Journal of Human Genetics (2009) 54, 340–348
© 2009 The Japan Society of Human Genetics All rights reserved 1434-5161/09 $32.00
www.nature.com/jhg

ORIGINAL ARTICLE

Diverse genetic origin of Indian Muslims: evidence from autosomal STR loci

Muthukrishnan Eaaswarkhanth[1,2], Bhawna Dubey[1], Poorlin Ramakodi Meganathan[1], Zeinab Ravesh[2], Faizan Ahmed Khan[3], Lalji Singh[2], Kumarasamy Thangaraj[2] and Ikramul Haque[1]

The origin and relationships of Indian Muslims is still dubious and are not yet genetically well studied. In the light of historically attested movements into Indian subcontinent during the demic expansion of Islam, the present study aims to substantiate whether it had been accompanied by any gene flow or only a cultural transformation phenomenon. An array of 13 autosomal STR markers that are common in the worldwide data sets was used to explore the genetic diversity of Indian Muslims. The austere endogamy being practiced for several generations was confirmed by the genetic demarcation of each of the six Indian Muslim communities in the phylogenetic assessments for the markers examined. The analyses were further refined by comparison with geographically closest neighboring Hindu religious groups (including several caste and tribal populations) and the populations from Middle East, East Asia and Europe. We found that some of the Muslim populations displayed high level of regional genetic affinity rather than religious affinity. Interestingly, in Dawoodi Bohras (TN and GUJ) and Iranian Shia significant genetic contribution from West Asia, especially Iran (49, 47 and 46%, respectively) was observed. This divulges the existence of Middle Eastern genetic signatures in some of the contemporary Indian Muslim populations. Our study reveals that the spread of Islamic faith in the Indian subcontinent was predominantly cultural transformation associated with minor gene flow from West Asia.

Journal of Human Genetics (2009) **54,** 340–348; doi:10.1038/jhg.2009.38; published online 8 May 2009

Keywords: autosomal STRs; genetic affinity; genetic signatures; Indian Muslims; Middle East

INTRODUCTION

Contemporary ethnic India is an accumulation of several cultures, religions, languages and evolutionary histories. This is attributed to several invasions and exodus from several parts of the world, thus bringing in concert, a great diversity of human genes and human cultures.[1] Similar to any other multicultured society, India offers a cauldron where the process of unification as well as fragmentation is perpetually taking place. Hence, human diversity in India is characterized by 4635 documented communities comprising 2205 major population groups,[2] 461 tribal communities,[3] 20 major languages and approximately 750 dialects.[4] Regardless of contiguous inhabitation for several years, people of India encompass diverse ethnic (Australian, Indo-Caucasian, East Asian and African descent) populations and linguistic groups (Indo-European, Dravidian, Austro-Asiatic and Tibeto-Burman). The majority of these ethnic populations belong to Hindu religious fold, communally regulated into endogamous hierarchical castes and subcastes.[5] In addition, there are known religious groups, namely, Islam, Christianity, Sikhism, Buddhism, Jainism and Judaism. Islam is the second-most practiced religion next to Hinduism with a populace of 138 million.[6]

Islamic influence first came to be felt in the Indian subcontinent during early seventh century with the advent of Arab military forces into Sindh, the lower part of the Indus valley, and included it into the Arabian empire. Subsequently, an Indo-Islamic state was established in Sindh and thereafter the region served as an Islamic outpost where Arabs established their trade links with the Middle East. For the next two and half centuries Islamic presence was hardly felt throughout the subcontinent. By the end of tenth century spectacular changes took place when Turkic tribes embraced and took up the mission of propagating Islam. These assertively expansive invaders initially began to move into Afghanistan and Iran and later into India through the northwest. In the thirteenth century, a Turkic kingdom was established in Delhi, which facilitated Persian and Afghan Muslim invaders to further spread across India. In the following 100 years, the Muslim empire extended its influence east to Bengal and south to the Deccan regions. Muslim sultanates ruling from Delhi, beginning in the eleventh century, and the great Mughal Empire (1526–1707 CE) that followed created a substantive Islamic legacy before India fell under British colonial rule.[7–8] Thus the emergence of Islam in the region is concurrent with the Turko-Muslim invasion of medieval India (which includes large parts of present day Pakistan and India), where these

[1]National DNA Analysis Centre, Central Forensic Science Laboratory, Kolkata, India; [2]Centre for Cellular and Molecular Biology, Hyderabad, India and [3]State Forensic Science Laboratory, Lucknow, India
Correspondence: Dr K Thangaraj, Centre for Cellular and Molecular biology, Hyderabad 500 007, India.
E-mail: thangs@ccmb.res.in or Dr I Haque, National DNA Analysis Centre, Central Forensic Science Laboratory, 30-Gorachand Road, Kolkata 700 014, India.
E-mail: haque_cfslk@yahoo.co.in
Received 6 January 2009; revised 28 March 2009; accepted 8 April 2009; published online 8 May 2009

rulers took over the administration of large parts of Indian subcontinent. Since its introduction into Indian subcontinent, Islam has made significant religious, artistic, philosophical, cultural, social and political influences to Indian history. Muslim traders, mystics, preachers and invaders have shaped and influenced Indian subcontinent for thirteen centuries ensuing significant cultural diffusion of Muslim traditions among the ethnic Indian populations.[7–10]

According to the historical facts, present-day Indian Muslims may perhaps be either the descents of local Hindu converts or the descendants of Iranian and Arabian men who married local Hindu women, possibly during the historical period of Muslim rulers in the past.[8,10–11] Presently, India is the largest Muslim populated country, after Indonesia constituting 13.4% of the total Indian population.[6] As any other Muslims in the world, Indian Muslims also follow two sects, Shia and Sunni, practicing strict endogamy.[12–13]

Consistent with the above historical and anthropological facts, classical markers data[11–12,14] coupled with molecular marker studies[15–18] suggested the subsistence of high level of regional genetic affinity in some of the north and south Indian Muslim populations. In contrast, a couple of classical marker studies have shown that some of the northern and north western Muslim populations genetically vary from the aboriginal Indian Hindu populations.[14,19] Even though, a few biparental marker studies[16,18] have been reported, a study covering different geographical regions with large sample size based on a subset of 13 CODIS core autosomal STR loci is altogether missing for Indian Muslim populations. In this context, no comprehensive study is available to explore the genetic relationships of Indian Muslims with Indian non-Muslims and the world populations with regard to the historically attested movements. Therefore, to explore the genetic diversity of Indian Muslims and to test population affinities and the level of admixture, we analyzed six Muslim populations belonging to three different geographical regions of India (Figure 1) with a battery of 13 biparental markers (commonly available autosomal STR markers in the worldwide data sets).

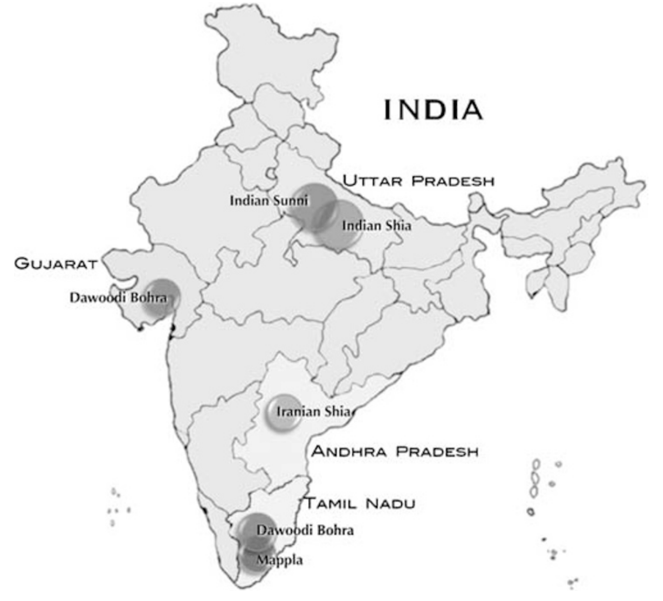

Figure 1 Map of India showing the geographical location of the six Indian Muslim populations included in this study.

MATERIALS AND METHODS

Samples

Blood samples were obtained from 477 unrelated healthy individuals belonging to six Muslim communities: Dawoodi Bohra (TN) (n=62), Mappla (n=62), Indian Shia (n=121), Indian Sunni (n=132), Dawoodi Bohra (GUJ) (n=50) and Iranian Shia (n=50) of three different geographical regions of India (see Figure 1) after informed consent was acquired. Ethical guidelines were followed as stipulated by both the institutions involved in this study.

To determine population affinities and genetic admixtures, we compared the STR diversity of Indian Muslims with the published allele frequencies data[20–49] of geographically targeted population groups listed in Supplementary information (SI) (Supplementary Table S1) based on the subset of 13 autosomal STR markers (common in the set of 15 STR markers).

DNA isolation and STR typing

Human genomic DNA was extracted using the standard Phenol-Chloroform procedure[50] and purified by ethanol precipitation. Multiplex PCR amplification was performed for 15 STR loci (D8S1179, D21S11, D7S820, CSF1PO, D19S433, vWA, TPOX, D18S51, D3S1358, THO1, D13S317, D16S539, D2S1338, D5S818 and FGA) using the AmpFlSTR Identifiler PCR Amplification Kit (Applied Biosystems, Foster City, CA, USA), according to the user's manual specifications[51] (Applied Biosystems). PCR products were genotyped using ABI 3100 automated DNA sequencer (Applied Biosystems). GeneScan 3.7 analysis software was used to determine the allele fragment size. Genotyper software 3.7 NT was used to designate alleles by comparison with the allelic ladder supplied with the kit.

Statistical analyses

Allele frequencies of the 15 STR loci were calculated using Genepop Version 3.4.[52] Arlequin software package Version 3.1[53] was used to determine Hardy–Weinberg P-values, observed heterozygosity (H_o), expected heterozygosity (He) and gene diversity (GD) values. Locus-wise GD and polymorphic information content were computed using Power Marker V3.25.[54] The autosomal STR markers used in this study are being widely used for forensic DNA typing and so several parameters of forensic and population genetic importance such as matching probability (pM), power of discrimination (PD), power of exclusion (PE) and paternity index (PI) were calculated using Excel PowerStats spreadsheet (www.promega.com/geneticidtools/powerstats). Genetic structuring among Indian Muslims was determined through hierarchical analysis of molecular variance by grouping them according to their sect (Shia and Sunni) and geographic location (north, west and south) using Arlequin Version 3.1.[53]

The allele frequencies of the 13 STR markers for the complied populations (Supplementary Table S1) were prearranged and used to generate NJ trees based on Nei's D_A genetic distances using various options in the PHYLIP 3.6 software.[55] Initially, bootstrapping of STR alleles was carried out with 1000 replications by the option SEQBOOT, and then the GENDIST option was used to calculate Nei's D_A genetic distances for the acquired multiple data sets of 1000 bootstrap matrices. The resultant genetic distances matrices were applied to create 1000 possible phylogenetic trees by the option NEGHBOR, while the best-fit tree (the consensus of 1000 trees) was generated using CONSENSE programs in PHYLIP 3.6 software.[55] The principal component analyses (PCA) was executed with the Nei's D_A genetic distances between pairs of populations and were plotted in a three-dimensional representations using Statistical Package for the Social Sciences (SPSS) software version 12.0.1.[56]

Various components of genetic variance, such as coefficient of gene differentiation (Gst), total GD (Ht) and GD within population (Hs) were calculated using DISPAN software.[57] Locus-wise exact test of population differentiation was performed using Arlequin Version 3.1[53] to analyze the extent of genetic diversity among the Indian Muslims and their geographically closest neighboring Hindu (including several caste and tribal populations) populations at the analyzed 13 STR loci.

ADMIX95 software based on gene identity method[58] was used to estimate the admixture proportions (mean ± s.e.) of Indian Muslim populations. 'Admixture tests reveal the amount of genetic contributions of putative parental populations to the gene pool of the collection being characterized; that is, the hybrid population. The contribution of a parental population in the gene pool

Table 1 Summary of Statistical parameters of population genetics interest

Population	*Total alleles*	*CMP*[a]	*CPE*	*CPD*	*Avg. He*	*HWE departures*[b]	*GD*
Dawoodi Bohra (TN)	150	8.182×10^{15}	0.999994886	[c]	0.770	D7S820, CSF1PO, TPOX, D13S317, D16S539, D2S1338 and FGA	0.799 ± 0.403
Mappla	157	7.196×10^{16}	0.999997863	[c]	0.780	FGA	0.814 ± 0.410
Indian Shia	142	1.137×10^{17}	0.999990906	[c]	0.752	D18S51 and FGA	0.783 ± 0.397
Indian Sunni	147	2.886×10^{17}	0.999996670	[c]	0.777	D18S51 and FGA	0.805 ± 0.404
Dawoodi Bohra (GUJ)	121	4.556×10^{15}	0.999997315	[c]	0.759	CSF1PO and FGA	0.799 ± 0.403
Iranian Shia	117	5.172×10^{15}	0.999969927	[c]	0.703	D18S51, D2S1338 and FGA	0.787 ± 0.398

Abbreviations: Avg. He, average heterozygosity; CMP, combined matching probability; CPE, combined probability of exclusion; CPD, combined probability of discrimination; GD, gene diversity; HWE, Hardy–Weinberg equilibrium.
[a] Expressed as 1 in....
[b] Indicates HWE departures after Bonferroni-like adjustment for number of loci tested (α: 0.05/15=0.0033).
[c] Indicates each of the six populations generated a CPD $>$0.999999999999999.

of a hybrid population, which arose by hybridization with one or more other populations, is estimated at the population level from the probability of gene identity, if there has been no genetic drift. However, they may not only reflect gene flow between the parent and hybrid populations but shared ancestry given that the test compares frequencies and allelic distributions to reach its conclusions. The dynamics of accumulation of such admixture is studied incorporating the fluctuations due to finite size of the hybrid population. All the tests assume that the admixture is a static, one-time phenomenon, whereas in reality there is usually a continued, long-term exchange of genes among populations.'[58] 'One of the major impediments to admixture estimations in humans has been the lack of accurate identities of parental populations of a hybrid population. Therefore, unless identities of, and allele frequencies in, parental populations are known accurately, estimates of admixture may be quite unreliable.'[59] Considering these assumptions and limitations, we chose three putative parental populations including (i) the geographically closest Indian Hindu population, and a pool of populations from (ii) Arabia and (iii) Iran. The parental populations, Arabia and Iran, were known to have historical significance pertaining to Indian Muslims.[8,10–11] Moreover, the CODIS core 13 STR markers included in this study were established to be informative that makes them useful for admixture analyses.[60]

RESULTS

STR diversity of Indian Muslims

Allelic frequency distributions of the 15 STR loci for Dawoodi Bohra (GUJ) and Iranian Shia Muslims are given as Supplementary materials (Supplementary Tables S2 and S3, respectively). The population-wise descriptive statistics for six Muslim populations are given in Supplementary Table S4. The combined matching probability ranges from 1 in 4.556×10^{15} in Dawoodi Bohra (GUJ) to 1 in 2.886×10^{17} in Indian Sunni Muslims (Table 1). The combined power of exclusion is 0.9999 in all studied populations (Table 1). In each of the six Muslim populations the combined power of discrimination value is $>$0.9999 (Table 1). The GD ranges from 0.783 ± 0.397 in Indian Shia to 0.814 ± 0.410 in Mappla Muslims (Table 1). The average heterozygosity values were found with a narrow range between 0.703 in Iranian Shia and 0.780 in Mappla Muslims (Table 1). The loci in each population that are found to depart from Hardy–Weinberg equilibrium even after applying Bonferroni adjustment for number of loci tested (α=0.05/15 or 0.0033) are listed in Table 1.

The phylogenetic relationship among the Indian Muslims was inferred by the neighbor-joining (NJ) method with Nei's D_A genetic distance. The NJ tree (Supplementary Figure S1a) explicitly portrayed the segregation of almost all the six Muslim populations with varied range of bootstrap values. In the PCA plot (Supplementary Figure S1b), it is notable that Dawoodi Bohra (TN) and Dawoodi Bohra (GUJ) are placed more distantly than in the NJ tree and they are separated out from the other studied groups. However, similar segregation pattern was observed in both PCA plot (Supplementary Figure S1b) and NJ tree (Supplementary Figure S1a). When the genetic structure among Indian Muslims was examined, the analysis of molecular variance yielded no statistically significant results for any group distinctions based on sect (Shia and Sunni), geography (north India, south India and west India) or other criteria investigated (Supplementary Table S5).

Inter-population STR diversity

Owing to the lack of population data, including entire set of 15 STRs, the comparative analyses were performed with the commonly available 13 CODIS core STR loci from the worldwide data set. With this subset of 13 STR markers, we analyzed relationships among Indian Muslims, non-Muslims and populations from Middle East, Europe and East Asia.

Genetic affinities and admixture of Indian Muslims with other populations

Indian non-Muslim populations. To test for any genetic differences of statistical significance, exact test of population differentiation was performed between Indian Muslim populations and their corresponding geographically closest neighboring Hindu religious (including several caste and tribal populations) populations. This pair-wise comparison test values are tabulated in Table 2. Indian Shia showed significant differences with Thakur at 10 of 13 loci and with Khatri, Kurmi and Thakur at locus D5S818. A similar trend was observed for Indian Sunni as well that significantly differed at 8 of 13 loci with Thakur and at loci D5S818 and FGA with Khatri, Kurmi and Thakur. Remarkably, Dawoodi Bohra (GUJ) showed most significant difference with both of their geographically closest neighboring populations, Patel and Gujarati at 12 of 13 loci, whereas Iranian Shia was found with significant difference from Andhra Brahmins, Kappu Naidu, Raju, Kamma Chaudhary and Kappu Reddy at TPOX and also from Andhra Brahmin, Komati, Raju and Kamma Chaudhary at locus D13S317. Dawoodi Bohra (TN) showed significant differences with their neighboring populations at 11 of 13 loci particularly the locus FGA showed significant differences with all their geographically closest neighboring populations, likewise locus CSF1PO also showed significant differences with all neighboring populations except Paraiyar. Mappla showed least significant difference with all their

Table 2 Population differentiation tests between Indian Muslims and their neighboring non-Muslim populations based on 13 autosomal STR markers

	D8S1179	*D21S11*	*D7S820*	*CSF1PO*	*vWA*	*TPOX*	*D18S51*	*D3S1358*	*THO1*	*D13S317*	*D16S539*	*D5S818*	*FGA*
Indian Shia vs													
Khatri	0.74821	0.15862	0.54241	0.18393	0.01083	0.00000	0.25777	0.11754	0.32418	0.17306	0.98280	0.00000	0.00343
Kurmi	0.07687	0.06649	0.63075	0.93698	0.69965	0.42350	0.10746	0.76314	0.01322	0.39702	0.52767	0.00142	0.08898
Thakur	0.00797	0.00000	0.03619	0.06902	0.00000	0.02651	0.01259	0.08125	0.00433	0.03723	0.82955	0.00000	0.01994
Kanyakubj Brahmin	0.59289	0.76342	0.91652	0.07743	0.18986	0.30423	0.00070	0.64707	0.74489	0.00013	0.67912	0.12111	0.06661
Teli	0.00009	0.52946	0.38192	0.60341	0.12421	0.24867	0.43592	0.54153	0.27746	0.26585	0.62416	0.25141	0.15707
Yadav	0.43547	0.30286	0.64626	0.66239	0.45538	0.25904	0.17320	0.95868	0.40356	0.39858	0.83632	0.38636	0.15130
Indian Sunni vs	*D8S1179*	*D21S11*	*D7S820*	*CSF1PO*	*vWA*	*TPOX*	*D18S51*	*D3S1358*	*THO1*	*D13S317*	*D16S539*	*D5S818*	*FGA*
Khatri	0.73303	0.09024	0.25096	0.34389	0.01348	0.00000	0.46770	0.11240	0.16198	0.30175	0.97208	0.00000	0.00014
Kurmi	0.17240	0.12022	0.15664	0.91086	0.37751	0.78098	0.20700	0.46075	0.01214	0.24673	0.23742	0.00041	0.01975
Thakur	0.07801	0.00014	0.02896	0.01598	0.00000	0.13696	0.21037	0.15041	0.00452	0.02128	0.73789	0.00000	0.00277
Kanyakubj Brahmin	0.26642	0.99059	0.47863	0.11144	0.38511	0.37434	0.00051	0.01420	0.93497	0.00075	0.24638	0.09869	0.87455
Teli	0.00552	0.68632	0.10122	0.77736	0.15914	0.73673	0.04586	0.06584	0.09052	0.30334	0.56185	0.54141	0.24198
Yadav	0.96134	0.88521	0.28563	0.84709	0.87546	0.39446	0.19561	0.51950	0.53162	0.48761	0.69281	0.77679	0.50069
Dawoodi Bohra (GUJ) vs	*D8S1179*	*D21S11*	*D7S820*	*CSF1PO*	*vWA*	*TPOX*	*D18S51*	*D3S1358*	*THO1*	*D13S317*	*D16S539*	*D5S818*	*FGA*
Patel	0.00000	0.00000	0.00000	0.00000	0.00000	0.00006	0.00000	0.00000	0.00000	0.08891	0.00002	0.00292	0.00000
Gujarati	0.00000	0.00000	0.00036	0.00000	0.00000	0.00000	0.00000	0.00208	0.00000	0.04063	0.00363	0.62406	0.00000
Iranian Shia vs	D8S1179	D21S11	D7S820	CSF1PO	vWA	TPOX	D18S51	D3S1358	THO1	D13S317	D16S539	D5S818	FGA
Andhra Brahmin	0.58705	0.00000	0.01260	0.17746	0.63470	0.00409	0.05655	0.04618	0.17978	0.00852	0.71400	0.19521	0.17511
Komati	0.15036	0.00625	0.09955	0.09957	0.00954	0.15219	0.10765	0.01947	0.08229	0.00093	0.14220	0.16444	0.09563
Kappu Naidu	0.23994	0.73302	0.42587	0.11652	0.54213	0.01947	0.08646	0.25231	0.38739	0.07990	0.47940	0.09976	0.18496
Raju	0.00008	0.33904	0.02455	0.21061	0.17182	0.00182	0.08485	0.31460	0.13330	0.02179	0.57109	0.86943	0.06679
Kamma Chaudhary	0.01656	0.84851	0.16310	0.16494	0.25357	0.00461	0.12423	0.16532	0.18906	0.02268	0.25813	0.07254	0.44803
Kapu Reddy	0.03582	0.93386	0.03575	0.06061	0.30311	0.00059	0.12945	0.08268	0.22259	0.12946	0.79164	0.59135	0.02251
Dawoodi Bohra (TN) vs	*D8S1179*	*D21S11*	*D7S820*	*CSF1PO*	*vWA*	*TPOX*	*D18S51*	*D3S1358*	*THO1*	*D13S317*	*D16S539*	*D5S818*	*FGA*
Gounder	0.20535	0.24371	0.11099	0.00030	0.02150	0.00009	0.35531	0.20563	0.69349	0.04747	0.02173	0.57085	0.00000
Irular	0.10509	0.05258	0.00000	0.00169	0.22931	0.00000	0.38886	0.26783	0.01792	0.00000	0.00167	0.05036	0.00000
Chakkiliyar	0.44924	0.30957	0.15196	0.00149	0.05262	0.01418	0.82636	0.00399	0.02960	0.05549	0.02238	0.04136	0.00103
Tanjore Kallar	0.26633	0.41738	0.00022	0.00000	0.01276	0.10236	0.01013	0.01248	0.28319	0.00000	0.00027	0.31110	0.00000
Vanniyar	0.21346	0.41515	0.02103	0.00000	0.10850	0.02312	0.12193	0.09012	0.43823	0.00007	0.00138	0.09747	0.00000
Pallar	0.18556	0.94418	0.34327	0.03601	0.00514	0.02402	0.38993	0.75334	0.14068	0.15731	0.24600	0.40093	0.00000
Paraiyar	0.70956	0.70833	0.50989	0.34531	0.20704	0.23654	0.63572	0.22933	0.84724	0.93544	0.45478	0.68000	0.03509

Table 2 Continued

	D8S1179	D21S11	D7S820	CSF1PO	vWA	TPOX	D18S51	D3S1358	THO1	D13S317	D16S539	D5S818	FGA
Mappla vs													
Gounder	0.55697	0.96954	0.35886	0.27677	0.00840	0.04344	0.71225	0.72800	0.45467	0.52518	0.81166	0.67325	0.00000
Irular	0.48094	0.21454	0.03354	0.40951	0.40841	0.00606	0.17859	0.87152	0.35821	0.00113	0.01035	0.55612	0.18057
Chakkiliyar	0.58838	0.26038	0.50813	0.28108	0.20045	0.66895	0.80147	0.24997	0.35789	0.74181	0.84785	0.59454	0.71952
Tanjore Kallar	0.56729	0.69853	0.00495	0.23731	0.25379	0.76252	0.51830	0.31234	0.92067	0.00270	0.48881	0.87413	0.07311
Vanniyar	0.87268	0.86476	0.44933	0.53294	0.62360	0.68187	0.68395	0.71318	0.84602	0.05415	0.75272	0.52720	0.23079
Pallar	0.91532	0.64216	0.36069	0.37603	0.11264	0.27914	0.99555	0.86613	0.51712	0.63953	0.88456	0.90229	0.19655
Paraiyar	0.91080	0.84167	0.63913	0.97859	1.00000	0.73391	0.88367	0.50038	0.93577	0.98646	0.99977	0.63658	0.99261

The significant differences observed between the studied and their neighboring populations at various loci are underlined.

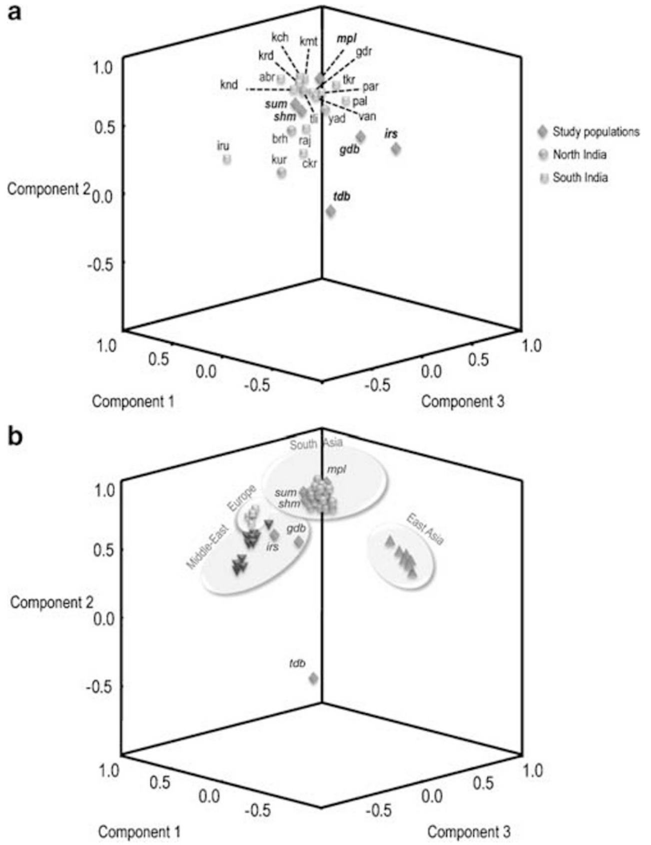

Figure 2 (**a**) Three-dimensional PCA plot depicting relationships between Indian Muslims and non-Muslims based on Nei's D_A genetic distances generated from the 13 CODIS core STR markers. Population codes: shm, Indian Shia; sum, Indian Sunni; kur, Kurmi; brh, Kanyakubj Brahmin; tli, Teli; yad, Yadav; gdb, Dawoodi Bohra (GUJ); irs, Iranian Shia; abr, Andhra Brahmin; kmt, Komati; knd, Kappu Naidu; raj, Raju; kch, Kamma Chaudhary; krd, Kapu Reddy; tdb, Dawoodi Bohra (TN); mpl, Mappla; gdr, Gounder; iru, Irular; ckr, Chakkiliyar; tkr, Tanjore Kallar; van, Vanniyar; pal, Pallar; par, Paraiyar. (**b**) Three-dimensional PCA plot portraying relationships between Indian Muslims, non-Muslims and populations from Middle East, East Asia and Europe. Population codes: shm, Indian Shia, sum, Indian Sunni, gdb, Dawoodi Bohra (GUJ), irs, Iranian Shia, tdb, Dawoodi Bohra (TN) and mpl, Mappla.

neighboring populations except Irular tribe at loci D7S820, TPOX, D13S317 and D16S539.

To ascertain the genetic relationships between Indian Muslims and non-Muslims, NJ tree was generated based on Nei's genetic distance (D_A). The phylogenetic relationships of the Indian Muslims and non-Muslims, inferred from the STR diversity, are shown in Supplementary Figure S2. In the consensus NJ tree, clusters depicting geographic regions (south, north and west) were observed. Exceptions were Yadav, Teli and Kurmi populations from northern region of India located among the south Indian cluster. The two groups from the western India (Gujarati and Patel) and northern India (Khatri and Thakur) bifurcated separately to the extreme bounds (bootstrap value=100 and 79%, respectively) from Kanyakubj Brahmin with a bootstrap value of 25%. Worthwhile stating was the cluster formed by the five Muslim populations (Indian Shia, Indian Sunni, Dawoodi Bohra (TN), Dawoodi Bohra (GUJ) and Iranian Shia) in the proximity of likely north Indian cluster. Within this cluster, Indian Shia and Indian Sunni bifurcated (bootstrap=15%) together initially and then Dawoodi Bohra (TN) and Dawoodi Bohra (GUJ) set apart from

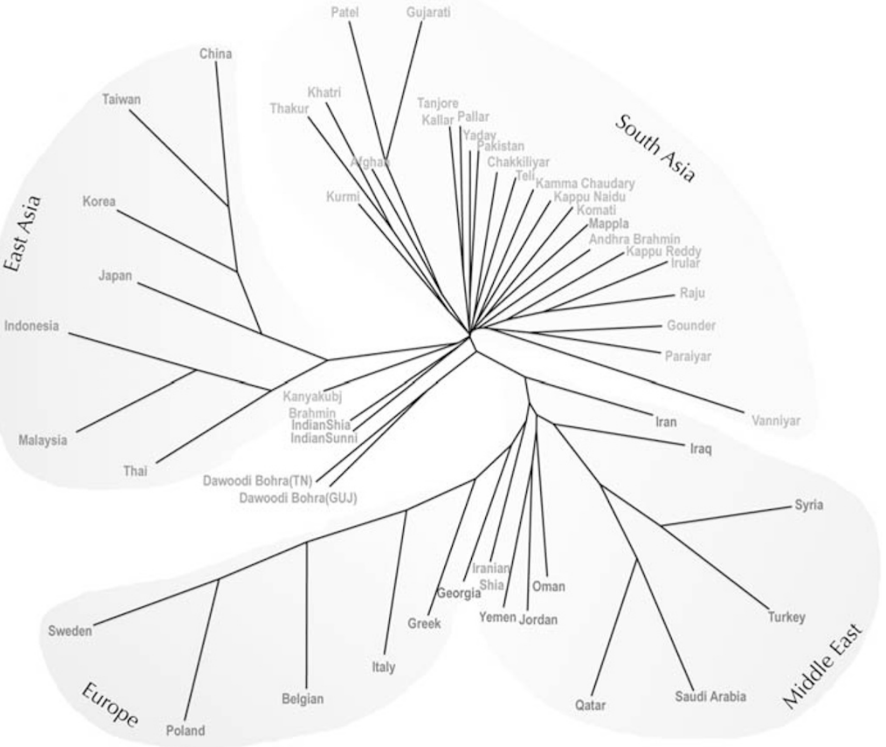

Figure 3 Neighbor-joining (NJ) tree based on Nei's D_A genetic distances between Indian Muslims, non-Muslims and populations from Middle East, East Asia and Europe generated from the 13 CODIS core STR markers.

each other with a 67% incidence, whereas Iranian Shia split (bootstrap=40%) midway between these two Muslim groups. As expected, Mappla Muslim was located in the south Indian population cluster.

The PCA plot displayed the D_A genetic distances between populations in three-dimensional space (Figure 2a) consistent with the overall topology of the NJ tree. In the PC analysis, to get an apparent picture of populations cluster, two populations from western India (Gujarati and Patel) and two populations from northern India (Thakur and Khatri) were excluded as they set distant from other Indian populations in the combined plot. The Iranian Shia, Dawoodi Bohra (TN) and Dawoodi Bohra (GUJ) were outliers, whereas Indian Shia and Indian Sunni grouped in the close proximity of Kanyakubj Brahmin and Mappla lie in the south Indian population cluster. Therefore, the PCA plot essentially mirrored the STR diversity pattern represented in the NJ dendrogram.

Populations of Middle East, Europe and East Asia. Genetic variance for Indian Muslims, Indian non-Muslims and populations from Middle East, Europe and East Asia was tested with the subset of 13 STR loci. The coefficient of gene differentiation (*Gst*), total GD (*Ht*) and GD within population (*Hs*) for overall 13 loci are 0.028, 0.807 and 0.785, respectively (Supplementary Table S6).

Phylogenetic analyses were conducted on the published worldwide data[20–49] based on Nei's D_A genetic distance. The topological arrangement of the populations within the worldwide consensus NJ tree (Figure 3) followed partitioning along the major geographical lines, as expected. Four distinct groups of populations can be identified, namely the South Asia, Middle East, East Asia and Europe clusters (Figure 3). Interestingly, Dawoodi Bohra (TN) and Dawoodi Bohra (GUJ) bifurcated (bootstrap value=46%) in central position to the close proximity of Middle East and European population cluster, whereas Indian Shia and Indian Sunni bifurcated in the opposing end of the South Asian population cluster with a low bootstrap value of 11% and the adjoining Kanyakubj Brahmin split from the South Asian cluster next to the East Asian population group with a low incidence of 10%. Notably, Iranian Shia positioned in the Middle Eastern population cluster in the close propinquity of Georgia and European population cluster. We observed a strong parallelism existing between the NJ dendrogram (Figure 3) and the three-dimensional PCA plot (Figure 2b). Figure 2b depicted clearly, four distinct clusters representing South Asia, Europe, Middle East and East Asia. The placements of Iranian Shia among the Middle East group, Dawoodi Bohras (GUJ and TN) near the Middle Eastern and European group, Mappla among the South Asian group, Indian Shia and Indian Sunni near the South Asian group mirrored unequivocally the phylogenetic relationships derived from the NJ method.

The genetic admixture estimates from the weighted least squares by gene identity method for Indian Muslim populations are shown in Table 3. On the basis of allele frequencies, we calculated the admixture proportions with three putative parental populations, including (i) the geographically closest Indian Hindu religious (including several caste and tribal populations) populations, and a pool of populations from (ii) Arabia and (iii) Iran. The admixture proportions varied markedly among populations (Supplementary Figure S3). In Indian Shia, Indian Sunni, and Dawoodi Bohra (TN) Muslims 50–56% contributions were from the closest Hindu parental populations, while 44–49% from Iranian and 0–3% from Arabian gene pools with correlation coefficients (R^2) of 0.922, 0.997 and 0.867, respectively (Table 3). Exceptionally, Dawoodi Bohra (GUJ) and Iranian Shia Muslims showed major contributions (47 and 46%, respectively) from Iran followed by

Table 3 Admixture proportions (±s.e.) based on 13 autosomal STR markers

	Local Indian closest neighbors						
Admixed\Parental	*Uttar Pradesh*	*Gujarat*	*Andhra Pradesh*	*Tamil Nadu*	*Iran*	*Arabia*	*R^2*
Indian Shia	0.496±0.082				0.473±0.080	0.031±0.085	0.922
Indian Sunni	0.563±0.026				0.441±0.023	−0.004±0.018	0.997
Dawoodi Bohra (TN)				0.503±0.076	0.492±0.159	0.005±0.154	0.867
Dawoodi Bohra (GUJ)		0.226±0.020			0.468±0.050	0.306±0.053	0.949
Iranian Shia			0.115±0.033		0.462±0.049	0.423±0.028	0.979
Mappla				0.765±0.122	0.151±0.164	0.084±0.091	0.939

R^2, Correlation coefficient of admixture contributions from the parental populations.

Arabia (30 and 42%, respectively), whereas there was least contribution from the closest Hindu parental populations (23 and 12%, respectively) with significant R^2 values of 0.949 and 0.979, respectively. Reflecting the phylogenetic analyses, Mappla Muslims acquire major contribution from the closest Hindu parental population's gene pool (77%) and relatively lower proportions from Iran (15%) and Arabia (8%) with R^2=0.939.

DISCUSSION

A number of molecular genetic marker studies reported, so far, with contemporary Indian populations focused only on geographic, ethnic and linguistic affiliations.[61–66] This study throws light on religious grounds for clear understanding of genetic diversity patterns in Indian Muslims. According to historians and anthropologists, the augmentation of Islamic faith in India could be through two discrete ways (i) military invasions that flourished Muslim kingdoms and subsequent migration of mercenaries, businessmen and political emissaries from Middle Eastern countries, Iran and Arabia followed by admixture with the local population; (ii) cultural diffusion as a result of absorption and dominance that resulted in a sizeable population embracing Islam.[7–10] In this perspective, genetic status and relationships of Indian Muslims distributed throughout the country are not yet well studied. We therefore sought out to study a compilation of six Muslim populations from three different geographical regions of India (Figure 1) that witnessed several conquests and immigrations.[7–10] For this we chose a battery of 13 autosomal STR markers that are commonly available for several worldwide populations.

One of the prominent facets observed from the microsatellite diversity analyses in each of the six Indian Muslim populations is the genetic singularity that subsisted at least for the set of microsatellite markers analyzed in this study. Remarkably, of 15 STR loci analyzed, 11 loci in Dawoodi Bohra (TN), Mappla, Indian Shia and Indian Sunni, 9 loci in Dawoodi Bohra (GUJ) and 12 loci in Iranian Shia showed lower observed heterozygosity (*Ho*) values than the expected heterozygosity (H_e) values (Supplementary Table S4). This is the indicative of a relevant heterozygosity deficit. Consanguineous marriage is widely practiced among Indian Muslims. It is conspicuous that such marriage preference varies by sect, that is, Sunni (60.9%) versus Shia (43.4%) versus Dawoodi Bohra (41.1%) and according to clan affiliation.[12–13] The scientific fact that homozygosity at genetic loci increases distinctly in populations practicing consanguinity seem to be reflected in the comparatively low observed heterozygosity (*Ho*) values for most of the STR loci analyzed in six Indian Muslim populations (Supplementary Table S4). Thus, the departures of Hardy–Weinberg equilibrium expectations detected in each of the six Muslim populations appear to be due to excess of homozygotes over heterzygotes, which most likely to be the consequence of high consanguinity rates reported for this populations.[12–13]

The austere endogamy revealed by the STR diversity in each of the Indian Muslim populations mirrored in the phylogenetic analyses. In the three-dimensional PCA plot and NJ dendrogram (Supplementary Figures S1a and S1b), all the six Muslim populations showed distinct separations from each other, reflecting a high level of genetic separation. Exceptionally, Shia and Sunni populations from northern region of India jointly isolated from other Muslims, suggesting genetic affinity among them.

To increase the resolution of genetic analyses of Indian Muslims, we performed inter-population comparisons, first with geographically closest Hindu religious (including several caste and tribal populations) populations and second, populations from Middle East, Europe and East Asia.

In the NJ tree (Supplementary Figure S2) derived from the genetic distances between Indian Muslims and their geographically closest Hindu religious (including several caste and tribal populations) populations, all the Muslim populations except Mappla clustered out separately with a couple of bifurcations (one in the beginning of the branch with Indian Shia-Indian Sunni and the second in the other end with Dawoodi Bohras TN-GUJ) and a split (between these two groups with Iranian Shia) in the close vicinity of north and west Indian branches. In agreement with our recent reports,[18] Mappla Muslims displayed close genetic affinity with the south Indian populations. The regional inter-population relationships in the PCA analysis reiterated most of the depictions inferred from the regional NJ tree; that is, downward diagonal dispersion of Dawoodi Bohra of Gujarat and Tamil Nadu, isolation of Iranian Shia in the right transverse position of Dawoodi Bohra (GUJ), location of grouped Indian Shia and Indian Sunni close to Kanyakubj Brahmins and placement of Mappla in the South Indian cluster (Figure 2a). This was also evident from the pair-wise genetic differentiation tests, in which Dawoodi Bohras (TN and GUJ) and Iranian Shia showed most significant genetic difference from their geographically close Hindu religious (including several caste and tribal populations) populations analyzed, whereas this significant genetic differentiation was least between Mappla, Indian Shia, Indian Sunni and their corresponding pairs of analyzed populations (Table 2). Coupling of Indian Shia and Indian Sunni close to Kanyakubj Brahmins was in congruent with previous reports based on mtDNA[17] and the separate set of microsatellite data analyses.[16]

The array of populations in the worldwide NJ tree (Figure 3) and three-dimensional PCA plot (Figure 2b) exemplified Dawoodi Bohras (TN and GUJ) in an intermediary genetic position, especially relative to the Middle Eastern and European population clusters. As expected, Iranian Shia (recent immigrants from Iran) joined the Middle East population cluster. Likewise in the regional phylogeny, Mappla was found in the South Asian cluster, especially among the south Indian populations. Admixture estimates strongly supports three emerging scenarios from the overall microsatellite diversity based phylogenetic outcome: (i) comparatively high level of genetic contributions from Iran to Iranian Shia and Dawoodi Bohra (GUJ) (ii) almost parallel contributions from local neighboring Hindu populations and Iran populations to Dawoodi Bohra (TN), Indian Shia and Indian Sunni (iii) major genetic contributions from local neighboring Hindu religious (including several caste and tribal populations) populations to Mappla Muslims. There is notable genetic variation between different Indian Muslim populations, some being very similar to local Indian populations and others being similar to outside populations, so that when they are all grouped according to their sect and geographical location in analysis of molecular variance (AMOVA) analyses, group difference was statistically insignificant (Supplementary Table S5).

Summarizing, the findings of our extensive analyses on population affinities and admixture contributions showed the traceable level of gene flow from Western Asia into India in congruent with the earlier reports.[61,67–68] In per view of our present study, attention-grabbing feature is the genetic signals of Middle East in some of the contemporary Indian Muslims. This could be attributed to various Islamic advents from Middle East, particularly Iran and Arabia, during the expansion of Islamic faith into the Indian subcontinent. Further, deep insight into the Middle East genetic signatures in Indian Muslims could be evolved from Y-chromosome (paternal) and mtDNA (maternal) markers.

ACKNOWLEDGEMENTS

We thank all the participants of this study. We thank Mustafa Fakruddin Viramgamwala for his help in sample collection. ME, PRM and BD thank Directorate of Forensic Science, Ministry of Home Affairs, Government of India for the fellowship.

1 Gadgil, M., Joshi, N. V., Shambu, Prasad U. V., Manoharan, S. & Suresh, Patil in The Indian Human Heritage (eds Balasubramanian, D. & Appaji Rao, N.) 100–129 (Universities Press, Hyderabad, India, 1997).
2 Singh, K. S. India's Communities. People of India National Series Vol IV (Oxford University Press, India, 1998).
3 Singh, K. S. *People of India: An introduction* (Anthropological Survey of India, Calcutta, India,, 1992).
4 Kosambi, D. D. *The culture and civilization of ancient India in historical outline* (Vikas Publishing House, New Delhi, 1991).
5 Majumder, P. P. People of India: biological diversity and affinities. *Evol. Anthrop.* **6,** 100–110 (1998).
6 Census of India, Census 2001 [http://www.censusindia.net/].
7 Schimmel, A. *Islam in India and Pakistan* (Brill Academic Publishers, Leiden, 1982).
8 Robb, P. *A History of India* (Palgrave Macmillan Publishers, Hampshire, England, 2002).
9 Naqvi, S. *The Iranian Afaquies contributions to the Qutub Shahi and Adil Shahi Kingdoms* (Hussain Book Shop, Hyderabad, India, 2003).
10 Shokoohy, M. *Muslim architecture of South India: the sultanate of Ma'bar and the traditions of the maritime settlers on the Malabar and Coromandel coasts (Tamil Nadu, Kerala and Goa)* (Routledge Curzon, New York, 2003).
11 Papiha, S. S. Genetic variation in India. *Hum. Biol.* **68,** 607–628 (1996).
12 Afzal, M. *Effects of consanguinity on reproductive fitness and certain behavioural traits among Bihar Muslims* (Ph.D. thesis, Bhagalpur University, Bihar, 1984).
13 Bittles, A. H. & Hussain, R. An analysis of consanguineous marriage in the Muslim population of India at regional and the state levels. *Annal. Hum. Biol.* **27,** 163–171 (2000).
14 Aarzoo, S. S. & Afzal, M. Gene diversity in some Muslim populations of North India. *Hum. Biol.* **77,** 343–353 (2005).
15 Gutala, R., Carvalho-Silva, D. R., Jin, L., Yngvadottir, B., Avadhanula, V., Nanne, K. *et al.* A shared Y-chromosomal heritage between Muslims and Hindus in India. *Hum. Genet.* **120,** 543–551 (2006).
16 Khan, F., Pandey, A. K., Tripathi, M., Talwar, S., Bisen, P. S., Borkar, M. *et al.* Genetic affinities between endogamous and inbreeding populations of Uttar Pradesh. *BMC Genet.* **8,** 12 (2007).
17 Terreros, M. C., Rowold, D., Luis, J. R., Khan, F., Agrawal, S. & Herrera, R. J. North Indian Muslims: enclaves of foreign DNA or Hindu converts? *Am. J. Phys. Anthropol.* **133,** 1004–1012 (2007).
18 Eaaswarkhanth, M., Vasulu, T. S. & Haque, I. Genetic Affinity between diverse ethno-religious communities of Tamil Nadu, India: a Microsatellite study. *Hum. Biol.* **80,** 601–609 (2008).
19 Balgir, R. S. & Sharma, J. C. Genetic Markers in the Hindu and Muslim Gujjars of Northwestern India. *Am. J. Phys. Anthropol.* **75,** 391–403 (1988).
20 Tandon, M., Trivedi, R. & Kashyap, V. K. Genomic diversity at 15 fluorescent labeled short tandem repeat loci in few important populations of state of Uttar Pradesh, India. *Forensic Sci. Int.* **128,** 190–195 (2002).
21 Dubey, B., Meganathan, P. R., Eaaswarkhanth, M., Vasulu, T. S. & Haque, I. Forensic STR profile of two endogamous populations of Madhya Pradesh, India. *Leg. Med. (Tokyo)* **11,** 41–44 (2009).
22 Sarkar, N. & Kashyap, V. K. Genetic diversity at two pentanucleotide STR and thirteen tetranucleotide STR loci by multiplex PCR in four predominant population groups of central India. *Forensic Sci. Int.* **128,** 196–201 (2002).
23 Ashma, R. & Kashyap, V. K. Genetic polymorphism at 15 STR loci among three important subpopulation of Bihar, India. *Forensic Sci. Int.* **130,** 58–62 (2002).
24 Mohapatra, B. K., Trivedi, R., Mehta, A. K., Vyas, J. M. & Kashyap, V. K. Genetic diversity at 15 fluorescent-labeled short tandem repeat loci in the Patel and other communities of Gujarat, India. *Am. J. Forensic Med. and Path.* **25,** 108–112 (2004).
25 Hima Bindu, G., Trivedi, R. & Kashyap, V. K. Population genetics of seventeen microsatellite loci in three major groups of Andhra Pradesh, India. *Forensic Sci. Comm.* **7** (2005).
26 Hima Bindu, G., Trivedi, R. & Kashyap, V. K. Allele frequency distribution based on 17 STR markers in three major Dravidian linguistic populations of Andhra Pradesh, India. *Forensic Sci. Int.* **170,** 76–85 (2007).
27 Eaaswarkhanth, M., Roy, S. & Haque, I. Allele Frequency Distribution for 15 Autosomal STR Loci in Two Muslim Populations of Tamilnadu, India. *Leg. Med. (Tokyo)* **9,** 332–335 (2007).
28 Sitalaximi, T., Trivedi, R. & Kashyap, V. K. Autosomal microsatellite profile of three socially diverse ethnic Tamil populations of India. *J. Forensic Sci.* **48,** 211–214 (2003).
29 Sitalaximi, T., Trivedi, R. & Kashyap, V. K. Genotype profile for thirteen tetranucleotide repeat loci and two pentanucleotide repeat loci in four endogamous Tamil population groups of India. *J. Forensic Sci.* **47,** 1168–1173 (2002).
30 Shepard, E. M. & Herrera, R. J. Genetic encapsulation among Near Eastern populations. *J. Hum. Genet.* **51,** 467–476 (2006a).
31 Berti, A., Barni, F., Virgili, A., Iacovacci, G., Franchi, C., Rapone, C. *et al.* Autosomal STR frequencies in Afghanistan population. *J. Forensic Sci.* **50,** 1494–1496 (2005).
32 Shepard, E. M. & Herrera, R. J. Iranian STR variation at the fringes of biogeographical demarcation. *Forensic Sci. Int.* **158,** 140–148 (2006b).
33 Barni, F., Berti, A., Pianese, A., Boccellino, A., Miller, M. P., Caperna, A. *et al.* Allele frequencies of 15 autosomal STR loci in the Iraq population with comparisons to other populations from the middle-eastern region. *Forensic Sci. Int.* **167,** 87–92 (2007).
34 Alshamali, F., Alkhayat, A. Q., Budowle, B. & Watson, N. D. STR population diversity in nine ethnic populations living in Dubai. *Forensic Sci. Int.* **152,** 267–279 (2005).
35 Pérez-Miranada, A. M., Alfonso-Sánchez, M. A., Peña, J. A. & Herrera, R. J. Qatari DNA variation at a crossroad of Human migrations. *Hum. Hered.* **61,** 67–79 (2006).
36 Abdin, L., Shimada, I., Brinkmann, B. & Hohoff, C. Analysis of 15 short tandem repeats reveals significant differences between the Arabian populations from Morocco and Syria. *Leg. Med. (Tokyo)* **5,** S150–S155 (2003).
37 Cakir, A. H., Celebioglu, A., Altunbas, S. & Yardimci, E. Allele frequencies for 15 STR loci in Van-Agri districts of the Eastern Anatolia region of Turkey. *Forensic Sci. Int.* **135,** 60–63 (2003).
38 Seah, L. H., Jeevan, N. H., Othman, M. I., Jaya, P., Ooi, Y. S., Wong, P. C. *et al.* STR data for the AmpF/STR Identifiler loci in three ethnic groups (Malay, Chinese, Indian) of the Malaysian population. *Forensic Sci. Int.* **138,** 134–137 (2003).
39 Wang, C. W., Chen, D. P., Chen, C. Y., Lu, S. C. & Sun, C. F. STR data for the AmpF/STR SGM Plus and profiler loci from Taiwan. *Forensic Sci. Int.* **138,** 119–122 (2003).
40 Hashiyadaa, M., Itakurab, Y., Nagashima, T., Nataa, M. & Funayama, M. Polymorphism of 17 STRs by multiplex analysis in Japanese population. *Forensic Sci. Int.* **133,** 250–253 (2003).
41 Chan, K. M., Chiu, C. T., Tsui, P., Wong, D. M. & Fung, W. K. Population data for the Identifiler 15 STR loci in Hong Kong Chinese. *Forensic Sci. Int.* **152,** 307–309 (2005).
42 Rerkamnuaychoke, B., Rinthachai, T., Shotivaranon, J., Jomsawat, U., Siriboonpiputtana, T., Chaiatchanarat, K. *et al.* Thai population data on 15 tetrameric STR loci—D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, D2S1338, D19S433, vWA, TPOX, D18S51, D5S818 and FGA. *Forensic Sci. Int.* **158,** 234–237 (2006).
43 Yoo-Li, K., Ji-Yeon, H., Yoo-Jin, K., Seok, L., Nak-Gyun, C., Hyun-Gyung, G. *et al.* Allele frequencies of 15 STR loci using AmpF/STR Identifiler kit in a Korean population. *Forensic Sci. Int.* **136,** 92–95 (2003).
44 Dobashi, Y., Kido, A., Fujitani, N., Hara, M., Susukida, R. & Oya, M. STR data for the AmpF/STR Identifiler loci in Bangladeshi and Indonesian populations. *Leg. Med. (Tokyo)* **7,** 222–226 (2005).

45 Montelius, K., Karlsson, A. O. & Holmlund, G. STR data for the AmpF/STR Identifiler loci from Swedish population in comparison to European, as well as with non-European population. *Forensic Sci. Int. Gen.* **2,** e49–e52 (2008).
46 Sánchez-Diz, P., Menounos, P. G., Carracedo, A. & Skitsa, I. 16 STR data of a Greek population. *Forensic Sci. Int. Gen.* **2,** e71–e72 (2008).
47 Decorte, R., Engelen, M., Larno, L., Nelissen, K., Gilissen, A. & Cassiman, J. J. Belgian population data for 15 STR loci (AmpF/STR SGM Plus and AmpF/STR Profiler PCR amplification kit). *Forensic Sci. Int.* **139,** 211–213 (2004).
48 Presciuttini, S., Cerri, N., Turrina, S., Pennato, B., Alu, M., Asmundo, A. *et al.* Validation of a large Italian Database of 15 STR loci. *Forensic Sci. Int.* **156,** 266–268 (2006).
49 Czarny, J., Grzybowski, T., Derenko, M. V., Malyarchuk, B. A. & Sliwka, D. M. Genetic variation of 15 STR loci (D3S1358, vWA, FGA, TH01, TPOX, CSF1PO, D5S818, D13S317, D7S820, D16S539, D2S1338, D8S1179, D21S11, D18S51, and D19S433) in populations of north and central Poland. *Forensic Sci. Int.* **147,** 97–100 (2005).
50 Sambrook, J., Fritsch, E. & Maniatis, T. *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1989).
51 Applied Biosystems. *Applied Biosystems AmpFlSTR® identifilerTM PCR amplification kit user's manual, instruction for use of products* (Applied Biosystems, Foster City, CA, 2001).
52 Raymond, M. & Rousset, F. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J. Hered.* **86,** 248–249 (1995).
53 Excoffier, L., Laval, G. & Schneider, S. Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* **1,** 47–50 (2005).
54 Lui, K. & Muse, S. V. Powermarker: integrated analysis environment for genetic marker data. *Bioinformatics* **21,** 2128–2129 (2005).
55 Felsenstein, J. *Phylogeny Inference Package (PHYLIP) version 3.6c/ distributed by the author* (Department of Genetics, University of Washington, Seattle, 1993).
56 *Statistical Package for the Social Sciences (SPSS) software 12.0.1*, Evaluation version for Windows. Chicago: SPSS Inc.
57 Ota, T. *DISPAN: Genetic Distance and Phylogenetic Analysis* (University Park, PA: Tatsuya Ota and Pennsylvania State University, USA, 1993).
58 Chakraborty, R. Gene Identity in racial hybrids and estimation of admixture rates In: *Genetic microdifferentiation in man and other animals* (eds Neel, J.V. and Ahuja, Y.R.) 171–180 (Delhi University, Delhi: Indian Anthropological Association,, 1985)..
59 Elston, R. C., Olson, J. M. & Lyle, Palmer *Biostatistical Genetics and Genetic Epidemiology* 1–5 (John Wiley and Sons, Inc., 2002).
60 Barnholtz-Sloan, J. S., Pfaff, C. L., Chakraborty, R. & Long, J. C. Informativeness of the CODIS STR loci for admixture analysis. *J. Forensic Sci.* **50,** 1322–1326 (2005).
61 Mukherjee, N., Nebel, A., Oppenheim, A. & Majumder, P. P. High-resolution analysis of Y-chromosomal polymorphisms reveals signatures of population movements from Central Asia and West Asia into India. *J. Genet.* **80,** 125–35 (2001).
62 Cordaux, R., Saha, N., Bentley, G. R., Aunger, R., Sirajuddin, S. & Stoneking, M. Mitochondrial DNA analysis reveals diverse histories of tribal populations from India. *Eur. J. Hum. Genet.* **11,** 253–264 (2003).
63 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).
64 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).
65 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).
66 Thanseem, I., Thangaraj, K., Chaubey, G., Singh, V. K., Bhaskar, L. V. K. S., Reddy, B. M. *et al.* Genetic affinities among the lower castes and tribal group of India: inference from Y chromosome and mitochondrial DNA. *BMC Genet.* **7,** 42 (2006).
67 Metspalu, M., Kivisild, T., Metspalu, E., Parik, J., Hudjashov, G., Kaldma, K. *et al.* Most of the extant mtDNA boundaries in South and Southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet.* **5,** 26 (2004).
68 Quintana-Murci, L., Chaix, R., Wells, R. S., Behar, D. M., Sayar, H., Scozzari, R. *et al.* Where West meets East: the complex mtDNA landscape of the Southwest and Central Asian corridor. *Am. J. Hum. Genet.* **74,** 827–845 (2004).

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