BMC Genetics

BioMed Central

Research article

Open Access

Most of the extant mtDNA boundari,

Table 1: Characteristics of the Indian and Iranian population samples whose mtDNA variation has been determined in the course of this study.

INDIA

IRAN

the southern states, and peaked at 86% in West Bengal (Table 8, see Additional file 3).

With the exception of the diverse set of largely Indian-specific R lineages, the most frequent mtDNA haplogroup in India that derives from the phylogenetic node N is haplogroup W [13]. The frequency peak of haplogroup W is 5% in the northwestern states – Gujarat, Punjab and Kashmir. Elsewhere in India its frequency is very low (from 0 to 0.9%) (Table 2) forming a significant spatial cline (Figure 4).

At 15% among the caste and 8% among the tribal populations haplogroup U is the most frequent sub-clade of R in India (Table 12, see Additional file 7). Approximately one half of the U mtDNAs in India belong to the Indian-specific branches of haplogroup U2 (U2i: U2a, U2b and U2c) [13,27] (Table 2). They are present throughout India without a clear geographical cline (Figure 2, panel U2i, SAA $p > 0.05$). However, the spread of another subset of U, haplogroup U7 [13], is similar to that of haplogroup W, peaking at 12% and 9% in Gujarat and Punjab, respectively (Table 11, see Additional file 6). The frequency of U7 is also high in neighboring Pakistan (6%) and particularly in Iran (9%) (Table 9, see Additional file 4).

MtDNA haplogroups in Iran

Over 90% of the mtDNAs found in Iran belong to haplogroups HV, TJ, U, N1, N2 and X, commonly found in West Eurasia (Table 2). In contrast to Europe, where H is predominant among the mtDNA haplogroups, in Iran the frequency of haplogroup U (29%) is higher than that of haplogroup H (17%) (Table 9, see Additional file 4). This difference accounts, at least partly for the presence in Iran of U sub-groups, such as U7 (9.4%), that are virtually absent in Europe.

Compared to India, haplogroup M frequency in Iran is marginally low (5.3%) and there are no distinguished Iranian-specific sub-clades of haplogroup M. All Iranian haplogroup M lineages can be seen as derived from other regional variants of the haplogroup: eleven show affiliation to haplogroup M lineages found in India, twelve in East and Central Asia (D, G, and M8) and one in northeast Africa (M1).

Table 2: Geographic, linguistic and socio-cultural distribution of major Indian-specific mtDNA haplogroups

	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Socio-cultural affiliation (Indian data only)														
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Language groups of India														
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

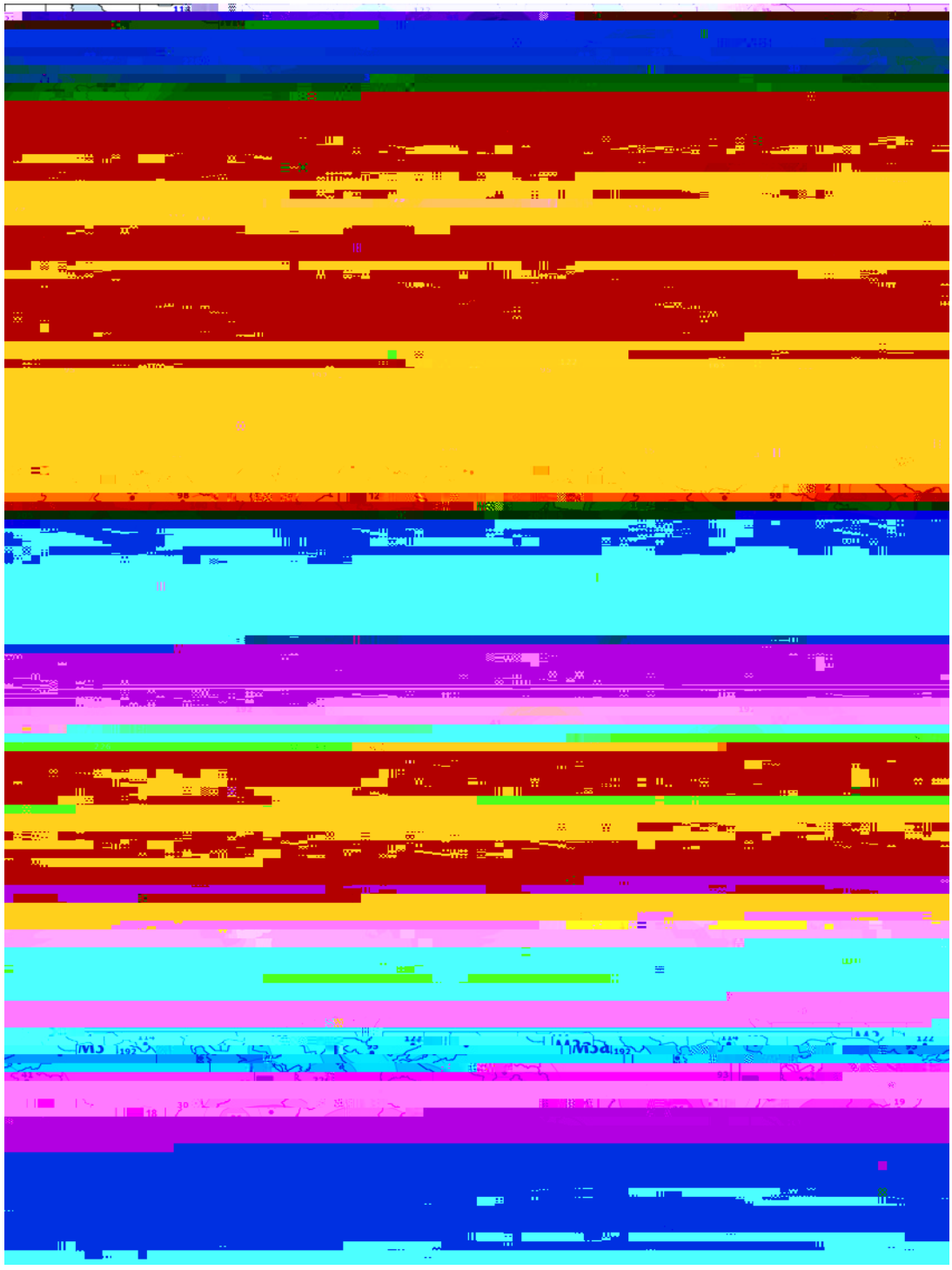

Figure 1
The spatial distribution of Indian-specific mtDNA haplogroups (M2, M6 and M3) and their sub-haplogroups.

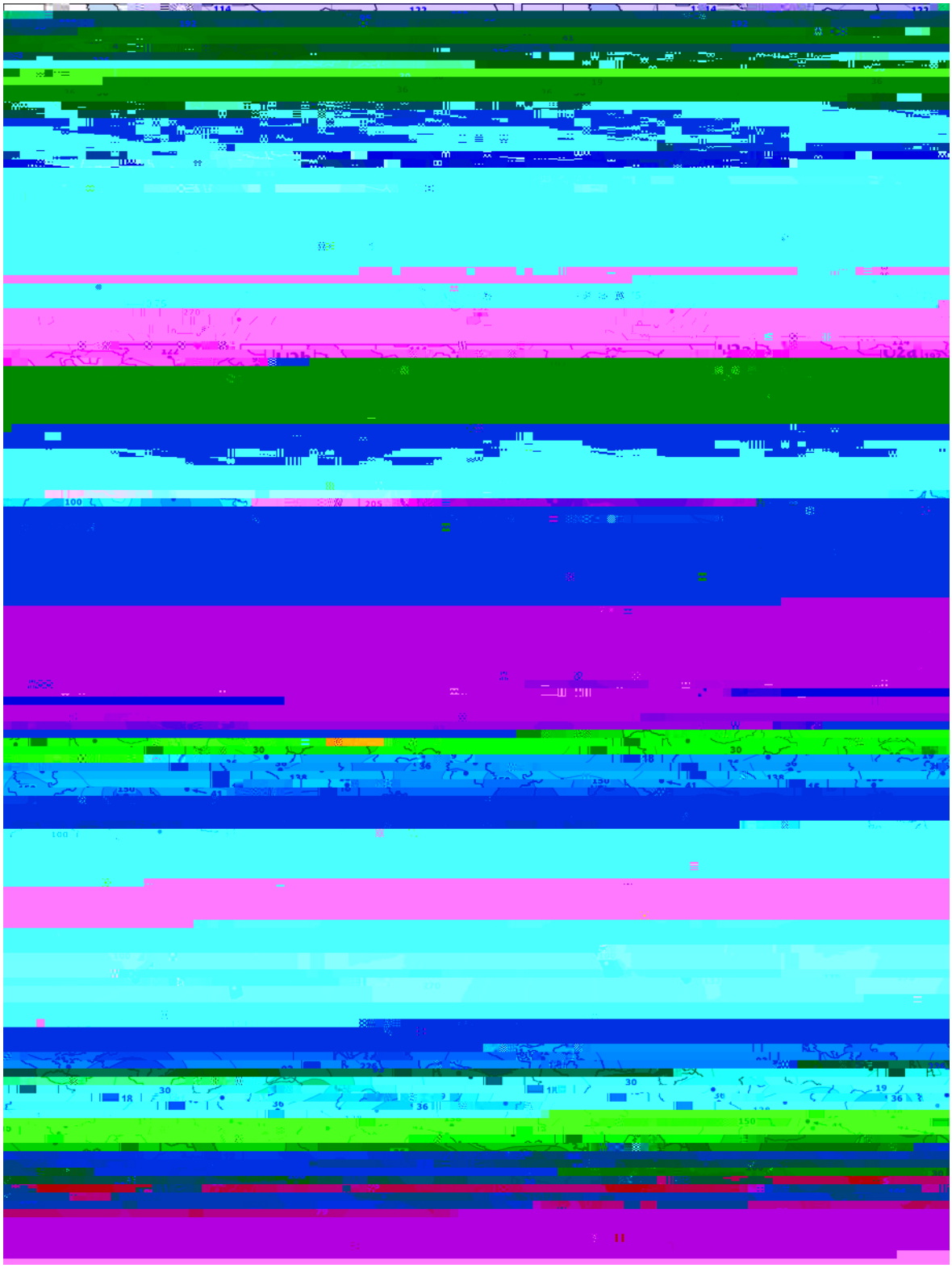

Figure 2

Indian-specific (R5 and Indian-specific M and U2 variants) and East Asian-specific (A, B and East Asian-specific M subgroups) mtDNAs, both, make up less than 4% of the Iranian mtDNA pool. We used Turkey (88.8 ± 4.0%) as the third parental population for evaluating the relative proportions of admixture from India (2.2 ± 1.7%) and China (9.1 ± 4.1%) into Iran. Therefore we can conclude that historic gene flow from India to Iran has been very limited.

The package of the most ancient mtDNA haplogroups in India

Approximately one tenth of the Indian haplogroup M mtDNAs fall into its major sub-clade M2, which is defined by the motif 477G-1780-8502-16319 [15]. M2 can be further subdivided into haplogroups M2a (transitions at nps 5252 and 8369) and M2b [15]. Haplogroup M2 and its two major sub-clades reveal coalescence times of 50 to 70 thousand years (Table 3). Due to the increased frequency towards the southern part of India (Figure 1, panel M2, SAA $p < 0.05$ Figure 4), M2 is significantly ($p < 0.05$) more frequent among the Dravidic speakers than among the Indo-European speakers who are spread mostly in the northern regions of India (Table 2). It is more plausible that geography rather than linguistics is behind this pattern, because the frequency of M2 amongst the Indo-European speaking populations in southern India is significantly higher than that in the north, while there is no significant difference between Dravidic and Indo-European speaking populations from the same geographic region (Table 2). It is also notable that the frequency of M2 among the Brahmins and the Kshatriyas of Andhra Pradesh (CR 3.3 – 19.2%) is not significantly ($p > 0.05$) different from that among the other castes or the tribal populations of the region (CR: 5–12.9%, 11.2–18.3%, respectively). On the other hand, none of the 159 Brahmins and Kshatriyas from the northern states of India (Punjab, Rajasthan, Uttar Pradesh and West Bengal) belong to M2 while the frequency reaches nearly 3% (CR: 1.6–4.6%) among the other castes and tribal populations of the region.

We found that R5, which is defined by transitions at nps 8594 [27], 16266 and 16304, is the second most frequent sub-clade of R in India after haplogroup U. The coalescence age estimate for R5 was similar to that of M2 (Table 3), whereas individual boughs within the R5 limb showed expansions from ca. 20,000 ybp to ca. 50,000 ybp (Figure 12, see Additional file 8). Our data indicate that this diverse and ancient haplogroup is present over mor mboJ0 -T*0.000

Figuare3.

(around 50,000 – 70,000 ybp). Together they constitute nearly 15% of the Indian mtDNAs. Importantly, these haplogroups are virtually absent elsewhere in Eurasia [13,15], this study]. Because most of Indian varieties of haplogroup M are still unclassified (M*), this package is likely to be extended when more mtDNA coding region information will become available for the M* lineages in India.

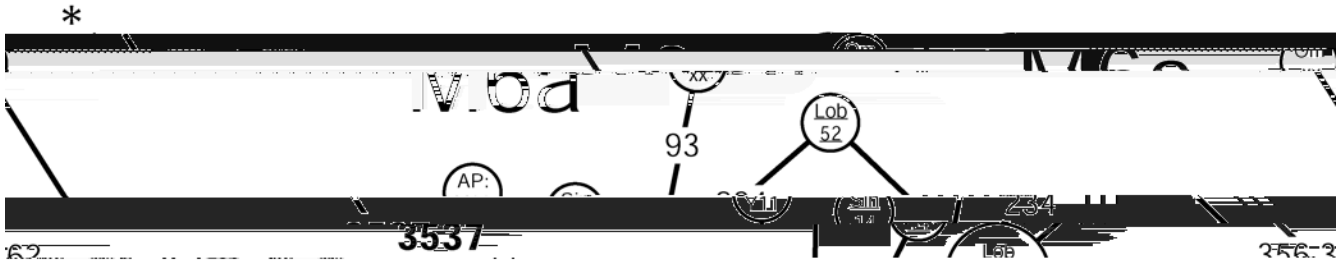

Figure 6
Network of HVS-I haplotypes belonging to haplogroup M6.

Figure 7
A network of haplogroup M3a haplotypes.

Lodha of West Bengal (Table 10, see Additional file 5) suggests a possible founder effect in this population. This explains the nearly two-fold difference between the coalescence estimates for this cluster calculated with and without the tribal data (Table 3).

The G to A transition at np 15928 has been spotted on different branches (e. g. haplogroups T and M) of the mtDNA phylogeny [3,36]. Quintana-Murci and colleagues observed this transition within haplogroup M in combination with the HVS-I motif 16048-16129-16223-16390 [27]. None of the mtDNAs in our study which har-

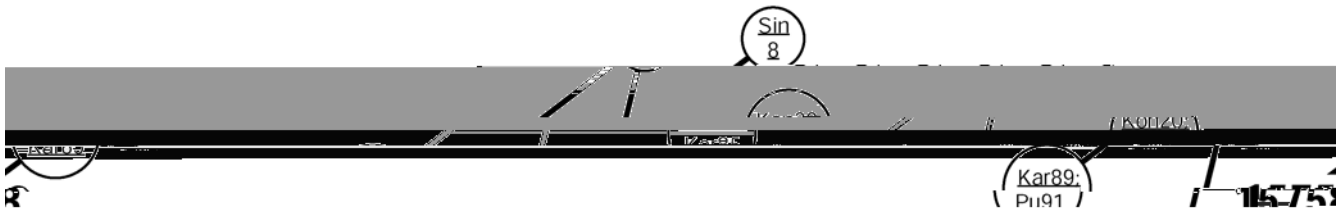

Figure 8
Network of HVS-I haplotypes belonging to haplogroup M4a.

bor – or stem from – the 16048-16129-16223 motif were positive for the 15928 transition, suggesting an additional occurrence. In addition, we recorded this transition associated with three other HVS-I motifs on the background of haplogroup M (M8-Z: 16185-16223-16260-16298; M*: 16223 and M*: 16086-16223-16335). These occurrences cannot be monophyletic for obvious reasons. Yet, when combined with the transition at np 16304, G15928A roots a star-like subclade of haplogroup M that we tentatively named M25 (Figure 10). In this case, monophylecity is the most parsimonious assumption. This haplogroup is moderately frequent in Kerala and Maharashtra but rather infrequent elsewhere in India (Figure 2, panel M25).

Coalescence estimates for these Indian-specific mtDNA haplogroups (M3a, M4a, M6, M25 and R6) fall largely between 20,000 and 30,000 ybp. These estimates overlap with those of many West Eurasian-specific (e.g. H, HV, preHV, U3, U4, K, X [9,34]) and East Eurasian-specific (A, F2, D4, M7c1, M7a1, M8a [7,22]) mtDNA clades, suggesting a rather synchronic worldwide demographic expansion event in the late Pleistocene, during an interglacial period preceding the LGM.

Several Indian-specific mtDNA clades demonstrate a similar spread-pattern in southern India. We found haplogroups M4a, M6a and M18 in southeastern Tamil Nadu and Andhra Pradesh while they were absent from neighboring Karnataka and Kerala (Figure 1 panel M6a and Figure 2 panels M4a and M18). One possible explanation is that admixture has been facilitated along the coastlines of the Arabian Sea and the Bay of Bengal. On the other hand, because the absolute frequencies of these haplogroups are rather low, it cannot be ruled out that an increase of sample sizes would disrupt the observed spread-pattern.

Were the Austro-Asiatic speaking tribal people the earliest inhabitants of India?

By calculating nucleotide diversities and expansion times (using the method from [37]) for different linguistic groups of India, some previous studies on mtDNA variation have distinguished the Austro-Asiatic speaking tribal groups as the carriers of the genetic legacy of the earliest

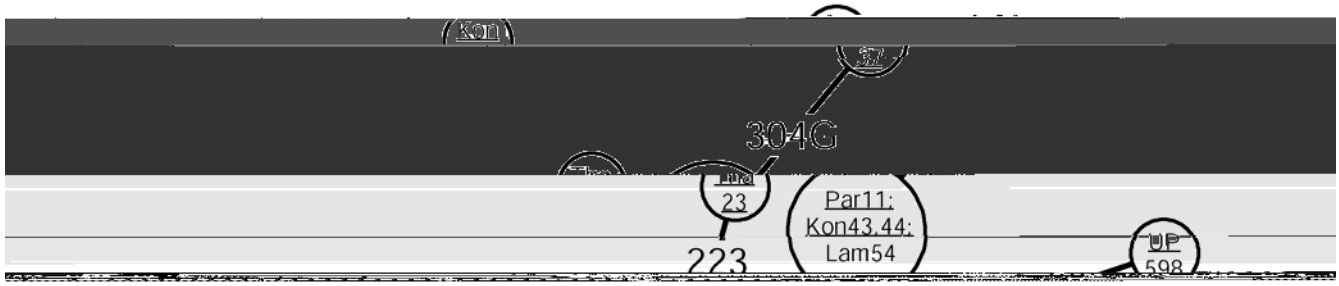

Figure 10
Network of HVS-I haplotypes belonging to haplogroup M25.

16319-16352), without information from the coding region it is not clear whether the other two sequences (HVS-I motifs: 16092-16179-16223-16289-16294-16319 and 16147G-16172-16223-16319) represent novel M2 sub-clades (because these sequences cannot be affiliated with M2a or M2b) or derive from two independent branches of haplogroup M where 16319 transition has arisen recurrently. HVS-I motif 16147G-16172-16223, for example, is commonly associated with haplogroup N1a. Since sequence data on the five M2s among the Austro-Asiatic speaking tribe Ho, reported by Basu et al. (2003), have not been made available in the publication, we cannot rely on their haplogroup classification. Thus, we are left with one Munda and one Santal mtDNA belonging to haplogroup M2. They make up just 5% of the Austro-Asiatic tribal sample of 37 subjects (excluding the ten Ho). Interestingly, we found no instances of haplogroup M2 among the 56 Lodhas analyzed in this study. Consequently, when excluding the recurrences of the 16319 transition on the background of other sequence motifs, the frequency of M2 among the Austro-Asiatic speaking tribal groups from West Bengal in the combined dataset (Table 7, see Additional file 2) is significantly reduced to about 2%. The corrected value is comparable to the M2 frequency (>3%) in tribal populations speaking Indo-European languages of Punjab and Uttar Pradesh, but is significantly lower than its frequency (>14%) among the Dravidic-speaking tribal groups of Andhra Pradesh (Table 8, see Additional file 3).

Language families present today in India, such as Indo-European, Dravidic and Austro-Asiatic, are all much younger than the majority of indigenous mtDNA lineages found among their present-day speakers at high frequencies (see Additional file 9). It would make it highly speculative to infer, from the extant mtDNA pools of their speakers, whether one of the listed above linguistically defined group in India should be considered more "autochthonous" than any other in respect of its presence in the subcontinent.

Additionally, we note that some recent linguistic and archaeological evidence place the spread of the Austro-Asiatic languages in the Neolithic, in conjunction with the

dispersal of rice cultivation from the Yangtze River basin [39]. If this were the case, it would imply that the arrival of this linguistic phylum in India was not associated with

Figure 11
The segregation of West Eurasian, East Eurasian and South Asian mtDNA pools.

extending north into Central Asia. The coalescence times of these haplogroups suggest that this continuum took shape somewhere between 30,000 to 50,000 ybp (Table 4), thus falling within the climatically favorable interglacial period. We notice that the extant U7 and W frequencies along the proposed continuum are not uniform. U7 is more predominant in Iran, Pakistan, northwestern India and the Arabian peninsula, while W is more frequent in the western Near-East, Anatolia and the Caucasus. The coalescence ages of the Indian- and Iranian-specific U7 clades suggest that the time-window of this continuum was closed by ca. 20,000 ybp. The inferred extreme aridity of eastern Iran and western India during the last glacial maximum, which is well documented in paleovegetation reconstructions [42] may explain the observed segregation.

It has been suggested that the Jews settled in southwest India on the coast of the Arabian Sea sometime during the early Middle Ages. However, the mtDNA pool of the extant Cochin Jews is overwhelmingly Indian-specific (Table 10, see Additional file 5). We found exact or close matches to the fourteen HVS-I haplotypes observed among the Cochin Jews in other Indian populations. It is not clear whether the Near Eastern mtDNA lineages have been lost or the initial Jewish settlers did not include women.

Gene flow from East Eurasia

The East Eurasian-specific mtDNA haplogroups are less common in India and more sharply geographically segregated than the haplogroups of western Eurasian ancestry (Table 2; Figure 11, panel C). Indian caste populations harbor only about 4% of such mtDNAs, compared to 17% of the West Eurasian ones (Table 2). Elevated frequencies of haplogroups common in eastern Eurasia are observed in Bangladesh (17%) and Indian Kashmir (21%) and may

The number of shared haplotypes between pairs of social, linguistic and geographic groups of Indian populations is slightly (but in most cases insignificantly) lower than that between random groups of Indian populations taken for reference (see Materials and Methods). Where the decline of shared haplotypes is significant relative to the reference, it is most probably caused by large differences in the sample sizes of the groups under comparison (Table 5).

An alternative method that assesses the degree of haplotype sharing between populations is to investigate the combined frequency of the shared haplotypes in two population groups. Thus, amongst the northern and the southern population groups the combined frequency of the haplotypes present also in the other group is significantly lower than that which we observed in the case of random groups. This is not surprising because West Eurasian-specific mtDNA haplogroups are rather frequent in northwest India. Because the Indo-European and the Dravidic speakers of India are largely concentrated to the northern and southern parts of the subcontinent, respectively, the differences arising from geographic division of the Indian populations also correspond to these linguistic groupings (Table 5).

Conclusions

Three Indian-specific haplogroups, M2, U2i and R5, which encompass about 15% of the Indian mtDNA pool, exhibit equally deep coalescence ages of about 50,000 – 70,000 years. Thus, their spread can be associated with the initial peopling of South Asia.

Haplogroups U7, W and R2 harbor a number of similar traits. Their overlapping geographic distributions and coalescence times suggest some degree of genetic continuum in the area spanning from the Near and Middle East through northwest India and reaching north into Central Asia. The coalescence estimates for these haplogroups are equally deep (around 30,000 – 50,000 years) in these different regions. That may be a result of either relatively more recent albeit large in scale migrations that brought along most of the diversity or may indeed reflect the region-specific expansions of these haplogroups. The former explanation could be ruled out since it is impossible to envisage a substantial movement of mtDNAs from South Asia that would not include haplogroup M. The same is true for the opposite – the share of U7, W and R2 within the West Eurasian-specific mtDNA haplogroups is two-fold higher in India than it is in Iran. Moreover, the South- and West Asian-specific sub-branches of haplogroup U7 predate the last glacial maximum. Therefore, deep autochthonous history of these haplogroups in the region remains to be the most parsimonious explanation. Through the use of mtDNA coding region markers, we were able to classify altogether a quarter of the Indian M and R mtDNAs into a number of Indian-specific mtDNA haplogroups, four of which we newly identified. Several of these are characterized by clear patterns in their geo-

and the neighboring Bijnor district of Uttar Pradesh. Both these tribal groups speak languages belonging to the Indo-European phylum [47].

The non-tribal Indian samples analyzed contained 105 West Bengalis of different caste rank, 58 Konkanastha Brahmins from Bombay, 53 Gujaratis, 50 Moors and 82 Sinhalese from Sri Lanka, 109 Punjabis of different caste rank from the Punjab, 25 Brahmins from Uttar Pradesh, 35 Rajputs from Rajasthan, 55 Parsees from Maharashtra and 100 subjects from Cochin, Kerala (including 45 Jews who have moved to Israel) (Table 1).

The Iranian sample of 436 individuals was collected in different locations mainly from southwestern and north-western Iran (Table 1).

The new Indian mtDNA sequence data was combined with that previously published on Indian populations [3,13,15,17,18,29,35,49-51] to produce a pooled dataset (n = 2572) in which the tribal populations constitute slightly over 50% (Table 6, see Additional file 1). By including also the data on Iranian and published data on Pakistani [13,27,29], Bangladeshi [29], Chinese [7,22,52-54] and Thai [53,55,56] populations (n = 145, 29, 919

slightly ova6btrea-6.G.diff. Kelogy with1(pu)-6.1(blroups speak)-55.8wl5.7(cludingno)ibal go1.2037 T00t2099 TD0.00HVS-I 7(d

8

Acknowledgments