Published in final edited form as:

Ann Hum Biol. 2009 ; 36(1): 46–59. doi:10.1080/03014460802558522.

Presence of three different paternal lineages among North Indians: A study of 560 Y chromosomes

Zhongming Zhao^{1,2,3}, Faisal Khan⁴, Minal Borkar⁴, Rene Herrera⁴, and Suraksha Agrawal⁴

¹Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia, USA

²Department of Human Genetics, Virginia Commonwealth University, Richmond, Virginia, USA

³Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, Virginia, USA

⁴Department of Medical Genetics, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow, (UP) 226014, India

Abstract

Background—The genetic structure, affinities, and diversity of the 1 billion Indians hold important keys to numerous unanswered questions regarding the evolution of human populations and the forces shaping contemporary patterns of genetic variation. Although there have been several recent studies of South Indian caste groups, North Indian caste groups, and South Indian Muslims using Y-chromosomal markers, overall, the Indian population has still not been well studied compared to other geographical populations. In particular, no genetic study has been conducted on Shias and Sunnis from North India.

Aim—This study aims to investigate genetic variation and the gene pool in North Indians.

Subjects and methods—A total of 32 Y-chromosomal markers in 560 North Indian males collected from three higher caste groups (Brahmins, Chaturvedis and Bhargavas) and two Muslims groups (Shia and Sunni) were genotyped.

Results—Three distinct lineages were revealed based upon 13 haplogroups. The first was a Central Asian lineage harbouring haplogroups R1 and R2. The second lineage was of Middle-Eastern origin represented by haplogroups J2*, Shia-specific E1b1b1, and to some extent G* and L*. The third was the indigenous Indian Y-lineage represented by haplogroups H1*, F*, C* and O*. Haplogroup E1b1b1 was observed in Shias only.

Conclusion—The results revealed that a substantial part of today's North Indian paternal gene pool was contributed by Central Asian lineages who are Indo-European speakers, suggesting that extant Indian caste groups are primarily the descendants of Indo-European migrants. The presence of haplogroup E in Shias, first reported in this study, suggests a genetic distinction between the two Indo Muslim sects. The findings of the present study provide insights into prehistoric and early historic patterns of migration into India and the evolution of Indian populations in recent history.

Correspondence: Suraksha Agrawal, PhD, Department of Medical Genetics, Sanjay Post Graduate Institute of Medical Sciences, Raebareli Road, Lucknow (UP) 226014 India. E-mail: suraksha@sgpgi.ac.in.

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

Publisher's Disclaimer: The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Keywords

Paternal lineages; Y-chromosomal markers; North Indians; migration

Introduction

India occupies a unique stage in human population evolution because one of the early waves of migration of modern humans was out of Africa, through West Asia, into India (Cann 2001). More recently, about 15 000–10 000 years before present (ybp), when agriculture developed in the Fertile Crescent region that extended from Israel through Northern Syria to Western Iran, there was an eastward wave of human migration (Renfrew 1989; Cavalli-Sforza et al. 1994). It has been postulated that this wave brought the Dravidian language into India (Renfrew 1989). Subsequently, the Indo-European (Aryan) language was introduced into India from the Iranian plateau approximately 4000–3000 ybp, where this language was probably brought by pastoral nomads from the Central Asian steppes (Renfrew 1989). Therefore, linguistic evidence suggests that West Asia and Central Asia were two major geographical sources contributing to the Indian gene pool.

Indian society predominantly revolves around the concept of caste, or the Caste System, a strong socio-cultural conglomerate of traditions that have created and maintained a great number of hierarchically arranged endogamous groups (Bamshad et al. 2001). This unique social system exists only in India. One impact of the system is that a person's fate, including even the choice of marriage partner, is largely determined at his/her birth. The Hindu caste system plays a major role in social and economic organization of the Indian population. In this system, the society is divided into four broad castes: (from low to high) Sudras, Vaishyas, Kshatriyas and Brahmins. The rules that generally prevent marriages between castes may have contributed to population substructure and the pattern of genetic diversity. Another important feature in Indian population history was the occurrence of four separate or distinct waves of migration into the subcontinent (Cordaux et al. 2004): (i) an ancient Palaeolithic migration by modern humans, (ii) an early Neolithic migration, probably via Proto-Dravidian speakers from the eastern horn of the Fertile Crescent, (iii) an influx of Indo-European speakers, and (iv) a migration from East/Southeast Asians, i.e. Tibeto-Burman speakers. In addition to these migrations, India has also experienced colonization by Europeans, which may have also contributed to the ethnic multiplicity. Furthermore, it has been reported (Cordaux et al. 2004) that the Y lineages of Indian castes are more closely related to Central Asians than to Indian tribal populations, suggesting that Indian caste groups are primarily the descendants of Indo-European migrants.

The Y chromosome is one of the most informative loci for investigating genetic diversity and population substructure. The DNA variation found within the non-recombining portion of the Y chromosome (NRY) reflects a simple paternal history revealed by the pattern of alleles at informative loci, i.e. markers, comprising the haplotype (Underhill et al. 2000). At present, NRY contains approximately 600 binary markers which form 311 distinct haplogroups in the Y chromosome tree (Karafet et al. 2008). When we combine DNA variation information with other population characteristics such as geographic, archaeological, and linguistic background, we may have more power to trace the histories of contemporary populations. Moreover, Y-chromosomal markers are appropriate for the genetic study of populations that have small effective population size or recent divergence time. Therefore, markers on the Y chromosome should provide a powerful tool for investigating the origins of Indian populations, as well as genetic substructure within these populations.

There have been a few recent studies of both South Indian and North Indian caste groups, and South Indian Muslims using Y-chromosomal markers (Gutala et al. 2006; Sahoo et al. 2006; Zerjal et al. 2007). However, no genetic study has been conducted on Shias and Sunnis from North India. Such knowledge is expected to provide insights on migration into and within India, as well as how Indian populations have evolved in recent history, and how genetic variation is distributed in the modern Indian population.

In this study we examined the genetic compositions of three endogamous North Indian upper caste populations (Brahmins and two sub-populations of Brahmins: Bhargavas and Chaturvedis) and two Muslims sects (Shias and Sunnis). Bhargavas and Chaturvedis practice strict surname endogamy (Agrawal et al. 2005) and Muslims sects (Shias and Sunnis) practice consanguinity. One major aim of the present study is to evaluate the impact of Muslim invasions and their admixturing with upper caste Hindus who otherwise claim to be highly endogamous. We selected Bhargava, Chaturvedis and Brahmins because they are highly homogeneous groups and follow strict endogamy, which does not apply to the lower caste populations. We examined a set of 32 Y-unique event polymorphisms (UEPs) including 27 single nucleotide polymorphisms (SNPs), four insertions/deletions (indels) and 1 *Alu* repeat in a total of 560 Y chromosomes from three upper caste and two Muslim populations in North India. Analysis of these markers revealed substantial genetic variation between groups and provided evidence of male-driven gene flow among these populations.

Materials and methods

Human samples

Blood samples from 560 unrelated male individuals belonging to five populations (Bhargavas, $n=96$; Chaturvedis, $n=88$; Brahmins, $n=118$; Shia Muslims, $n=154$; and Sunni Muslims, $n=104$) were collected randomly from different regions of Uttar Pradesh, India. The collection sites included districts of Lucknow, Kanpur, Raebareilly, Barabanki, Faizabad, Agra, Jhansi, Gonda and Basti. All participants were normal adults and the mean age was 38.8 ± 3.4 years. All individuals sampled belonged to the Indo-Aryan linguistic family and were at least third generation residents of Uttar Pradesh.

Prior to the sample collection, regional addresses and detailed computerized lists of the populations were prepared from different districts of Uttar Pradesh. Random numbers were generated with the help of a computer program and adult individuals living in different parts of Uttar Pradesh were questioned about their ethnicity, caste affiliations and surnames, and the birthplaces of their parents, as per the order of the random number. Three generation pedigree charts were prepared to ascertain un-relatedness in all the samples. The study was performed after the approval of the institutional ethical reviewing committee of Sanjay Gandhi Post Graduate Institute of Medical Sciences (SGPGIMS), Lucknow. Informed consent was obtained from all the subjects and 5 mL blood was collected in EDTA by venepuncture. DNA was extracted by standard protocol (Corney et al. 1994).

Y-haplogroup

Genotypes of 32 binary Y-markers in 560 individuals were obtained by PCR-RFLP (Luis et al. 2004), the polymorphic *Alu* insertion (PAI), and standard direct sequencing methods. The primer sets and other conditions used were based on Underhill et al. (2000). The 32 binary markers included 28 SNPs, one Y-chromosome *Alu* polymorphism (YAP), and three indels. ABI DNA Sequencing Analysis (v 3.4.1) was used for sequence examination and analysis. The nucleotide sequences (amplicon sequences) were aligned with the published sequence of the UEP (Underhill et al. 2001; The Y Chromosome Consortium 2002). The presence or absence

of the mutation was scored using the software BioEdit (version 5.0, <http://www.mbio.ncsu.edu/BioEdit/BioEdit.html>).

Data analysis

Allele and haplotype frequencies were obtained by the direct counting method from the observed number of alleles at a locus divided by the total number of gametes or haplotypes, as described in Chakraborty and Jin (1993). We used the PHYLIP package (version 3.57c) (Felsenstein 2005) to estimate the genetic distance and reconstruct phylogenetic trees. AMOVA (analysis of molecular variance) was carried out by using the software ARLEQUIN (version 2.0) (Schneider et al. 2000).

Results

We investigated 32 Y-chromosome markers in 560 Indian males from five Indian subpopulations. We observed 13 haplogroups (C[★], E1b1b1, F[★], G[★], H1[★], J2[★], K[★], L[★], O[★], P[★], R1a1[★], R1b1b2[★] and R2) in these samples. The detailed haplogroups and their frequencies are shown in Figure 1. The common haplogroups included 'R1a1[★]', 'R2', 'J2[★]', 'H1[★]', and 'C[★]', which accounted for 78.0% of all the Y chromosomes. Four haplogroups (F[★], K[★], O[★] and P[★]) had a low frequency in all five populations. Their frequencies ranged from 2.7 to 3.0%. Interestingly, haplogroup E1b1b1 was present only in Shia Muslims (11.0% of the Shia sample). Moreover, none of the Y lineages carried the derived alleles for M174, M170, TAT, M70, M4 and M3 UEPs. This lack of derived alleles of the above SNPs led to an absence of haplogroups D, I, K2, M, N and Q in the North Indian gene pool. The description of each haplogroup is given below.

Haplogroup R (R1 and R2)

There were a total of 256 Y chromosomes carrying the allele's characteristic of haplogroup R, which accounted for 45.7% of the total sample studied. Haplogroup R was segregated into R1 by the presence of the M173-C allele and R2 by the presence of the M124-T allele. Haplogroup R1, which was observed 133 times, was the most frequent (23.8%) and was found in all five populations. The frequency in each population was 24.0% (Bhargavas), 23.9% (Chaturvedis), 29.7% (Brahmins), 15.6% (Shias) and 28.8% (Sunnis), respectively (Figure 1). Interestingly, 130 of the 133 R1 lineages belonged to R1a1[★] haplogroup, which carried a 1 bp deletion at M17 UEP in the lineage of M173-C allele. R1b1b2[★] was only found in Bhargavas (once) and Brahmins (twice). Haplogroup R2 had a similar high frequency in the sample – it was observed 123 times. Its frequency in each group was 32.3% in Bhargavas, 31.8% in Chaturvedis, 20.3% in Brahmins, 13.0% in Shias, and 19.2% in Sunnis, respectively (Figure 1).

Haplogroup J2[★]

Haplogroup J2[★] was observed 77 times. Its frequency was higher in the Shia (19.5%) and Sunni (15.4%) sample set than in the other three upper caste populations (6.3% in Bhargavas, 12.5% in Chaturvedis, and 11.9% in Brahmins).

Haplogroup H1[★]

Haplogroup H1[★] was observed 59 times. Its frequency ranged from 6.8% (Chaturvedis) to 16.7% (Bhargavas) (Figure 1). The frequencies in the Brahmin, Shia and Sunni groups were 11.9%, 7.8% and 10.6%, respectively.

Haplogroup C[★]

Haplogroup C[★] was observed 48 times and present in all populations. Its frequencies varied in a narrow range (6.3–10.2%).

Haplogroup E

We found that 17 individuals, all of whom were Shia Muslims, carried YAP retrotransposition insertion derived haplogroup E. This is the first report of haplogroup E observed in Shias, and its high frequency excludes the possible false positives. Presence of YAP lineage was further confirmed by the G→A transition at M145 and the G→C transition at M203 UEP. Further resolution of the YAP + lineage in our population was into E haplogroup, which was identified by M40-A allele. All 17 individuals had a G→C mutation at M35, which is the designated marker to distinguish E1b1b1 (major subclade of E) from other E sub-haplogroups (Karafet et al. 2008). Importantly, E1b1b1 is the most frequently observed haplogroup in the Africa/Mid-East region. Therefore, our observation suggests that Shia Muslims might carry some African/Middle-Eastern ancestral alleles.

Haplogroup K★ and P★

Haplogroup K★ is a default haplogroup (paragroup) formed when there are no further sub-lineages found associated with the M9-G allele. Haplogroup P★ is also a default haplogroup and is a result of no sub-lineages in 92R7-T/M45-A/M74-A. Both these haplogroups presented in all populations selected in this study but at a low frequency (~3%) (Figure 1).

Haplogroup F★ and G★

Haplogroup F★ is another default haplogroup characterized by a T allele at M89 UEP. This haplogroup could not be resolved into any subsequent haplogroups (i.e. from G to R2). The frequency of this haplogroup was low (17 counts); however, it was present in all populations. Its frequency ranged between 1.9% in Sunnis to 4.2% in Bhargavas. Haplogroup G★ was derived from the T allele at M201 UEP. It was also observed in low frequency (23 counts) in our sample. Interestingly, it was found only twice in Brahmins and was absent in Bhargavas and Chaturvedis, thereby occurring at a very low frequency in those groups, compared to the frequency in Shias (9.7%) and Sunnis (5.8%) (Figure 1).

Haplogroup L★ and O★

Both haplogroups were present in low frequency in our sample. Haplogroup L★ was identified by a G allele at M11 and M20 UEPs. Its frequency ranged from 0 in Bhargavas to 3.9% in Shias. Haplogroup O★ was identified by a 5 bp deletion at M175. It was present in all the populations with a frequency ranging from 1.9 to 4.2% (Figure 1).

Analysis of molecular variance (AMOVA)

We performed an AMOVA using the 32 Y-chromosomal markers in the five populations from North India. The results provided quantitative support of the genetic affiliations of these populations (Table I). The five populations were in principle divided into two groups (upper caste and Muslims). We further divided the upper caste group into two subgroups: Brahmins who follow endogamy and Bhargavas and Chaturvedis who follow surname endogamy (i.e. marriage is only allowed within the same sub-population). The fraction of variation among the upper caste populations was moderately high (1.08%). We then compared each upper caste population with Muslims. Bhargavas had the highest fraction of variation (1.54%) with the Muslim group, while Brahmins had the smallest fraction of variation (0.18%). Overall, the fraction of variance between the upper caste as a whole group and Muslims was low (0.86%). This may reflect a moderate level of gene flow between the ancestral populations of upper caste and Muslims.

F-statistics

Table II shows the F_{st} genetic distance between the five populations. The results revealed a differential genetic relationship of the two Muslim groups, compared to the upper caste populations. Shia Muslims were distinct from the other populations probably due to the presence of its unique Y-lineage E1b1b1 and the high frequency of haplogroup J2★.

Phylogenetic reconstruction

We compared the five North Indian populations in this study with an additional 36 world populations published in three separate studies (Underhill et al. 2000; Ramana et al. 2001; Wells et al. 2001). We manually collected the polymorphism data overlapping with the markers used in our study from these papers, estimated genetic distance between each pair of populations using Nei's method (Nei 1973), and then reconstructed the phylogenetic tree using the neighbour-joining method (Saitou and Nei 1987). The analysis was performed by the PHYLIP package (Felsenstein 2005). Interestingly, the 41 populations could be separated into six clusters; each represents South East-Asian, African, European, Middle Eastern, Indian and Central-Asian populations, respectively (Figure 2). All Indian populations were clustered together, but there was further bifurcation between North Indian Brahmins and South Indian Brahmins (Vizag Brahmins). The five populations selected in this study (i.e. three Brahmin groups and two Muslim sects) were distributed along a single branch. Interestingly, not unexpectedly, the Central Asian populations were clustered; however, they were overall closer to Indian populations, depicting the gene flow from Central Asia.

Discussion

In the last 3000–10 000 years, Northern-Western India has experienced an enormous amount of gene flow from different parts of the world, with the majority of this gene flow being male-driven (Bamshad et al. 2001). Therefore, one of the major aims of the present study was to create a male-derived phylogeny of three higher caste groups (Brahmins, Chaturvedis and Bhargavas) and two Muslims groups (Shia and Sunni) from North India. To examine the genetic variation in these populations, we used evolutionarily stable binary markers (UEPs) that define the unique haplogroup signature of a particular Y-chromosome. The significant feature of these UEPs is that they segregate in the form of haplotypes to formulate unique haplogroups. These haplogroups can be used to understand the haplotype geography, affinity and diversification existing in the assorted human populations (Lahr and Foley 1998). Our examination of the 32 UEPs in 560 North Indian Y chromosomes revealed 13 different haplogroups (C★, E1b1b1, F★, G★, H1★, J2★, K★, L★, O★, P★, R1a1★, R1b1b2★ and R2), of which nine (C★, F★, H1★, J2★, K★, O★, P★, R1a1★ and R2) were present in all the studied populations.

The observed haplogroups in the five North Indian populations could be apportioned into three Y-lineages based on their possible origin. These lineages include Central Asian (or west Eurasian), Middle Eastern and possibly indigenous Indian Y-lineages. Distribution of these lineages is illustrative of past human migrations in the Indian subcontinent. We discuss this in more detail below.

Central Asian (or west Eurasian lineage)

The Central Asian or west Eurasian Y-lineages are depicted in terms of presenting a similar high frequency of sibling clades of R haplogroups (R1a1★ and R2) in the studied populations. A total of 256 of the 560 individuals (45.7%) in this study belonged to European Y-lineages, i.e. R1a1★ (M173/M17), R1b1b2★ (M173) and R2 (M124) clades (Figure 1). Similar results were reported in a previous study of the Indian subcontinent (Kivisild et al. 2003). Haplogroup R reflects the impact of expansion and migration of Indo-European pastoralists from Central

Asia, thus linking haplogroup frequency to specific historical events (Sengupta et al. 2006). Haplogroup R is widely spread in central Asian Turkic-speaking populations and in eastern European Finno-Ugric and Slavic speakers and is less frequent in populations from the Middle East and Sino-Tibetan regions of northern China (Karafet et al. 1999; Underhill et al. 2000).

Interestingly, the high frequency of the R1a1[★] haplogroup seems to be concentrated around the elevated terrain of central and western Asia. Several migratory routes of *H. sapiens* are illustrated in Figure 3. Although haplogroup R1a in Central Asians depicted a low genetic diversity estimate, many researchers (Kivisild et al. 2003; Zerjal et al. 2003) have suggested a recent founder effect or drift that led to the high frequency of R1a in the Southeastern Central Asia. It has also been suggested that R1a might have an independent origin in the Indian subcontinent (Kivisild et al. 2003). We have observed a low frequency of R1b1b2[★] (0.5%). An additional signature of the Central Asian lineage is haplogroup R2. Its frequency was 22.0% in our sample. This haplogroup is mainly found in Indian, Iranian, and Central Asian populations and has been postulated to have a Central Asian origin (Quintana-Murci et al. 2001; Wells et al. 2001; Kivisild et al. 2003). However, our results have shown that high incidence of R2 clade was also observed in other North Indian populations, which was similarly reported in other studies (Cordaux et al. 2004; Cavalli-Sforza 2005). Overall, we suggest that Central Asia is the most likely source of North Indian Y lineage considering the historical and genetic background of North India (Karve 1968; Balakrishnan 1978).

Middle East (West Asian) lineages

The Middle East is often called the Fertile Crescent due to the emergence of agriculture during the Neolithic era and is one of the most important geographical areas contributing to the initial population and re-population of Europe (Cavalli-Sforza and Feldman 2003). There were two putative mutations found in the Middle Eastern populations: YAP/PN2/ M35 and 12f2/M172 (Semino et al. 2000; Underhill et al. 2001). The first mutation creates haplogroup E1b1b1 while the other mutation defines haplogroup J2[★]. The Middle Eastern populations might have contributed differentially to the South Asian gene pool during the last 8000–10 000 years (Lahr and Foley 1998). The Y-lineages observed in the present study may suggest two major episodes of migrations: One carried J2 and to some extent L and G with the Neolithic farmers (Underhill et al. 2001) and the other arrived with the Muslims carrying E1b1b1 and a few more haplogroups such as J2[★] and G[★]. Kivisild et al. (2003) also reported the presence of a J2 clade and postulated that the origin of the J2 clade in India was probably Central Asia. Their hypothesis is based on eight populations taken from different parts of India. They observed the J2 clade in ~13% of the sample. The major Middle Eastern lineage present in our study was J2[★] with an average frequency of 13.8% and its frequency among Shias was the highest (19.5%). We suggest that the J2[★] lineage of the studied populations might be derived from the Middle East. This might have been due to two different episodes of migrations, one concomitant with the development and spread of agriculture ~8000–10 000 years ago (Renfrew 1989; Cavalli-Sforza 2005), and the other more recent migration being the arrival of Muslim rulers 1000 years ago. The supporting evidence of the Middle East or West Asian migrations in Indian Muslims was demonstrated by the presence of 11.0% of haplogroup E1b1b1 in Shia Muslims. Our results revealed that Shia Muslims are different from Sunnis and other upper caste populations. They possess a relatively high frequency of the E1b1b1 haplogroup which was not observed in any other population selected for the present study. It appears that gene pool of extant Shia Muslims reflects the contributions of earlier Islamic invaders who might have maintained the founder population features. Zerjal et al. (2007) have also recently reported the low frequency of E3b3a[★] (old nomenclature in YCC2002) in lower caste populations, i.e. Panchamas and Vaishyas populations of Uttar Pradesh, India.

Moreover, haplogroup G is primarily a Middle Eastern, Caucasus Region, and Mediterranean haplogroup that presents in northwestern Europe in only ~2% of males (Cruciani et al. 2002; Cinnioglu et al. 2004; Alonso et al. 2005; International Society of Genetic Genealogy 2006). The founder of haplogroup G is thought to have lived ~30 000 years ago along the eastern edge of the Middle East. In our study, the frequency of haplogroup G[★] was 4.1% in the total sample. However, its frequency varied greatly among the populations we studied. It was 9.7% in Shias, supporting the genetic signature of the Middle East in Shias. In contrast to the Muslims, haplogroup G[★] was absent in Chaturvedis and Bhargavas and in low frequency in Brahmins (1.7%).

Indigenous Indian lineage

The putative earliest inhabitants of India have been postulated to be tribal populations and more specifically, Austro-Asiatic tribes (Majumder 1998). These original inhabitants of India carry some of the Y-lineages dispersed from Africa (Underhill et al. 2000, 2001) that include unresolved F[★], H1[★], C[★] (without M217 transversion) and O[★] (Cordaux et al. 2004). Of these four haplogroups, H1[★] was found at a higher frequency among Indian tribes (30%) (Cordaux et al. 2004) and represents the major indigenous Indian haplogroup. Haplogroup H1[★] has rarely been seen in Central Asian or West Asian populations (Wells et al. 2001). In our study, we found haplogroup H1[★] in all five populations and it represented one of the major Indian haplogroups. Haplogroups C[★] (without M217) and O[★] are distributed widely in Southeast Asia and in earlier inhabitants of India. We observed a moderate frequency of haplogroup C[★] and a low frequency of haplogroup O[★] in our studied sample, suggesting that there might have been some recent gene flow among earlier inhabitants and caste groups as well as from Southeast Asia to India. The high level of gene flow that Northern India has experienced during the last 3000–8000 years (Majumder 1998) may have left a strong signature in the present day North Indian gene pool. The patterns of haplogroup/haplotype distribution observed in our studied populations may be reflective of the heavy contribution from Central Asian (or West Eurasian) and Middle Eastern lineages, and result from the constant gene flow between the earliest settlers of the Indian subcontinent and later immigrants and from Southeast Asia to India.

Genetic homogeneity in populations

Phylogenetic reconstruction has shown distinct genetic variation between South Indian Brahmins and North Indian Brahmins, as there was a clear bifurcation in the Indian population cluster (Figure 2). The two Muslim sects clustered with North Indian Brahmins, indicating that there was gene flow between these groups during invasions by Muslims into India ~1000–1500 years ago (Gutala et al. 2006). A closer examination of the phylogram shows that Indian populations tended to harbour lineages from Central Asia (Figure 2).

AMOVA results indicated that the per cent variance between upper castes and Muslim (Shias and Sunnis) groups was 0.86, again confirming that genetic variation between these two major groups is not large (Table I). In particular, the variation between Brahmins versus Muslim groups (0.18) was much less than that between other two groups.

Comparison with lower castes

So far, the origin of the caste system in India has still remained a mystery. A recent systematic comparison of Y lineages indicated that the Indian lower castes showed more similarity with tribal groups than with the upper caste populations, suggesting a tribal origin for the Indian lower castes (Thanseem et al. 2006). Furthermore, Zerjal et al. (2007) recently reported genetic isolation and drift within the Jaunpur upper castes, but not the lower castes, suggesting the influence of founder effects and social factors in the upper castes. In Thanseem et al. (2006), there were six major haplogroups (R, H, L, J, F[★] and O) compared among upper and lower

castes and tribal groups (see their Table IV). The frequencies of these six haplogroups in our North Indian upper caste sample were similar overall to those of upper castes in Thanseem et al. (2006), but both were very different from the corresponding frequencies in Indian lower castes or tribes. Thus, our data supports Thanseem et al.'s findings. Moreover, the frequencies of the six haplogroups in our Shia and Sunni sample were close to those in the upper castes but very different from the lower castes or tribal groups. For example, the frequency of haplogroup H was 8.9% in our Muslim sample (Shias and Sunnis), close to that in our North Indian upper castes (11.9%) and Indian upper castes in Thanseem et al. (9%), but markedly smaller than that in the lower castes (25%) or tribes (30%). Our results suggest genetic admixture between Muslims and upper caste Hindus.

Conclusions

The synthesis of Y-genealogy and estimated diversity of several clades demonstrated that North Indians carry three Y-lineages, one derived from Central Asia or West Eurasia (R1a1[★], R1b1b2[★] and R2 haplogroups), one derived from the Middle East (J2[★], Shia-specific E1b1b1, and to some extent G[★] and L[★] haplogroups), and the indigenous Indian Y-lineage marked by H1[★], F[★], O[★] and C[★] without the M217 transversion. Our data revealed that there may have been admixture between Sunni Muslims and Brahmins in North India. However, a recent study has shown the presence of the YAP + element in lower caste groups, namely Panchamas and Vaishyas of North India (Uttar Pradesh) (Zerjal et al. 2007). It may be postulated that there was admixture between Shia Muslims with both higher and lower caste groups from Uttar Pradesh in the past. Our previous results based on mtDNA analysis (Terrerros et al. 2007) revealed that the two Muslim sects (Shia and Sunni) appeared to lack significant levels of the haplogroups (M2, U2, R5) which are believed to represent the proto-Indians involved in the initial migration out of Africa along the southern Asian coast 60 000–80 000 ybp. This suggests that admixture between Islamic travellers who introduced their religion, culture, and language into Indian groups might have also shaped the social status and geographical proximity of the existing Indian populations. Interestingly, results on both mtDNA and Y chromosome indicate rather similar findings.

Acknowledgements

This research work was supported by a NIH grant (LM009598), the Thomas F. and Kate Miller Jeffress Memorial Trust Fund, Indian Council of Medical Research (ICMR), New Delhi, and Sanjay Gandhi Post Graduate Institute of Medical Sciences, Lucknow, India. The authors would like to thank two anonymous reviewers for their valuable comments.

References

- Agrawal S, Khan F, Pandey A, Tripathi M, Herrera RJ. YAP, signature of an African Middle Eastern migration into northern India. *Curr Sci* 2005;88:174–179.
- Alonso S, Flores C, Cabrera V, Alonso A, Martin P, Albarran C, Izagirre N, de la Rua C, Garcia O. The place of the Basques in the European Y-chromosome diversity landscape. *Eur J Hum Genet* 2005;13:1293–1302. [PubMed: 16094307]
- Balakrishnan V. A preliminary study of genetic distances among some populations of the Indian subcontinent. *J Hum Evol* 1978;7:67–75.
- Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, Naidu JM, Prasad BV, Reddy PG, Rasanayagam A, Papiha SS, Vilems R, Redd AJ, Hammer MF, Nguyen SV, Carroll ML, Batzer MA, Jorde LB. Genetic evidence on the origins of Indian caste populations. *Genome Res* 2001;11:994–1004. [PubMed: 11381027]
- Cann RL. Genetic clues to dispersal in human populations: Retracing the past from the present. *Science* 2001;291:1742–1748. [PubMed: 11249820]

- Cavalli-Sforza LL. The Human Genome Diversity Project: Past, present and future. *Nat Rev Genet* 2005;6:333–340. [PubMed: 15803201]
- Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. *Nat Genet* 2003;33(Suppl):266–275. [PubMed: 12610536]
- Cavalli-Sforza, LL.; Menozzi, P.; Piazza, A. The history and geography of human genes. Princeton: Princeton University Press; 1994.
- Chakraborty, R.; Jin, L. A unified approach to study hypervariable polymorphisms: Statistical considerations of determining relatedness and population distances. In: Pena, SDJ.; Chakraborty, R.; Epplen, JT.; Freys, AJ., editors. *DNA fingerprinting: State of the science*. Switzerland: Birkhauser: Basel; 1993. p. 153–175.
- Cinnioglu C, King R, Kivisild T, Kalfoglu E, Atasoy S, Cavalleri GL, Lillie AS, Roseman CC, Lin AA, Prince K, Oefner PJ, Shen P, Semino O, Cavalli-Sforza LL, Underhill PA. Excavating Y-chromosome haplotype strata in Anatolia. *Hum Genet* 2004;114:127–148. [PubMed: 14586639]
- Comey CT, Koons BW, Presley KW, Smerick JB, Sobieralski CA, Stanley DM, Baechtel FS. DNA extraction strategies for amplified fragment length polymorphism analysis. *J Forensic Sci* 1994;39:1254–1269.
- Cordaux R, Auger R, Bentley G, Nasidze I, Sirajuddin SM, Stoneking M. Independent origins of Indian caste and tribal paternal lineages. *Curr Biol* 2004;14:231–235. [PubMed: 14761656]
- Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V, Wallace DC, Oefner PJ, Torroni A, Cavalli-Sforza LL, Scozzari R, Underhill PA. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet* 2002;70:1197–1214. [PubMed: 11910562]
- Felsenstein, J. Distributed by the author Department of Genome Sciences. Seattle: University of Washington; 2005. PHYLIP (Phylogeny Inference Package) version 3.57c.
- Gutala R, Carvalho-Silva DR, Jin L, Yngvadottir B, Avadhanula V, Nanne K, Singh L, Chakraborty R, Tyler-Smith C. A shared Y-chromosomal heritage between Muslims and Hindus in India. *Hum Genet* 2006;120:543–551. [PubMed: 16951948]
- International Society of Genetic Genealogy. Y-DNA Haplogroup Tree 2006. 2006http://www.isogg.org/tree/Main06.html. Version: 3.06, Date: 11 May 2008.
- Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res* 2008;18:830–838. [PubMed: 18385274]
- Karafet TM, Zegura SL, Posukh O, Osipova L, Bergen A, Long J, Goldman D, Klitz W, Harihar S, de Knijff P, Wiebe V, Griffiths RC, Templeton AR, Hammer MF. Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. *Am J Hum Genet* 1999;64:817–831. [PubMed: 10053017]
- Karve, I. Kinship organization in India. Bombay: Asia Publishing House; 1968.
- Kivisild T, Rootsi S, Metspalu M, Mastana S, Kaldma K, Parik J, Metspalu E, Adojaan M, Tolk HV, Stepanov V, Golge M, Usanga E, Papiha SS, Cinnioglu C, King R, Cavalli-Sforza L, Underhill PA, Villems R. The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am J Hum Genet* 2003;72:313–332. [PubMed: 12536373]
- Lahr MM, Foley RA. Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution. *Am J Phys Anthropol Suppl* 1998;27:137–176.
- Luis JR, Rowold DJ, Regueiro M, Caeiro B, Cinnioglu C, Roseman C, Underhill PA, Cavalli-Sforza LL, Herrera RJ. The Levant versus the Horn of Africa: Evidence for bidirectional corridors of human migrations. *Am J Hum Genet* 2004;74:532–544. [PubMed: 14973781]
- Majumder PP. People of India: Biological diversity and affinities. *Evol Anthropol* 1998;6:100–110.
- Nei M. Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci USA* 1973;70:3321–3323. [PubMed: 4519626]
- Quintana-Murci L, Krausz C, Zerjal T, Sayar SH, Hammer MF, Mehdi SQ, Ayub Q, Qamar R, Mohyuddin A, Radhakrishna U, Jobling MA, Tyler-Smith C, McElreavey K. Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. *Am J Hum Genet* 2001;68:537–542. [PubMed: 11133362]

- Ramana GV, Su B, Jin L, Singh L, Wang N, Underhill P, Chakraborty R. 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* 2001;9:695–700. [PubMed: 11571559]
- Renfrew C. The origins of Indo-European languages. *Sci Am* 1989;261:82–90.
- Sahoo S, Singh A, Himabindu G, Banerjee J, Sitalaximi T, Gaikwad S, Trivedi R, Endicott P, Kivisild T, Metspalu M, Vilems R, Kashyap VK. A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. *Proc Natl Acad Sci USA* 2006;103:843–848. [PubMed: 16415161]
- Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;4:406–425. [PubMed: 3447015]
- Schneider, S.; Roessli, D.; Excoffier, L. Arlequin: A software for population genetics data analysis. Ver 2.000. University of Geneva: Genetics and Biometry Lab, Department of Anthropology; 2000.
- Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Underhill PA. The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: A Y chromosome perspective. *Science* 2000;290:1155–1159. [PubMed: 11073453]
- Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow CE, Lin AA, Mitra M, Sil SK, Ramesh A, Usha Rani MV, Thakur CM, Cavalli-Sforza LL, Majumder PP, Underhill PA. 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* 2006;78:202–221. [PubMed: 16400607]
- Terreros MC, Rowold D, Luis JR, Khan F, Agrawal S, Herrera RJ. North Indian Muslims: Enclaves of foreign DNA or Hindu converts? *Am J Phys Anthropol* 2007;133:1004–1012. [PubMed: 17427927]
- Thanseem I, Thangaraj K, Chaubey G, Singh VK, Bhaskar LV, Reddy BM, Reddy AG, Singh L. Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA. *BMC Genet* 2006;7:42. [PubMed: 16893451]
- The Y Chromosome Consortium. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res* 2002;12:339–348. [PubMed: 11827954]
- Underhill PA, Passarino G, Lin AA, Shen P, Mirazon Lahr M, Foley RA, Oefner PJ, Cavalli-Sforza LL. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet* 2001;65:43–62. [PubMed: 11415522]
- Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ. Y chromosome sequence variation and the history of human populations. *Nat Genet* 2000;26:358–361. [PubMed: 11062480]
- Wells RS, Yuldasheva N, Ruzibakiev R, Underhill PA, Evseeva I, Blue-Smith J, Jin L, Su B, Pitchappan R, Shanmugalakshmi S, Balakrishnan K, Read M, Pearson NM, Zerjal T, Webster MT, Zholoshvili I, Jamarjashvili E, Gambarov S, Nikbin B, Dostiev A, Aknazarov O, Zalloua P, Tsoy I, Kitaev M, Mirrakhimov M, Chariev A, Bodmer WF. The Eurasian heartland: A continental perspective on Y-chromosome diversity. *Proc Natl Acad Sci USA* 2001;98:10244–10249. [PubMed: 11526236]
- Zerjal T, Pandya A, Thangaraj K, Ling EY, Kearley J, Bertoneri S, Paracchini S, Singh L, Tyler-Smith C. Y-chromosomal insights into the genetic impact of the caste system in India. *Hum Genet* 2007;121:137–144. [PubMed: 17075717]
- Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, Qamar R, Ayub Q, Mohyuddin A, Fu S, Li P, Yuldasheva N, Ruzibakiev R, Xu J, Shu Q, Du R, Yang H, Hurles ME, Robinson E, Gerelsaikhan T, Dashnyam B, Mehdi SQ, Tyler-Smith C. The genetic legacy of the Mongols. *Am J Hum Genet* 2003;72:717–721. [PubMed: 12592608]

Figure 1.

Distribution of NRY-binary haplogroups in five North Indian populations. The markers used in this study are shown on each branch. The default haplogroups are marked with asterisks.

Figure 2.

Neighbour-joining tree of the five Indian populations with other 36 world populations. The tree was reconstructed by Nei's genetic distance (Nei 1973) using the PHYLIP package (Felsenstein 2005). The bootstrap values (1000 replicates) are indicated on the branches. The five north Indian populations in this study are located together to the right of the figure – Shia, Sunni, Chaturvedies, Bhargavas, and Brahmins.

Three haplogroups — E — R — F

Figure 3.
Migratory routes of paternal lineages of Indian upper caste and Muslim populations.

Table I

AMOVA analysis based on Y-UEP haplogroup frequencies.

Source of variation	Sum of squares	Variance components	Percentage of variation
Among upper castes	1.00	0.0045	1.08
Between Bhargavas and Muslims	2.07	0.0068	1.54
Between Chaturvedis and Muslims	1.32	0.0018	0.42
Between Brahmins and Muslims	0.67	0.0009	0.18
Between upper castes and Muslims	1.94	0.0037	0.86

Table II

Fst genetic distances between five populations based on 32 Y-UEPs.

	Chaturvedis	Brahmins	Shia	Sunni
Bhargavas	0.0036	0.0034	0.0828	0.0090
Chaturvedis		0.0012	0.0753	0.0034
Brahmins			0.0454	0.0072
Shia				0.0375