), H1 (19.1%), R2 (10.5%), O (10.1%), L (9.5%), J*/J2 (8.3%) and F* (6.9%). These haplogroups remained the most frequent haplogroups even after the distribution of Y-chromosomes within respective groups of Brahmins, schedule castes and tribals, but with significant percentage differences (Supplementary Table 2). Five haplogroups out of 18 were found to be most frequent (>5%) in Brahmins (R1a1* (35.7%), J*/J2 (12.4%), L (11.3%), R2 (10.8%) and H1 (8.0%)) and represented 78.2% of the total number of samples (n=767), whereas haplogroup O was found to be very less frequent (0.7%) in Brahmin Y-chromosomes. Seven out of 14 haplogroups (with percentage >5%) (H1 (24.2%), R1a1* (17.2%), R2 (14.2%), L (12.2%), F* (9.8%), J*/J2 (6.4%) and K* (5.3%)) represented 89.3% of the total number of Dalit Y-chromosomes (n=674). Tribal Y-chromosomes represented by seven out of 20 haplogroups displayed percentages >5%: O (25.5%), H1 (25.3%), R1a1* (10.2%), F* (7.5%), R2 (6.4%), J*/J2 (6.1%) and L (5%) (86% of the total number of samples (n=1368)). All other observed haplogroups had their percentages <5% (Supplementary Table 2). The study was further extended, dividing the samples into four main linguistic categories (Indo-European (IE), Dravidian (DR), Tibeto-Burman (TB) and Austro-Asiatic (AA)) present in India as well as five regional categories (Central, East, North, South and West India). Y-haplogroup distributions as per these categories are presented in Supplementary Figures 3a and b. AMOVA was also done using the compiled dataset and by characterizing the populations into social, geographical and linguistic groups (Table 2). Geographical regions showed very less variation (0.79%) among the groups but higher variation between populations within groups (16.94%). In contrast, linguistic groups showed higher variation among the groups (15.56%) but lower variation between populations within linguistic groups (6.15%). Interestingly, when the TB linguistic group was removed from the analysis, the percentage variation among the

Table 2 AMOVA results based on compiled dataset

Group	*Total variation (%)*		
	Within populations	*Between populations (within groups)*	*Between groups*
Social structure	80.06	11.70	8.24
Geographical region	82.27	16.94	0.79
Linguistic group	78.29	6.15	15.56
Linguistic group[a]	84.41	6.16	9.43

Social structure: Brahmins, Scheduled Castes, Tribes.
Geographical regions: Central, East, North, South and West India.
Linguistic groups: Indo-European (IE), Dravidian (DR), Austro-Asiatic (AA) and Tibeto-Burman (TB).
[a]Without TB.

groups reduced (9.43%) but variation between populations remained almost the same (Table 2). It was observed that by either of the grouping most of the variation was within the population groups.

The overall observations, either by comparison of Brahmins and scheduled castes/tribes from different Indian regions (Supplementary Table 2) or from the analyses of the pooled dataset of Indian Y-chromosomes, showed no consistent pattern of the exclusive presence and distribution of Y-haplogroups to distinguish the higher-most caste, Brahmins, from the lower-most ones, schedule castes and tribals (Supplementary Figures 3a–c).

Origin of Y-haplogroup R1a1*

However, a peculiar trend in distribution of the highest frequency of Y-haplogroup R1a1* (Table 1) in Brahmins, H1 in tribals and schedule castes, and O in tribals was also observed. Whereas on the one hand a consensus has developed in the literature among all schools of thought in assigning Indian origin to haplogroup H1 and in the association of haplogroup O with either Austro-Asiatic or Tibeto-Burman tribals, the widespread geographic distribution of R1a1* and reasonably high frequency across Eurasia (Figure 1a), with scanty representation of its ancestral (R*, R1* and R1a*) and derived lineages (R1a1a, R1a1b and R1a1c) across the region, leaves obscure the question of origin of R1a1*. This becomes more complex with the claims[7,9,12,23] proposing a scenario of the recent major gene flow from Central Asia to India and the antagonistic observations[9,12] of its highest variance in India, suggesting the gene flow in opposite direction. Further, the observation of a very high frequency (upto 72.22%) in this study (Table 1) and in the literature (Supplementary Figures 3a and b) of this haplogroup in all of the Brahmins may indicate its presence as a founder lineage for this caste group (irrespective of the geographical and linguistic affiliation of Brahmins), thus making this haplogroup of extreme importance and a key haplogroup in answering the question of origin of caste systems in India.

Although the geographic origins of haplogroup expansions can be inferred from the frequencies, associated diversity[51] and clinal patterns of distribution, past inferences from literature indicate that such relations are not so simple to interpret. It is observed that regions of high frequency and high variance are not always the same. Regions with highest haplogroup frequencies are not always sites of its origin and clinal patterns are not obvious in binary HG frequency data;[33] also, the highly associated microsatellite variance, exclusively, may not always be an indicator of *in-situ* diversification and could result as a consequence of repeated gene flow from different sources[52,53] as observed by Y-chromosomal diversity in Central Asia.[32] This suggests that many analytical parameters should be included and potential causes of a wrong interpretation should be taken care of before reaching any conclusion.

All rival models of the origin of caste system were taken into consideration and results were analyzed to the highest Y-SNP marker

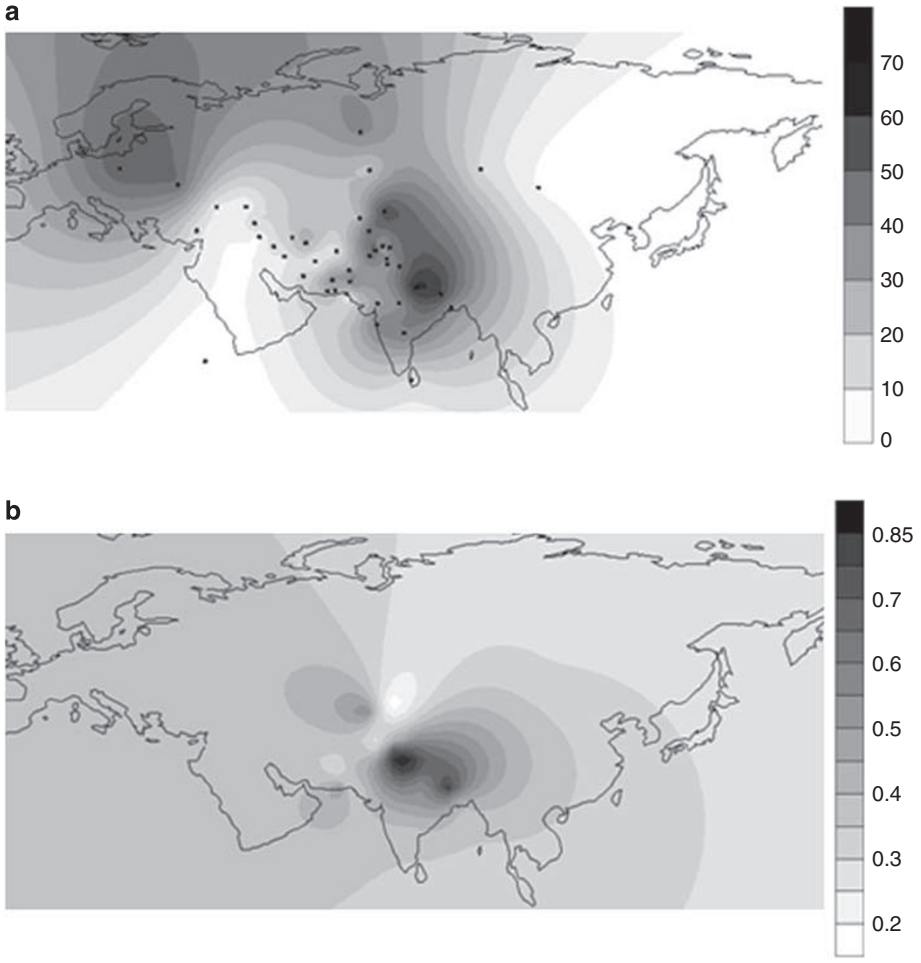

Figure 1 The spatial distribution maps of Y-haplogroup R1a1 generated by the Kriging procedure using SURFER version 8.0. (**a**) Spatial frequency distribution of Y-haplogroup R1a1* across Eurasia, Central Asia and the Indian subcontinent. (**b**) Spatial distribution of Y-haplogroup R1a1*-associated diversity based on microsatellite markers.

resolution for the R1a haplogroup (Supplementary Figure 1), in addition to adding data and information from the literature, making a pooled dataset of the R1a1* haplogroup containing ~1030 individuals (530 Indians, 224 Pakistanis, and 276 Central Asians and Eurasians) from around the Indian-subcontinent, Central Asia and Europe. Using different phylogenetic tools and parameters (mentioned in Materials and methods) and concentrating on the distribution of R1a1* and its ancestral lineages, the answer to the source and expansion of this haplogroup, across the globe, was explored.

Spatial frequency and molecular diversity distribution of R1a1* in Eurasia, Central Asia and the Indian subcontinent

The spatial frequency distribution of R1a1* across Eurasia along with spatial representation of associated diversity based on microsatellite markers within the haplogroup are given in Figures 1a and b. It was interesting to find that by adding information regarding the frequency and diversity of R1a1* from different population groups of North India (Information from North Indian population groups was scanty in earlier publications from India.) to the pooled data from different published sources, a clearer picture emerged, with overlapping high frequency and molecular diversity of R1a1* within India.

Admixture and diversity analysis

Considering the very high frequency of R1a1* (upto 72.22% as in WB) in Brahmins, irrespective of their geographical and linguistic affiliations, admixture analysis[41] based on pooled data was performed. Three models of potential parental contributions of R1a1* (Figure 2) were tested, to evaluate the concepts of Central Asian introduction of the Indian caste system[7] by Indo-Aryans (appointing themselves to the castes of higher ranks),[14] as well as of rank-related West Eurasian admixture.[11,21] The observed proportions of contributions, taking all populations (Europeans (EU), Central Asians (CA) and Indian Brahmins (IB)) alternatively as source populations under different models (Figure 2), suggested model 3 (CA+IB→EU) as the best fit model (tested by 1000 bootstraps) and model 2 also as a possibility, for contributions of R1a1*, based on both proportion of frequency distribution as well as molecular divergence. Admixture analysis in light of other genetic evidences from this study did not seem to favor either Central Asian origin of the haplogroup or rank-related Eurasian admixture; instead it supported the Indian origin of this haplogroup and its contributions to other regions.

Further, the average diversity of the R1a1* haplogroup in Central Asians, Europeans and Indians was also calculated. The highest diversity of 0.52 (for both sampling and stochastic processes s.d.=0.32) was observed in Indians when compared with Europeans (0.40, s.d.=0.27) and Central Asians (0.32, s.d.=0.23). The calculation of Spearman's rank correlation coefficients[46] between the latitude and longitude with haplogroup R1a1* frequency (r^2=−0.13, 0.30) did not show any significant correlation, The same observation for R1a1* diversity (r^2=−0.25, 0.20) has been reported earlier as well.[9] This observation is again in favor of the suggestion that there has been no bulk migration from Central Asia to India.

Nested cladistic analysis

To investigate the patterns of genetic diversity within different segments of R1a1* (separately), nested cladistic analysis[42] was performed. This method attempts to disentangle historical and geographical explanations of patterns of genetic diversity.[54] The cladogram was designed by using haplotypic data of Y-STRs and the total cladogram ended in 5-step clade level. Details of analyses, significant observations and inferences and interpretations are given in Supplementary Table 3. It was very interesting to observe that the analyzed haplotypic data within R1a1* encompassed pooled data of different Indian, European and Asian population groups from different geographical locations, but that all the observed clades showing significant geographical differentiations were in the Indian subcontinent (Supplementary Table 3 and Supplementary Figure 4). After applying the latest predefined inference keys (provided with the GEODIS 2.4 software), the significant clades resulted in two concluding inferences (Supplementary Table 3) of either restricted gene flow with isolation by distance (RGF with IBD) or contiguous range expansion (CRE).

Molecular evidences for the origin of R1a1* in the Indian subcontinent

The median joining network[40] was also constructed. This algorithm provides the best results when applied on datasets of multi-state markers but within closely related haplotypes[54] as is the case, using

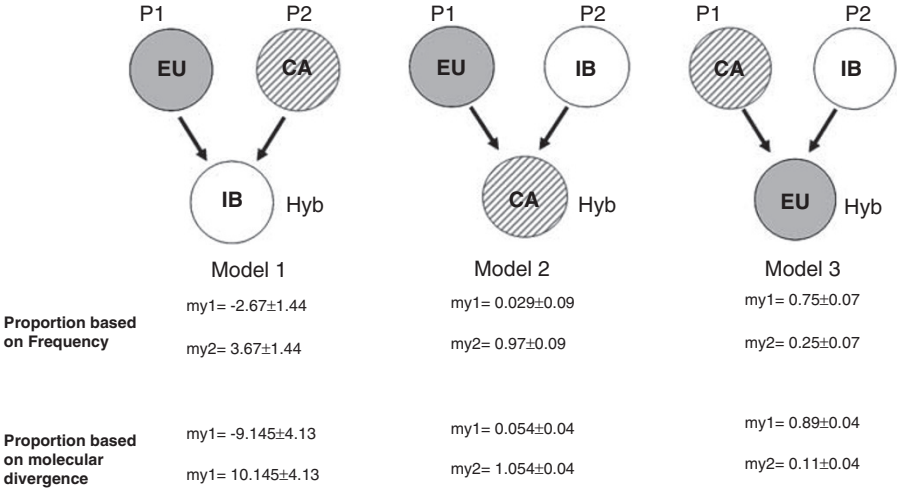

Figure 2 Admixture proportions were estimated using ADMIX2 software under different models. All populations (Europeans (EU), Central Asians (CA) and Indian Brahmins (IB)) were considered alternatively as source populations and the respective proportions of contributions were estimated. mY1 and mY2 are the estimated admixture coefficients, corresponding to the relative contribution to the hybrid population (Hyb) from the parental populations (P1 and P2, respectively).

pooled data of R1a1* haplogroup. The inferences from the analysis (Figure 3) were again in favor of our earlier observations. The Indian haplotypes were observed to be the most diverse, and haplotypes spanning Central Asia and Eurasia, along with some Indian regional haplotypes, seemed to be derived as a subset of this diversity. The extremely high level of sharing of haplotypes across the regions as well as reticulations, mostly with one step difference, in this subset suggests parallel evolution of different haplotypes, which appears more plausible after their geographical distribution and expansions. However, the diversity within the Indian populations, represented by the long branches and links connecting many haplotypes, is also an indicator of their ancestry, geographical differentiation and severe bottlenecks within India, suggesting loss of many of the intermediate haplotypes, thus reducing the reticulation and increasing the branches' length. The observed genetic distances FST[38] and 1−PSA[44] within the R1a1* haplogroup, between Central Asians (CA), Europeans (EU), as well as pooled populations of the Indian subcontinent (IS) showed overlapping trends of distribution. FST is based on the total variance in allele frequencies among populations and 1−PSA considers shared allele frequencies. IS populations showed less sharing with the CA (FST=0.095, 1−PSA=0.61) as compared with the EU (FST=0.021, 1−PSA=0.73) populations. AMOVA for these three pooled population groups (EU, CA, IS) showed that 94.07% of the total variation is present within the population, whereas only 5.93% of the differences are observed among population groups.

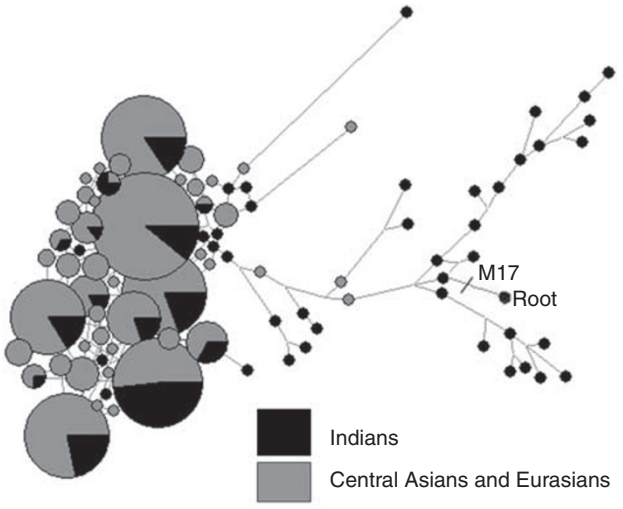

Figure 3 Median joining network based on Y-STR haplotypes within Y-haplogroup R1a1*, showing the relationship between Indian, Central Asian and Eurasian population groups. *Biallelic marker M17 was included with the highest weight. The root of the network represents an individual with SRY10831b-R1a* (x M17-R1a1).

Age estimates for Y-haplogroup R1a1*

The age of microsatellite variations was re-calculated using Y-STRs data and by applying mutation rates and generation times (discussed in Materials and methods) within R1a1* lineage in Central Asia, Eurasia, Pakistan, as well as Indian populations (Table 3), and

Table 3 Age of Microsatellite variations based on pooled variance of 6Y-STRs within R1a1* and R1a* lineages in different population groups

			Age[a]		*Age*[b]	
Haplogroup	*Population group (n)*	*Variance*	*Age*	*[95% CI] (in years)*	*Age*	*[95% CI] (in years)*
R1a1*	Overall population (979)	0.36	13043.48	(9473.68–20930.23)	16695.65	[12126.32–26790.70]
	Central Asia (127)	0.24	8695.65	(6315.79–13953.49)	11130.43	[8084.21–17860.47]
	Europe and Near East (119)	0.31	11231.88	(8157.89–18023.26)	14376.81	[10442.11–23069.77]
	Pakistan (224)	0.40	14492.75	(10526.32–23255.81)	18550.72	[13473.68–29767.44]
	India (Total) (509)	**0.38**	**13768.12**	**(10000–22093.02)**	**17623.19**	**[12800–28279.07]**
	India (Tribes + Dalits) (256)	0.41	14855.07	(10789.47–23837.21)	19014.49	[13810.53–30511.63]
	India (Tribes) (140)	0.40	14492.75	(10526.32–23255.81)	18550.72	[13473.68–29767.44]
	India (Brahmins) (253)	0.34	12318.84	(8947.37–19767.44)	15768.12	[11452.63–25302.33]
	Kashmiri Pandits (10)	0.52	18840.58	(13684.21–30232.56)	24115.94	[17515.79–38697.67]
	Himachal Brahmins (9)	0.43	15579.71	(11315.79–25000.00)	19942.03	[14484.21–32000.00]
	Bihar Brahmins (23)	0.30	10869.57	(7894.74–17441.86)	13913.04	[10105.26–22325.58]
	UP Brahmins (21)	0.35	12681.16	(9210.53–20348.84)	16231.88	[11789.47–26046.51]
	Maharashtra Brahmins (40)	0.28	10144.93	(7368.42–16279.07)	12985.51	[9431.58–20837.21]
	Gujarat Brahmins (21)	0.31	11231.88	(8157.89–18023.26)	14376.81	[10442.11–23069.77]
	Punjab Brahmins (10)	0.32	11594.20	(8421.05–18604.65)	14840.58	[10778.95–23813.95]
	MP Brahmins (16)	0.27	9782.61	(7105.26–15697.67)	12521.74	[9094.74–20093.02]
	Orissa Brahmins (10)	0.37	13405.80	(9736.84–21511.63)	17159.42	[12463.16–27534.88]
	WB Brahmins (13)	0.40	14492.75	(10526.32–23255.81)	18550.72	[13473.68–29767.44]
	Dravidian Brahmins (80)	0.29	10507.25	(7631.58–16860.47)	13449.28	[9768.42–21581.40]
	Saharia Tribe (16)	0.36	13043.48	(9473.68–20930.23)	16695.65	[12126.32–26790.70]
R1a*						
	Kashmir (9[a])	0.43	15579.71	(11315.79–25000.00)	19942.03	[14484.21–32000.00]
	Saharia Tribe (13)	0.60	21739.13	(15789.47–34883.72)	27826.09	[20210.53–44651.16]
	India (22)	**0.51**	**18478.26**	**(13421.05–29651.16)**	**23652.17**	**[17178.95–37953.49]**

Number of samples present in Haplogroup R1a1* or R1a*
[a]Seven extra R1a* samples were pooled (unpublished data).
Age[a]: age calculated using generation time of 25 years. Age[b]: age calculated using generation time of 32 years.

compared with the already published ages. The ages of the haplogroup, within the various population groups of India as well as after distributing them to social groups, were also calculated (Table 3). It was observed that the age of R1a1* was the highest in the Indian subcontinent. Interestingly, among different groups, the age of Y-haplogroup R1a1* was highest in scheduled castes/tribes when compared with Central Asians and Eurasians. These observations weaken the hypothesis of introduction of this haplogroup and the origin of Indian higher most castes from Central Asian and Eurasian regions, supporting their origin within the Indian subcontinent. Further, a particular population group of northern India, the Kashmiri Pandits (KPs), showed the highest variance (0.52) and thus the respective age (Table 3). Another north Indian population group, Himachal Brahmins, also showed higher variance (0.43) than that of the average Indian population.

High frequency of Y-haplogroup R1a1* in tribal populations and ancestral Y-haplogroup R1a* in the Indian subcontinent

Y-haplogroup R1a1* has been reported to be present in the tribal population in many of the earlier studies, but with very less frequency. In this study, a tribe named Saharia from Madhya Pradesh (Central India) showed the presence of R1a1* with high diversity in 19/71 males (26.76%), negating the idea of later admixture or some founder effect. Similar observations were made in the Chenchu tribe of Andhra Pradesh,[24] with a high percentage (26.82%) of R1a1*.

Apart from the observation of a simultaneous presence of R1a*, the ancestral haplogroup of R1a1* was also observed in this study with a highest ever known frequency in the two population groups KPs and Saharia. Incidentally, KPs are Brahmins, whereas Saharia is a tribal population group. Scanty representation of the R1a* haplogroup and its ancestral lineages (R*, R1*) in any of the geographical regions and the presence of the R1a1* haplogroup at high frequency across Central Asia and Eurasia had kept alive the question of the origin of R1a1* and associated conflicts. With the high-resolution analyses of the haplogroup (R1) in some population groups that were absent in the earlier studies and with the addition of published datasets, we were able to provide a clearer picture of the origin of R1a1* haplogroup and solve the existing conflict in literature. The calculated age for the haplogroup R1a* in both the population groups showed fascinating results. It was observed that the variance (0.43) of R1a*, and hence the respective age of this ancestral haplogroup, was far less in Kashmiris than the observed variance (0.52) and age of the derived R1a1*. However, a variance of 0.6 was observed in the Saharia tribe for R1a*, providing the age of 21 739.13 with 95% CI 15 789.47–34 883.72 years to this haplogroup. The haplogroup R1a1* was found to have an age of 13 043.48 and 95% CI 9473.68–20 930.23 years. To resolve the contradiction in these observations, we tried to explore the whole of the R1a lineage in these two population groups. By providing higher weight to the SNP (M17 that defines R1a1*) in the median joining network of Y-STR haplotypes within the R1a lineage among KPs and Saharia (Figure 4), we were able to elucidate some important inferences based on the clustering of the haplotypes. Two main clusters differentiating R1a* and R1a1* haplogroups were observed at the first instance. Further, subclustering based on population groups could be seen within these major clusters. However, few individuals belonging to KPs were seen in the Saharia population group clusters and vice versa, representing both R1a* and R1a1* haplogroups. It was particularly interesting to observe close overlaps in R1a1* cluster. Further, the long branches and less networking in both of the clusters (R1a* and R1a1*) again indicated bottlenecks and expansions, eliminating many of the haplotypes and resulting in long branches in the median joining tree. The exclusive high presence of the ancestral R1a* lineage in KPs and Saharias, their level of sharing, observed by way of a PSA of 0.51 (based on the average of Y-STRs within R1a*) and clustering in the network, suggested their deep common ancestry, a probable source population for the origin of R1a1* and for Brahmins, which later on differed in the two population groups. This observation of a close relationship was reflected in the MDS plot based on FST values obtained from a haplotypic analysis of 6Y-STRs within R1a1* (Supplementary Figure 5). Some of the other evidences hinting at this closeness are reflected in the cultural practices as well as folklores of these population groups.

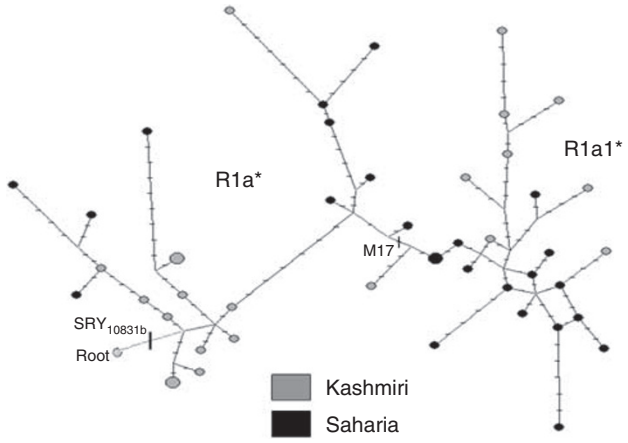

Figure 4 Median joining network based on Y-STR haplotypes showing the relationship between Kashmiri and Saharia Y-chromosomes bearing Y-haplogroups R1a* and R1a1*. Biallelic markers M17 and SRY10831b were also included and given the highest weight. The root of the network represents an individual with M173-R1* (x SRY10831b-R1a).

Conclusions

The observation of R1a* in high frequency for the first time in the literature, as well as analyses using different phylogenetic methods, resolved the controversy of the origin of R1a1*, supporting its origin in the Indian subcontinent. Simultaneously, the presence of R1a1* in very high frequency in Brahmins, irrespective of linguistic and geographic affiliations, suggested it as the founder haplogroup for the population. The co-presence of this haplogroup in many of the tribal populations of India, its existence in high frequency in Saharia (present study) and Chenchu tribes, the high frequency of R1a* in Kashmiri Pandits (KPs—Brahmins) as well as Saharia (tribe) and associated phylogenetic ages supported the autochthonous origin and tribal links of Indian Brahmins, confronting the concepts of recent Central Asian introduction and rank-related Eurasian contribution of the Indian caste system.

However, there is a scanty representation of Y-haplogroup R1a1 subgroups in the literature as well as in this study. The known subgroups (R1a1a, R1a1b and R1a1c), which are defined by binary markers M56, M157 or M87, respectively (Supplementary Figure 1), were not observed. In such a situation, it is likely that this haplogroup (R1a1*) is a polyphyletic (or paraphyletic) group of Y-lineages. It is, therefore, very important to discover novel Y chromosomal binary marker(s) for defining monophyletic subhaplogroup(s) belonging to Y-R1a1* with a higher resolution to confirm the present conclusion. Further, the under-representation of phylogenetic data of the population groups of North India in the literature and our observations hint at the immense need of phylogenetic explorations in the northern most Himalayan regions of India, which might have acted as an

incubator of many ancient lineages, to obtain a clearer picture of the peopling of India and Eurasia.

ACKNOWLEDGEMENTS

The financial assistance to SS as SRF (CSIR), ER as JRF (UGC), PS as SRF (DBT) and MJ as project assistant (DBT) is acknowledged. The financial assistance (UGC, India) and project grant (DBT, India) to the National Centre of Applied Human Genetics are also acknowledged.

1 Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton University Press, New Jersey, 1994).
2 Karve, I. *Kinship Organisation in India* (Asia Publishing House, Mumbai, India, 1968).
3 Majumder, P. P. Ethnic populations of India as seen from an evolutionary perspective. *J. Biosci.* **26,** 533–545 (2001).
4 Sharma, S., Saha, A., Rai, E., Bhat, A. & Bamezai, R. Human mtDNA hypervariable regions HVR I and II, hint at deep common maternal founder and subsequent maternal gene flow in Indian population groups. *J. Hum. Genet.* **50,** 497–506 (2005).
5 Das, K., Malhotra, K. C., Mukherjee, B. N., Walter, H., Majumder, P. P. & Papiha, S. S. Population structure and genetic differentiation among 16 tribal populations of central India. *Hum. Biol.* **68,** 679–705 (1996).
6 Basu, A., Mukherjee, N., Roy, S., Sengupta, S., Banerjee, S., Chakraborty, M. *et al.* Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Res.* **13,** 2277–2290 (2003).
7 Cordaux, R., Aunger, R., Bentley, G., Nasidze, I., Sirajuddin, S. M. & Stoneking, M. Independent origins of Indian caste and tribal paternal lineages. *Curr. Biol.* **14,** 231–235 (2004).
8 Singh, K. S. (ed.) *The Scheduled Tribes* (Oxford University Press, Oxford, UK, 1997).
9 Sengupta, S., Zhivotovsky, L. A., King, R., Mehdi, S. Q., Edmonds, C. A., Chow, C. E. *et al.* Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian Pastoralists. *Am. J. Hum. Genet.* **78,** 202–221 (2006).
10 Wooding, S., Ostler, C., Prasad, B. V., Watkins, W. S., Sung, S., Bamshad, M. *et al.* Directional migration in the Hindu castes: inferences from mitochondrial, autosomal and Y-chromosomal data. *Hum. Genet.* **115,** 221–229 (2004).
11 Bamshad, M., Kivisild, T., Watkins, W. S., Dixon, M. E., Ricker, C. E., Rao, B. B. *et al.* Genetic evidence on the origins of Indian caste populations. *Genome Res.* **11,** 994–1004 (2001).
12 Sahoo, S., Singh, A., Himabindu, G., Banerjee, J., Sitalaximi, T., Gaikwad, S. *et al.* A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. *Proc. Natl. Acad. Sci. USA* **103,** 843–848 (2006).
13 Renfrew, C. *Archaeology and Language: The Puzzle of Indo-European Origins* (Cambridge University Press, Cambridge, 1988).
14 Poliakov, L. *The Aryan Myth* (Basic Books, New York, USA, 1974).
15 Thapar, R. *A History of India* (Penguin, Middlesex, UK, 1966).
16 Ratnagar, S. In *Recent Perspectives of Early Indian History* (ed. Thapar, R.) 1–52 (Popular Prakashan, Bombay, India, 1995).
17 Cordaux, R., Saha, N., Bentley, G. R., Aunger, R., Sirajuddin, S. M. & Stoneking, M. Mitochondrial DNA analysis reveals diverse histories of tribal populations from India. *Eur. J. Hum. Genet.* **11,** 253–264 (2003).
18 Bamshad, M. J., Watkins, W. S., Dixon, M. E., Jorde, L. B., Rao, B. B., Naidu, J. M. *et al.* Female gene flow stratifies Hindu castes. *Nature* **395,** 651–652 (1998).
19 Bhattacharyya, N. P., Basu, P., Das, M., Pramanik, S., Banerjee, R., Roy, B. *et al.* Negligible male gene flow across ethnic boundaries in India, revealed by analysis of Y-chromosomal DNA polymorphisms. *Genome Res.* **9,** 711–719 (1999).
20 McElreavey, K. & Quintana-Murci, L. A population genetics perspective of the Indus Valley through uniparentally inherited markers. *Ann. Hum. Biol.* **32,** 154–162 (2005).
21 Thanseem, I., Thangaraj, K., Chaubey, G., Singh, V. K., Bhaskar, L. V., Reddy, B. M. *et al.* Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA. *BMC Genet.* **7,** 42 (2006).
22 Ammerman, A. J. & Cavalli-Sforza, L. L. *Neolithic Transition and the Genetics of Populations in Europe* (Princeton University Press, New Jersey, 1984).
23 Wells, R. S., Yuldasheva, N., Ruzibakiev, R., Underhill, P. A., Evseeva, I., Blue-Smith, J. *et al.* The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc. Natl. Acad. Sci. USA* **98,** 10244–10249 (2001).
24 Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., Parik, J. *et al.* The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am. J. Hum. Genet.* **72,** 313–332 (2003).
25 Heinz, C. B. *Asian Cultural Traditions* (Waveland Press, Prospect Heights, IL, 1999).
26 Dutta, R., Reddy, B. M., Chattopadhyay, P., Kashyap, V. K., Sun, G. & Deka, R. Patterns of genetic diversity at the nine forensically approved STR loci in the Indian populations. *Hum. Biol.* **74,** 33–49 (2002).
27 Misra, V. N. Prehistoric human colonization of India. *J. Biosci.* **26,** 491–531 (2001).
28 Zerjal, T., Pandya, A., Thangaraj, K., Ling, E. Y., Bertoneri, S., Paracchini, S. *et al.* Y-chromosomal insights into the genetic impact of the caste system in India. *Hum. Genet.* (2006).
29 Ramana, G. V., Su, B., Jin, L., Singh, L., Wang, N., Underhill, P. *et al.* Y-chromosome SNP haplotypes suggest evidence of gene flow among caste, tribe, and the migrant Siddi populations of Andhra Pradesh, South India. *Eur. J. Hum. Genet.* **9,** 695–700 (2001).
30 Karafet, T., Xu, L., Du, R., Wang, W., Feng, S., Wells, R. S. *et al.* Paternal population history of East Asia: sources, patterns, and microevolutionary processes. *Am. J. Hum. Genet.* **69,** 615–628 (2001).
31 Karafet, T. M., Osipova, L. P., Gubina, M. A., Posukh, O. L., Zegura, S. L. & Hammer, M. F. High levels of Y-chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. *Hum. Biol.* **74,** 761–789 (2002).
32 Zerjal, T., Wells, R. S., Yuldasheva, N., Ruzibakiev, R. & Tyler-Smith, C. A genetic landscape reshaped by recent events: Y-chromosomal insights into central Asia. *Am. J. Hum. Genet.* **71,** 466–482 (2002).
33 Cinnioglu, C., King, R., Kivisild, T., Kalfoðlu, E., Atasoy, S., Cavalleri, G. L. *et al.* Excavating Y-chromosome haplotype strata in Anatolia. *Hum. Genet.* **114,** 127–148 (2004).
34 Al-Zahery, N., Semino, O., Benuzzi, G., Magri, C., Passarino, G., Torroni, A. *et al.* Y-chromosome and mtDNA polymorphisms in Iraq, a crossroad of the early human dispersal and of post-Neolithic migrations. *Mol. Phylogenet. Evol.* **28,** 458–472 (2003).
35 Y-ChromosomeConsortium. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res.* **12,** 339–348 (2002).
36 Underhill, P. A., Passarino, G., Lin, A. A., Shen, P., Mirazón Lahr, M., Foley, R. A. *et al.* The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* **65,** 43–62 (2001).
37 Sharma, S., Rai, E., Bhat, A. K., Bhanwer, A. S. & Bamezai, R. N. A novel subgroup Q5 of human Y-chromosomal haplogroup Q in India. *BMC Evol. Biol.* **7,** 232 (2007).
38 Slatkin, M. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139,** 457–462 (1995).
39 Excoffier, L., Laval, L. G. & Schneider, S. Arlequin ver 3.0: an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* **1,** 47–50 (2005).
40 Bandelt, H. J., Forster, P. & Rohl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16,** 37–48 (1999).
41 Dupanloup, I. & Bertorelle, G. Inferring admixture proportions from molecular data: extension to any number of parental populations. *Mol. Biol. Evol.* **18,** 672–675 (2001).
42 Templeton, A. R. Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Mol. Ecol.* **7,** 381–397 (1998).
43 Posada, D., Crandall, K. A. & Templeton, A. R. GeoDis: a program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. *Mol. Ecol.* **9,** 487–488 (2000).
44 Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R. & Cavalli-Sforza, L. L. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368,** 455–457 (1994).
45 Delfiner, P. In *Advanced Geostatistics in the Mining Industry* (eds. Guarasio, M., David, M. & Haijbegts, C.) 49–68 (Reidel, Dordrecht, Austria, 1976).
46 Spearman, C. The proof and measurement of association between two rings. *Am J. Psychol.* **15,** 72–101 (1904).
47 Quintana-Murci, L., Krausz, C., Zerjal, T., Sayar, S. H., Hammer, M. F., Mehdi, S. Q. *et al.* Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. *Am. J. Hum. Genet.* **68,** 537–542 (2001).
48 Zhivotovsky, L. A., Underhill, P. A., Cinnioðlu, C., Kayser, M., Morar, B., Kivisild, T. *et al.* The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am. J. Hum. Genet.* **74,** 50–61 (2004).
49 Tremblay, M. & Vezina, H. New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *Am. J. Hum. Genet.* **66,** 651–658 (2000).
50 Helgason, A., Hrafnkelsson, B., Gulcher, J. R., Ward, R. & Stefansson, K. A population-wide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. *Am. J. Hum. Genet.* **72,** 1370–1388 (2003).
51 Barbujani, G. Geographic patterns: how to identify them and why. *Hum. Biol.* **72,** 133–153 (2000).
52 Carvalho-Silva, D. R., Zerjal, T. & Tyler-Smith, C. Ancient Indian roots? *J. Biosci.* **31,** 1–2 (2006).
53 Tambets, K., Tolk, H. V., Kivisild, T., Metspalu, E., Parik, J., Reidla, M. *et al. Examining the Farming/Language Dispersal Hypothesis* (eds. Bellwood, P. & Renfrew, C.) (McDonald Institute for Archaeological Research, Cambridge, UK, 2003).
54 Jobling, M. A., Hurles, M. & Tyler-Smith, C. *Human Evolutionary Genetics-Origins, Peoples and Disease* (Garland Science, New York, 2004).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)