Journal of Human Genetics (2011) 56, 765–771

www.nature.com/jhg

ORIGINAL ARTICLE

Y-chromosome haplogroup diversity in the sub-Himalayan Terai and Duars populations of East India

Monojit Debnath[1], Malliya G Palanichamy[1,2], Bikash Mitra[2,3], Jie-Qiong Jin[1], Tapas K Chaudhuri[3] and Ya-Ping Zhang[1,2]

The sub-Himalayan Terai and Duars, the important outermost zones comprising the plains of East India, are known as the reservoirs of ethnic diversity. Analysis of the paternal genetic diversity of the populations inhabiting these regions and their genetic relationships with adjacent Himalayan and other Asian populations has not been addressed empirically. In the present investigation, we undertook a Y-chromosome phylogeographic study on 10 populations (n=375) representing four different linguistic groups from the sub-Himalayan Terai and Duars regions of East India. The high-resolution analysis of Y-chromosome haplogroup variations based on 76 binary markers revealed that the sub-Himalayan paternal gene pool is extremely heterogeneous. Three major haplogroups, namely H, O and R, are shared across the four linguistic groups. The Indo-European-speaking castes exhibit more haplogroup diversity than the tribal groups. The findings of the present investigation suggest that the sub-Himalayan gene pools have received predominant Southeast Asian contribution. In addition, the presence of Northeast and South Asian signatures illustrate multiple events of population migrations as well as extensive genetic admixture amongst the linguistic groups.
Journal of Human Genetics (2011) **56,** 765–771; doi:10.1038/jhg.2011.98; published online 8 September 2011

Keywords: admixture; Duars; gene pool; haplogroup; sub-Himalayan populations; Terai; Y chromosome

INTRODUCTION

The Himalaya is a complex and vast mountain system with a contradictory evidence of human settlement. Although majority of the linguistic as well as genetic studies strongly favor Neolithic settlement,[1–5] archeological findings along with a recent mitochondrial DNA study advocate that the modern human reached Tibetan plateau by the late Paleolithic age.[6,7] The factors that influenced the settlement of human populations in different mountain ranges of the Himalaya included not only the climatic conditions but also the cultural and religious affinities. With the exception of western regions, some parts of the central and the entire eastern ranges of the Himalaya show domination of the Tibeto-Burman (TB) speakers, the origin of which has been proposed to be the upper and middle Yellow River basin by one group,[3] whereas van Driem[5,8] argued in favor of Yangtse River in Sichuan province. Multiple genetic studies based on the markers present in autosomes, mitochondrial DNA and Y-chromosome in the populations from Tibet have demonstrated that the Tibetan gene pool has largely been contributed by the immigrants from Northeast Asia.[1,3,9] The earlier proposition of Central Asian origin of Tibetans due to high frequency of D*-M174 sub-haplogroup of Y-chromosome[10] has been opposed by a recent study,[11] suggesting its southern origin in East Asia followed by a northward migration. Similarly, genetic diversity studies in the TB-speaking populations inhabiting Nepal and eastern Himalayan regions of Northeast India have indicated that their gene pools not only received East Asian contribution but also bear signatures of South Asian influence.[12–18] It is noteworthy that these geographic regions played a pivotal role in shaping the genetic landscapes as well as peopling of the Himalaya. This is reflected by genetic studies that suggested Northeast India as an important corridor for human dispersals, whereas Nepal Himalaya as a barrier for bidirectional gene flow.[16,19]

Apart from these Himalayan regions, there are some interesting outermost sub-Himalayan zones in the plains of East India, which are known as Terai and Duars (also Dooars). Unlike the Himalaya, these areas have characteristically unique and diverse ethnic populations belonging to four different linguistic groups such as TB, Austro-Asiatic (AA), Indo-European (IE) and Dravidian (DR). Because of their strategic geographic location and close proximity with Nepal, Bhutan

[1]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China; [2]Laboratory for Conservation and Utilization of Bio-resource, Yunnan University, Kunming, China and [3]Cellular Immunology Laboratory, University of North Bengal, Raja Rammohanpur, Darjeeling, India
Correspondence: Dr M Debnath, State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China.
Email: monozeet@gmail.com
or Professor Y-P Zhang, State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China and Laboratory for Conservation and Utilization of Bio-resource, Yunnan University, Kunming 650091, China.
E-mail: zhangyp1@263.net.cn
Received 30 March 2011; revised 20 July 2011; accepted 26 July 2011; published online 8 September 2011

and other eastern territories of the Himalaya, these sub-Himalayan regions might have witnessed multiple waves of population migrations and extensive population admixture. These processes might have transformed the sub-Himalayan Terai and Duars to important reservoirs of genetic diversity. Apart from few preliminary investigations based on classical genetic and mitochondrial DNA markers,[18,20–21] there is a dearth of high-resolution Y-haplogroup data in the sub-Himalayan Terai and Duars populations. Therefore, the analysis of the paternal genetic diversity of the ethnic groups from these regions will not only help to resolve their genetic origins but also elucidate their phylogenetic relationships with other Asian populations.

In the present investigation, we considered 10 populations representing four linguistic groups from the sub-Himalayan Terai and Duars for the first time for high-resolution Y-chromosome haplogroup diversity studies. A total of 76 binary markers from the non-recombining region of Y-chromosome were typed in 375 males. The results of our analyses suggest that the sub-Himalayan Terai and Duars paternal gene pools are extremely heterogeneous, and received East as well as South Asian genetic contribution.

MATERIALS AND METHODS

Samples

We considered a total number of 375 unrelated males from 10 ethnic groups comprising both the tribes and castes from Sikkim and the sub-Himalayan Terai and Duars regions of East India. The tribes included three Mundari-speaking AA groups, such as Kol (n=62), Santhal (n=51) and Kharia (n=34); four TB-speaking groups, namely, Dhimal (n=36), Rabha (n=26), Mech (n=19) and Lachungpa (n=11); and one DR-speaking Oraon (n=31), whereas castes included IE-speaking Bengali (n=54) and Rajbanshi (n=51). The geographic locations of the above populations have been depicted in Figure 1. Blood samples were collected from the volunteers with informed consent. All ethical guidelines were followed, as stipulated by the institutions involved in the study. DNA was extracted using phenol–chloroform method and was stored in 10 mM Tris-1 mM EDTA (pH 8.0) at −80 °C.

Y-chromosome haplogrouping

A total of 76 binary markers were genotyped by standard methods such as denaturing high-performance liquid chromatography, direct sequencing (ABI 3730 Genetic Analyzer, Applied Biosystems, Foster City, CA, USA) and GenomeLab SNPstream Genotyping System (Beckman Coulter, Fullerton, CA, USA) (Supplementary Table S1). We followed a hierarchical genotyping strategy identifying the major haplogroups first and then the downstream mutations along the haplogroup tree. The nomenclature followed for the Y-haplogrouping is as recommended by Y Chromosome Consortium[22] and Karafet *et al*.[23] The primer and sequence information for the insertion, deletion as well as single-nucleotide polymorphisms (SNPs) genotyped by denaturing high-performance liquid chromatography and direct sequencing were taken from Underhill *et al*.[24] and Karafet *et al*.[23] For the SNPs genotyped by the GenomeLab SNPstream genotyping system, we have used 12-plex format and performed the multiplex PCR as well as extension reactions following manufacturer's recommendations. Genotyping was carried out in an automated multiplexed system that employed oligonucleotide microarrays manufactured in a 384-well format on a glass-bottomed plate. The 12-plex PCR and extension primers were designed by Autoprimer software (Beckman Coulter).

Data analyses

Altogether 25 reference populations from the published literature were included for phylogenetic analyses. Reference populations comprised eight Southeast Asians (SEAS), eight Northeast Asians (NEAS) and nine South Asian groups. The South Asian populations included DR tribes (n=16), DR castes (n=6), IE castes (n=37), IE tribes (n=9), AA tribes (n=9), TB tribes (n=12) from India and two TB tribes, one IE caste and one IE-speaking tribe from Nepal (Supplementary Table S2). The frequency data taken from published literature were normalized at the phylogenetic resolution of major haplogroups. The description of the populations analyzed have been represented in Table 1. The frequencies of major Y-haplogroups in the reference populations are shown in Supplementary Table S3. Statistical analyses were done by using the graphPad software (http://statpages.org/).

RESULTS AND DISCUSSION

In the present investigation, we observed altogether 22 Y-chromosome haplogroups defined by 35 polymorphic loci, the phylogeny and frequencies of which are illustrated in Figure 2.

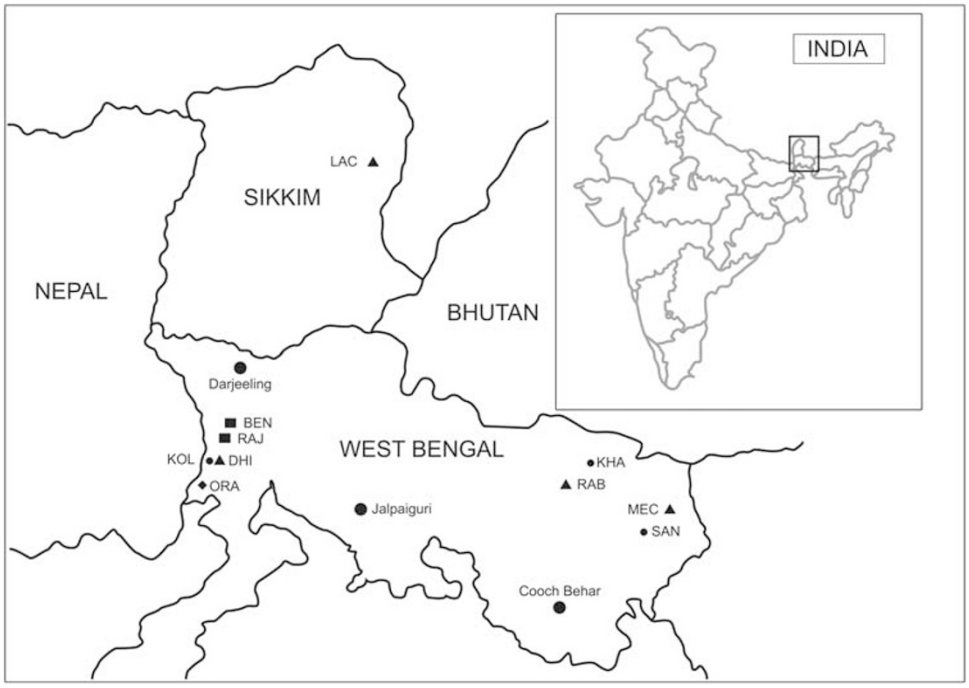

Figure 1 Geographic locations of the East Indian populations examined in this study. The codes used for the studied populations are: Kol (KOL), Santhal (SAN), Kharia (KHA), Dhimal (DHI), Rabha (RAB), Mech (MEC), Lachungpa (LAC), Oraon (ORA), Bengali (BEN) and Rajbanshi (RAJ). A full color version of this figure is available at the *Journal of Human Genetics* journal online.

Table 1 Populations analyzed

Group	*No*	*Population*	*Code*	*Size*	*Language family*	*References*
South Asian	1	Kol	KOL	62	Austro-Asiatic	This study
	2	Santhal	SAN	51	Austro-Asiatic	This study
	3	Kharia	KHA	34	Austro-Asiatic	This study
	4	Dhimal	DHI	36	Tibeto-Burman	This study
	5	Rabha	RAB	26	Tibeto-Burman	This study
	6	Mech	MEC	19	Tibeto-Burman	This study
	7	Lachungpa	LAC	11	Tibeto-Burman	This study
	8	Oraon	ORA	31	Dravidian	This study
	9	Bengali	BEN	54	Indo-European	This study
	10	Rajbanshi	RAJ	51	Indo-European	This study
	11	TB Tribe	TBT	138	Tibeto-Burman	Sengupta *et al.*,[13] Sahoo *et al.*[14]
	12	DR Tribe	DRT	292	Dravidian	Sengupta *et al.*,[13] Sahoo *et al.*[14]
	13	AA Tribe	AAT	147	Austro-Asiatic	Sengupta *et al.*,[13] Sahoo *et al.*[14]
	14	IE Tribe	IET	125	Indo-European	Sengupta *et al.*,[13] Sahoo *et al.*[14]
	15	DR Caste	DRC	173	Dravidian	Sengupta *et al.*[13]
	16	IE Caste	IEC	834	Indo-European	Sengupta *et al.*,[13] Sahoo *et al.*,[14] Zhao *et al.*[25]
	17	Nepal TB Tribe	NTT	111	Tibeto-Burman	Gayden *et al.*[16]
	18	Nepal IE Caste	NCC	77	Indo-European	Gayden *et al.*[16]
	19	Nepal IE Tribe	NIT	171	Indo-European	Fornarino *et al.*[26]
SEAS	20	Balinese	BLI	641	Austronesian	Karafet *et al.*[27]
	21	Philippines	PHI	48	Austronesian	Karafet *et al.*[28]
	22	Vietnam	VIE	70	Austro-Asiatic	Karafet *et al.*[28]
	23	Malaysia	MAL	32	Austronesian	Karafet *et al.*[28]
	24	Taiwan Han	TAN	82	Sino-Tibetan	Karafet *et al.*[29]
	25	Chinese She	SHE	45	Hmong-Mien	Zhong *et al.*[30]
	26	Chinese Zhuang	ZHU	61	Daic	Zhong *et al.*[30]
	27	Hani	HAN	60	Tibeto-Burman	Zhong *et al.*[30]
NEAS	28	Korea	KOR	74	Altaic	Karafet *et al.*[29]
	29	Japan	JAP	259	Altaic	Hammer *et al.*[31]
	30	Tibet	TIB	156	Sino-Tibetan	Gayden *et al.*[16]
	31	Uygur	UYG	71	Altaic	Zhong *et al.*[30]
	32	Manchu	MAN	35	Altaic	Xue *et al.*[32]
	33	Mongolia Outer	MOO	147	Altaic	Karafet *et al.*[29]
	34	Mongolia Inner	MOI	22	Altaic	Zhong *et al.*[30]
	35	Han Chinese	HAC	56	Sino-Tibetan	Zhong *et al.*[30]

Abbreviations: AA, Austro-Asiatic; DR, Dravidian; IE, Indo-European; NEAS, Northeast Asian; SEAS, Southeast Asian; TB, Tibeto-Burman.

The Y-chromosome haplogroup diversity data indicate that overall DR-speaking Oraon shows high homogeneity with only six haplogroups. A highly heterogeneous composition of paternal lineages is exhibited by IE-speaking groups with 16 haplogroups, followed by AA- and TB-speakers with 11 and 14 haplogroups, respectively. Of the 22 haplogroups, only three major clades like H, O and R are found to be shared across the four linguistic groups and accounted for >85% of the samples.

Haplogroup O

Haplogroup O is found to be most prevalent, which occurs at an average frequency of 43.5%. in the studied populations. When compared across the four linguistic groups, the frequencies of O2 and O3 sub-haplogroups varied significantly. Majority of the TB-speakers (56.5%) is represented by O3, whereas 44.9% of the AA-speaking individuals accounted for O2 sub-haplogroup. In addition to this, surprisingly O3 turned out to be the second most abundant haplogroup (24.8%) in the IE-speaking castes.

Haplogroup H

The second highest average frequency (26.3%) is seen for H haplogroup. It is observed that 71% of the DR-speaking Oraon tribe accounted for H haplogroup, followed by 36.7% of AA-, 14.3% of IE- and 8.6% of TB-speaking groups.

We have detected a previously uncharacterized mutation, an A-G transition at nucleotide position 147, while genotyping M89 and named it as M505. Fornarino *et al.*[26] also detected this variation in two samples from Tharu population of Nepal but could not define it phylogenetically. Similar to the observation of Fornarino *et al.*,[26] we also noticed that M505 is neither present in H*-M69 nor in H2*-Apt chromosomes. After screening all the M69-, M52-, M370- and M82-bearing chromosomes, we find that all the M82 chromosomes are M505 positive. On the basis of this finding, we place M505 under H1 haplogroup as H1a1*. It is worth noting that the phylogenetic tree of H haplogroup has recently been revised by Fornarino *et al.*,[26] showing M370 as H1a* and M82 as H1a1*. This change deviates from the earlier tree depicting M82 as H1a* and M370 as H1b.[23] Our findings although reveal that all the M82-bearing chromosomes are M370 derived, the identification of M505 leads to further modification of the H-haplogroup tree. The earlier H1a1*-M82 now becomes H1a1*-M505 and M82 has been marked as H1a1a*. The revised Y-haplogroup tree is depicted in Figure 2. We also typed the terminal SNPs

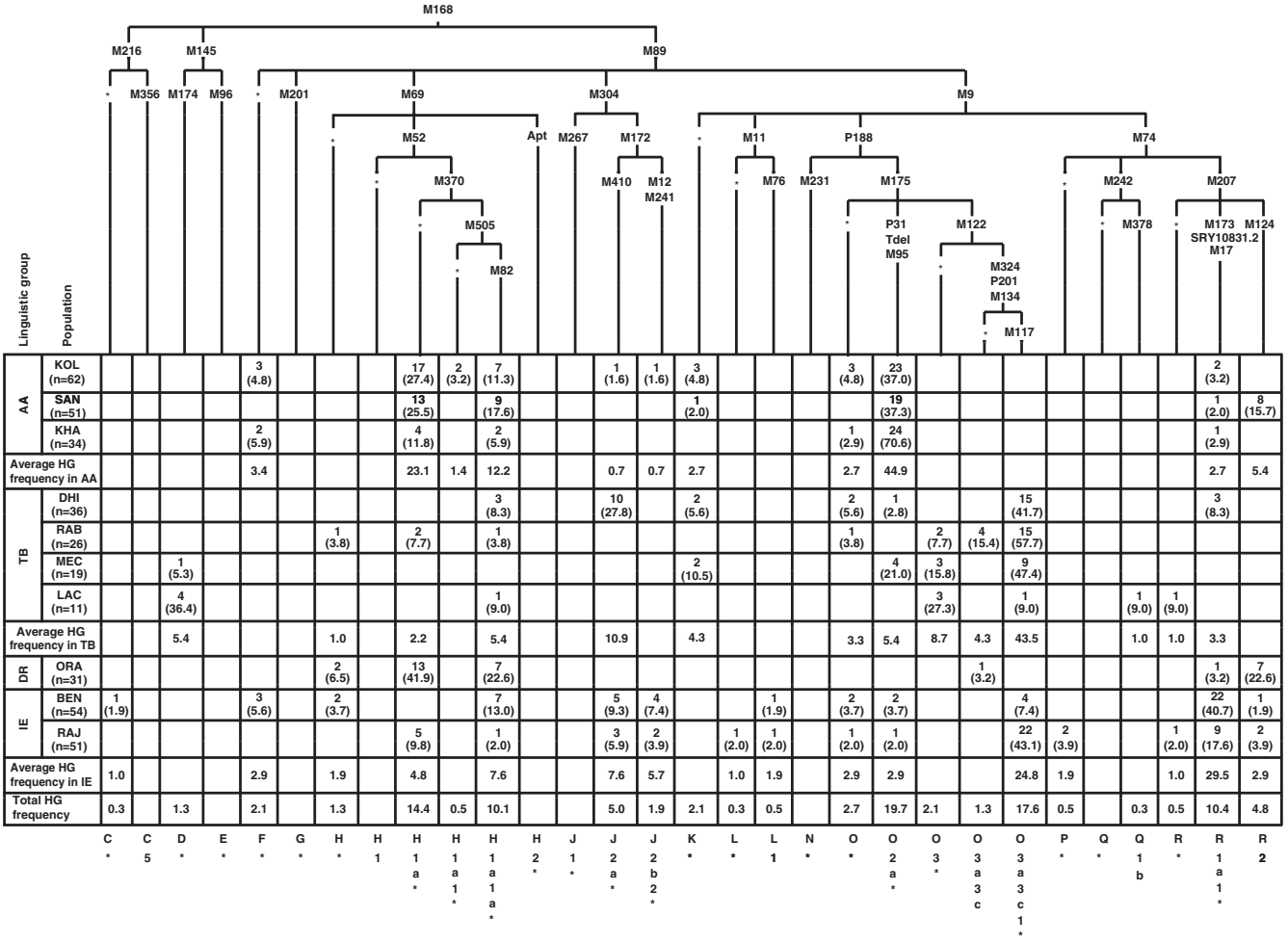

Figure 2 The hierarchical phylogenetic relationships, actual number and percent frequencies of the 22 paternal haplogroups observed in Terai and Duars populations. The average haplogroup frequencies in each linguistic group have been shown separately. AA, Austro-Asiatic; BEN, Bengali; DHI, Dhimal; DR, Dravidian; HG, haplogroup; IE, Indo-European; KHA, Kharia; KOL, Kol; LAC, Lachungpa; MEC, Mech; ORA, Oraon; RAB, Rabha; RAJ, Rajbanshi; SAN, Santhal.

such as M36, M97 and M39 of H1a1a*-M82, but none of the M82-positive chromosomes are found to exhibit these SNPs. In addition to this, none of the individuals of the studied groups showed the presence of H2*-Apt.

The clade H defined by M69 marker is very common in many Indian populations. It is represented by two sub-haplogroups, H1*-M52 and H2*-Apt along with eight additional downstream markers. Multiple studies based on the frequency distribution in different populations suggested that H*-M69, its sub-clades H1*-M52 and H2*-Apt have Indian origin.[13,14,33] Of these lineages, H1*-M52 has peak variance distribution in Western India, whereas H2*-Apt and H*-M69 display higher variance patterns in South and Northeastern India, respectively.[13]

The identification of new SNPs not only increases the resolution of the phylogenetic tree but also helps to trace the prehistoric migrations more accurately and distinguish sub-population within the haplogroup. As H haplogroup has been postulated to have indigenous origin, extensive high-resolution genotyping of SNPs as well as their associated short tandem repeat (STR) markers in populations across India would help to understand its exact place and time of origin. Moreover, as a non-East Asian lineage, HG-H could be highly informative for understanding not only the peopling of India but also the spread of this lineage to other geographical areas, as this has also been reported with similar high frequencies in Malaysian Indian and Hungarian populations.[34]

Haplogroup R

The third-most abundant haplogroup is R, accounting for an average frequency of 15.7%. The high incidence of this haplogroup is restricted only to the IE- and DR-speaking groups. Within these populations, IE-speakers exhibit an average frequency of 29.5% of R1a1*, whereas DR-speakers account for 22.6% of R2 sub-haplogroup.

Other haplogroups

Apart from these three major clades, we have also observed few other important haplogroups, such as C, D, F, J, K, L, P and Q, in the studied populations. Among these haplogroups, only J shows an average frequency of 6.9%, whereas the remaining seven haplogroups such as C (0.3%), D (1.3%), F (2.1%), K (2.1%), L (0.8%), P (0.5%) and Q (0.3%) are found to have negligible frequencies.

The PC analyses of the haplogroup frequencies in 35 populations (25 reference plus 10 from the present study) examined in this study (Table 1) have been depicted in Figure 3. In the plot, all the South Asian groups except Kharia, Rabha, Mech and Lachungpa clustered together on the upper left quadrant of the graph. Lachungpa plots

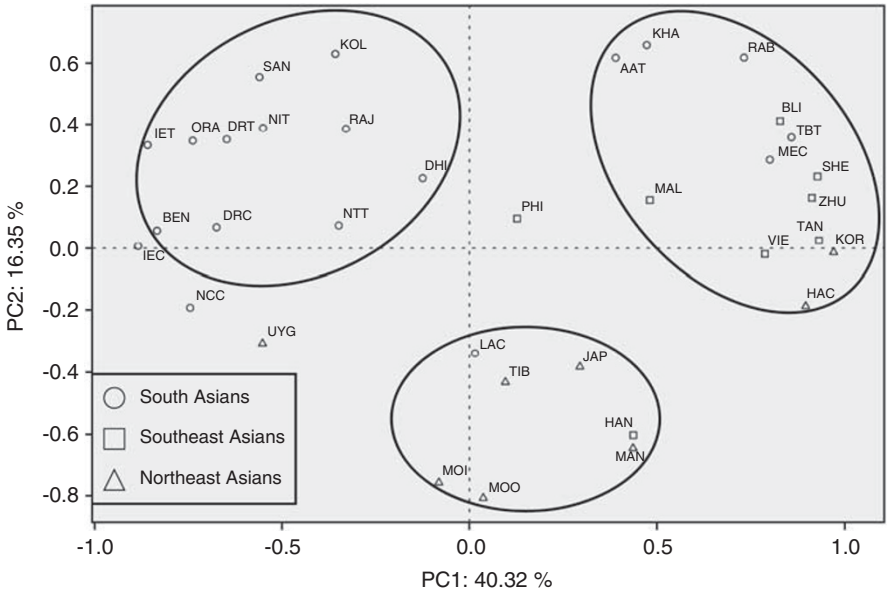

Figure 3 Principal component analysis (PCA) of populations in South, Southeast and Northeast Asia (Table 1). PC map is based on Y-haplogroup frequencies. The absolute frequencies and information on populations have been presented in the Supplementary Table S3. The frequencies were transformed as Richards *et al.*[35] suggested to standardize against the different effect of genetic drift on haplogroups of different frequencies. The contribution of each haplogroup was calculated as the factor for PC1 and PC2 with regression method in SPSS13.0 software.

within the NEAS cluster, which may be because of extensive sharing of haplogroup D. The reason for the assembly of the remaining three ethnic groups within the SEAS cluster may be due to sharing of O haplogroup. Apart from this, it was interesting to note that two NEAS populations, that is, Korean and Han Chinese plotted closer to SEAS cluster, whereas Hani plotted within NEAS cluster.

Sub-Himalayan gene pools: reservoirs of variation

The gene pools of the sub-Himalayan ethnic groups are found to be heterogeneous. This is partly contributed by paternal lineages specific to different geographic regions. Although majority of the lineages are derived mainly from East and South Asia, the incidence of extensive admixture among the linguistic groups unequivocally made the genetic architecture of the studied populations very interesting.

East and South Asian connection of the TB-speaking sub-Himalayan tribes. Altogether, 14 haplogroups are observed in the sub-Himalayan TB tribes considered for the present study. Among them, O3 lineage and its derivatives dominate the haplogroup composition with average frequency of 56.5%. After genotyping several sub-haplogroups of O3 lineage, it is found that the downstream mutation, that is, O3a3c1-M117 of O3a3c-M134 to be most prevalent with the frequency ranging from 9.0–57.7%. It is evident from the earlier studies that O3a3c-M134 has high frequencies in the TBs from Northeast India[12–14] as well as Nepal Himalaya.[16] The undifferentiated O3*-M122 is a dominant East Asian lineage (average frequency 44.3%) and proposed to have southern origin in East Asia followed by northward migration around 25 000–30 000 years ago.[36] Regarding the origin of TB language, genetic evidences suggest that proto-TB languages arose 5000–6000 years ago in East Asia. The upper and middle Yellow River basin was considered as the ancestral homeland of TB people, from there they marched westward and then southward, leading to the inhabitation of the Himalayas and peopling of Bhutan, Nepal and northeastern India.[3] It has also been suggested that the Baric branch of TB family was the first inhabitants of the Himalayan region. Age estimation of O3a3c-M134 indicates that though it has a more ancient age (25.36 ± 1.59 KYA (thousand years ago)) in East Asian populations,[36] in Himalayan populations it is found to have a relatively recent age of 8.1 ± 2.9 KYA,[16] which further supports Neolithic settlement of TB groups in the Himalaya. Based on these evidences, it can be suggested that the sub-Himalayan TB-speaking groups also might have received a recent East Asian genetic contribution. East Asian influence on the sub-Himalayan gene pools has also been supported by a recent mitochondrial DNA-based study in some of the sub-Himalayan TB tribes.[18] These findings although demonstrate East Asian connection of the sub-Himalayan TB-speakers, but to know the exact time period of TB settlement in sub-Himalayan regions further studies considering Y-STR markers are warranted.

Based on the high frequency of D*-M174, earlier it was hypothesized that the Tibetan gene pool bears affinity with the Central Asians. This is contradicted by a recent study pointing southern origin of this haplogroup in East Asia.[11] In the TB-speaking Himalayan collections, although D*-M174 has been found to be negligibly low[12,16] but to a startling surprise, high frequencies (8–65%) of YAP have been reported in some Northeast Indian tribes.[37] In the present study, we detected D*-M174 in two TB-speaking populations, Mech (5.3%) and Lachungpa (36.4%). The presence of this lineage in these populations though supports East Asian influence on the sub-Himalayan gene pools but considering its southern region in East Asia and its subsequent northward migration about 60 000 years ago,[11] raised question about the arrival time of this lineage in the sub-Himalayan regions. Moreover, D*-M174 is not as predominantly distributed as O3. It is plausible that the more recent and extensive invasion by O3-bearing groups might have wiped out D*-M174, as also pointed out by Shi *et al.*[11] The sporadic presence of D*-M174 in the sub-Himalayan populations suggests that they might have retained the traces of past events of population migrations and interactions.

In addition to O3 lineage, we also observed O2 with a frequency of 21.0% in TB-speaking Mech. Though it is completely absent in

majority of the TB-speakers living in the Tibetan plateau and Nepal Himalaya, few studies showed it to be in high frequency in some of the TB groups from Eastern Himalaya.[3,13,14,16,38] The infrequent occurrence of O2*-M95 in the TBs might be due to male gene flow from the AA-speaking populations found across the sub-Himalayan and Northeast Indian regions rather than speakers from Southeast Asia carrying this.

It is noteworthy that the West Eurasian- or Indian-specific haplogroups such as H, J and R are also found in the TB-speaking individuals. The most prominent is J2a*-M410 sub-haplogroup with a frequency of 27.8% in Dhimal, albeit found to be absent or negligibly present in other TBs from India.[13] Incidentally, it was found to be 6.1% in one of the TB groups (Newar) from Nepal.[16] In a previous study, it has been demonstrated that with the exception of AA tribes having ~11% of J2b, the J2 clade is nearly absent in other Indian tribals, and has about 1% frequency in East Asian populations.[13] In addition to this, an Indian-specific upper caste marker R1a1*-M17, which is known to be either absent or, if present, only with negligible frequency in Indian tribal groups, is found with a frequency of 8.3% in Dhimal population. Such pattern of haplogroup composition undoubtedly made the genetic history of Dhimal more complex. Although Dhimal bears predominant East Asian genetic signatures, they plot within South Asian cluster (Figure 3). This scenario can be explained by the fact that the caste populations living in close proximity for several generations might have exerted substantial genetic influence on the Dhimal gene pool.

East Asian-specific paternal lineages in Indo-European caste populations of Terai. The IE-speaking caste populations, namely, Rajbanshi and Bengali together displayed 16 haplogroups, the highest number of paternal lineages observed when compared with other linguistic groups. Among them, the most abundant is R (33.3%), followed by O (30.5%), H (14.2%) and J (13.3%) haplogroups. Although most of the major haplogroups are found to be shared between Rajbanshi and Bengali, the frequencies of O ($P<0.001$) and R ($P<0.05$) differed significantly.

The incidence of R1a1*-M17 is much higher (40.7 vs 17.6%) in Bengali than Rajbanshi, where as O3a3c1*-M117 is comparatively more frequent in Rajbanshis (43.1 vs 7.4%) than Bengali. The elevated frequency of O3a3c1*-M117 in Rajbanshi raises question about its genetic ancestry and phylogenetic relationships with other populations, as O3 haplogroup in IE-speaking castes is seen to be totally absent in East India while negligibly present in other parts of India.[13,14] Rajbanshis have wide-spread distribution across East India and are considered to be original inhabitants of this region. Majority of Rajbanshis live in the Terai and Duars regions. Although historical evidences indicate their tribal connection,[39] the same notion has not been substantiated by genetic studies. On the basis of O3 haplogroup sharing between Rajbanshis (43.1%) and TB groups (average frequency 56.5%) of sub-Himalayan regions and Nepal Himalaya (average frequency 47.7%),[16] it might be concluded that the current paternal gene pool of Rajbanshi has received a significant TB influence. This observation partially supports the tribal especially TB connection of the Rajbanshi population. The presence of South-Asian-specific paternal lineages in Rajbanshis might have been the result of extensive male gene flow from IE- to TB-speaking groups. The IE languages have been introduced to the present day India relatively recently (~3.6 KYA),[40] whereas the age of O3 haplogroup in the adjoining TB-speaking Himalayan collections is 8.1 ± 2.9 KYA.[16] Based on these facts, it can be suggested that the settlement of TB-speaking groups in the sub-Himalayan regions might have occurred much before the arrival of the IEs.

Y-chromosome haplogroup diversities in AA- and DR-speaking tribes of the Terai and Duars. The three AA-speaking Mundari groups exhibit altogether 11 haplogroups, of which O2a*-M95 is the most predominant (frequency range 37.0–70.6%). We have also detected T del mutation while genotyping O2*-P31 at nucleotide position 133 in the 6T stretch, which lies immediately after P31 T to C transition at nucleotide position 127 in all the O2-derived samples. This T del change has also been observed recently by Fornarino *et al.*[26] in Tharus of Nepal and one tribal group from Andhra Pradesh, India. They further showed that these samples were positive for PK4, which was identified in the Pakistani Pathans.[41] When we typed M88 and PK4, the downstream SNP markers of M95, none of the samples are found to possess these markers.

Haplogroup O2a*-M95 is known to dominate the haplogroup composition of the AA populations in India and Southeast Asia.[13,19,38] Although AA-speaking populations are considered as the oldest groups in India, there are many competing theories on their origin and subsequent settlement. The AAs have three sub-families such as Mundari, Mon-Khmer and Khasi-Khmuic in India and they display region-specific distribution. A recent study has proposed that O2a*-M95 had originated ~65 000 yrs bp in Mundari-speaking Indian AA tribes.[38] Unlike Eastern Himalaya where majority of the AA-speaking tribes are Khasi-Khmuic, the sub-Himalayan Terai and Duars regions are mainly inhabited by the Mundari speakers. Considering the age of O2a*-M95 in other Mundari groups, it might be possible that the AA-speaking Mundari tribes were the first to reach the sub-Himalayan regions. The presence of this haplogroup in some of the TB-speaking groups reflect gene flow from AA- to TB-speaking groups who might have inhabited these regions much later than the AA speakers.

The second most abundant (36.7%) haplogroup in the Mundari tribes is H, which has been reported to be comparatively less frequent (25.3%) in Mundari-speaking AA groups.[38] In addition to this, R2-M124 is also observed to be higher (15.7%) in Santhal population compared with its average frequency (4.9%) in Indian Mundari groups.[38] It is worth noting that, Sahoo *et al.*[14] found it to be entirely absent in the Santhals from East India.

In the DR-speaking Oraon tribe, H has been the most abundant haplogroup with a frequency of 71.0%, followed by R2*-M124 (22.6%). Although DR tribes were shown in a number of studies to have high incidence of H (average 40%), the frequency range of R2*-M124 varied from 8 to10%.[13,33] Similarly, the low frequency of R2 (<6%) has also been noticed in other linguistic such as TB and AA groups from Eastern Himalayan regions.[13,38] Surprisingly, R2 was found to be elevated (25.8%) in neighboring TB-speaking Newar population from Nepal.[16] Although, R2 has been suggested to have originated in Indian subcontinent,[13] the exact place of its origin is yet to be deciphered.

In addition to these haplogroups, no other Indian-specific clade could be detected in Oraon. It is noteworthy that although O2*-M95 has a high frequency (26.6%) in DR tribes,[13] it is found to be completely absent in Oraon from the Terai region. DR-speaking tribes are mostly concentrated in the Southern India, whereas they are sparsely distributed in Eastern India. The quite unusual pattern of haplogroup diversity and complete absence of O2* in the Oraon tribe could be the result of geographic isolation or a case founder effect.

The genetic ancestry of AA- and DR-speaking tribes is well established and distinct. Our findings on the Y-lineages are suggestive of the fact that there have been extensive population interactions and substantial gene flow among these groups in the sub-Himalayan regions.

CONCLUSION

In conclusion, the paternal lineages that have been observed in the ethnic groups from the sub-Himalayan Terai and Duars reveal a number of interesting facts (i) the sub-Himalayan gene pool is extremely heterogeneous, TB groups though received significant SEAS contribution, the genetic influence from Northeast and South Asian has also been substantial; (ii) one of the IE-Speaking castes, Rajbanshi might share a common ancestry with the TBs and, finally, (iii) the sub-Himalayan regions have experienced multiple events of population migrations and interactions and some of the studied groups bear more close genetic ties with the Himalayan collections. Further studies considering Y-STR markers are required to be carried out to gain better insights about the time of settlement of various groups into the sub-Himalayan regions.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We gratefully acknowledge Dr Peter Underhill of Stanford University for his constructive suggestions and criticism on our work. We thank Drs Min-Sheng Peng and Hai-Bing Xie of Kunming Institute of Zoology for their assistance in the analysis of some of the data. We also thank Dr Simona Fornarino, Pasteur Institute, Paris, for her help in constructing the phylogenetic tree. This study is partly supported by a grant from the China Postdoctorate Science Fund (No. 20060400308) to Dr Monojit Debnath.

1 Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton University Press, Princeton, NJ, USA, 1994).
2 Su, B., Xiao, J., Underhill, P., Deka, R., Zhang, W., Akey, J. *et al.* Y-chromosome evidence for a northward migration of modern humans into eastern Asia during the last Ice Age. *Am. J. Hum. Genet.* **65,** 718–1724 (1999).
3 Su, B., Xiao, C., Deka, R., Seielstad, T., Kangwanpong, D., Xiao, J. *et al.* Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. *Hum. Genet.* **107,** 582–590 (2000).
4 van Driem, G. L. *Languages of the Himalayas: an Ethnolinguistic Handbook of the Greater Himalayan Region Containing an Introduction to the Symbiotic Theory of Language* (Brill, Leiden, The Netherlands, 2001).
5 van Driem, G. L. in *Examining the Farming/Language Dispersal Hypothesis* (eds Bellwood, P. & Renfrew, C.) 233–249 (McDonald Institute for Archaeological Research, Cambridge, UK, 2002).
6 Aldenderfer, M. & Yinong, Z. The prehistory of the Tibetan plateau to the seventh century AD.: perspectives and research from China and the West since 1950. *J. World Prehist.* **18,** 1–55 (2004).
7 Zhao, M., Kong, Q. P., Wang, H. W., Peng, M. S., Xie, X. D., Wang, W. Z. *et al.* Mitochondrial genome evidence reveals successful late Paleolithic settlement on the Tibetan Plateau. *Proc. Natl Acad. Sci. USA* **106,** 21230–21235 (2009).
8 van Driem, G. L. in *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics* (eds Sagart, L., Blench, R. & Sanches-Mazas, A.) 81–106 (Routledge Curzon, Taylor and Francis group, London, 2005).
9 Wen, B., Xie, X., Gao, S., Li, H., Shi, H., Song, X. *et al.* Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. *Am. J. Hum. Genet.* **74,** 856–865 (2004).
10 Qian, Y., Qian, B., Su, B., Yu, J., Ke, Y., Chu, Z. *et al.* Multiple origins of Tibetan Y chromosomes. *Hum. Genet.* **106,** 453–454 (2000).
11 Shi, H., Zhong, H., Peng, Y., Dong, Y. L., Qi, X. B. & Zhang, F. Y Chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. *BMC Biol.* **6,** 45 (2008).
12 Cordaux, R., Weiss, G., Saha, N. & Stoneking, M. The northeast Indian passageway: a barrier or corridor for human migrations? *Mol. Biol. Evol.* **21,** 1525–1533 (2004).
13 Sengupta, S., Zhivotovsky, L. A., King, R., Mehdi, S. Q., Edmonds, C. A. *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).
14 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).
15 Parkin, E. J., Kraayenbrink, T., van Driem, G. L., Tshering Of Gaselo, K., de Knijff, P. & Jobling, M. A. 26-Locus Y-STR typing in a Bhutanese population sample. *Forensic Sci. Int.* **161,** 1–7 (2006).
16 Gayden, T., Cadenas, A. M., Regueiro, M., Singh, N. B., Zhivotovsky, L. A., Underhill, P. A. *et al.* The Himalayas as a directional barrier to gene flow. *Am. J. Hum. Genet.* **80,** 884–894 (2007).
17 Gayden, T., Mirabal, S., Cadenas, A. M., Lacau, H., Simms, T. M., Morlote, D. *et al.* Genetic insights into the origins of Tibeto-Burman populations in the Himalayas. *J. Hum. Gen.* **54,** 216–223 (2009).
18 Chandrasekar, A., Kumar, S., Sreenath, J., Sarkar, B. N., Urade, B. P. & Mallick, S. Updating phylogeny of mitochondrial DNA macrohaplogroup M in India: dispersal of modern human in South Asian corridor. *PLoS One* **4,** e7447 (2009).
19 Reddy, B. M., Langstieh, B. T., Kumar, V., Nagaraja, T., Reddy, A. N. S. *et al.* Austro-Asiatic tribes of Northeast India provide hitherto missing genetic link between South and Southeast Asia. *PLoS One* **2,** e1141 (2007).
20 Debnath, M. & Chaudhuri, T. K. Distribution of HLA-A and B loci allele in Toto: a sub-Himalayan vanishing Indian tribe. *Tissue Antigens* **67,** 64–65 (2006).
21 Agrawal, S., Srivastava, S. K., Borkar, M. & Chaudhuri, T. K. Genetic affinities of north and northeastern populations of India: inference from HLA-based study. *Tissue Antigens* **72,** 120–130 (2008).
22 Y Chromosome Consortium. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res.* **12,** 339–348 (2002).
23 Karafet, T. M., Mendez, F. L., Meilerman, M. B., Underhill, P. A., Zegura, S. L. & Hammer, M. F. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res.* **18,** 830–838 (2008).
24 Underhill, P. A., Passarino, G., Lin, A. A., Shen, P., Mirazon Lahr, M., Foley, R. A. *et al.* The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* **65,** 43–62 (2001).
25 Zhao, Z., Khan, F., Borkar, M., Herrera, R. & Agrawal, S. Presence of three different paternal lineages among North Indians: a study of 560 Y chromosomes. *Ann. Hum. Biol.* **36,** 46–59 (2009).
26 Fornarino, S., Pala, M., Battaglia, V., Maranta, R., Achilli, A., Modiano, G. *et al.* Mitochondrial and Y-chromosome diversity of the Tharus (Nepal): a reservoir of genetic variation. *BMC Evol. Biol.* **9,** 154 (2009).
27 Karafet, T. M., Hallmark, B., Cox, M. P., Sudoyo, H., Downey, S., Lansing, J. S. *et al.* Major east-west division underlies Y chromosome stratification across Indonesia. *Mol. Biol. Evol.* **27,** 1833–1844 (2010).
28 Karafet, T. M., Lansing, J. S., Redd, A. J., Reznikova, S., Watkins, J. C., Surata, S. P. *et al.* Balinese Y-chromosome perspective on the peopling of Indonesia: genetic contributions from pre-neolithic hunter-gatherers, Austronesian farmers, and Indian traders. *Hum. Biol.* **77,** 93–114 (2005).
29 Karafet, T., Xu, L., Du, R., Wang, W., Feng, S., Wells, R. S. *et al.* Paternal population history of East Asia: sources, patterns, and microevolutionary processes. *Am. J. Hum. Genet.* **69,** 615–628 (2001).
30 Zhong, H., Shi, H., Qi, X. B., Duan, Z. Y., Tan, P. P., Jin, L. *et al.* Extended Y-chromosome investigation suggests post-Glacial migrations of modern humans into East Asia via the northern route. *Mol. Biol. Evol.* **28,** 717–727 (2011).
31 Hammer, M. F., Karafet, T. M., Park, H., Omoto, K., Harihara, S., Stoneking, M. *et al.* Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. *J. Hum. Genet.* **51,** 47–58 (2006).
32 Xue, Y., Zerjal, T., Bao, W., Zhu, S., Shu, Q., Xu, J. *et al.* Male demography in East Asia: a north-south contrast in human population expansion times. *Genetics* **172,** 2431–2439 (2006).
33 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).
34 Volgyi, A., Zalan, A., Beres, J., Chang, Y. M. & Pamjav, H. Haplogroup distribution of Hungarian population and the largest minority group. *Forensic Sci. Int. Genetics Supplement Series* **1,** 383–385 (2008).
35 Richards, M., Macaulay, V., Torroni, A. & Bandelt, H. J. In search of geographical patterns in European mitochondrial DNA. *Am. J. Hum. Genet.* **71,** 1168–1174 (2002).
36 Shi, H., Dong, Y. L., Wen, B., Xiao, C. J., Underhill, P. A., Shen, P. D. *et al.* Y chromosome evidence of southern origin of the East Asian-specific haplogroup O3-M122. *Am. J. Hum. Genet.* **77,** 408–419 (2005).
37 Chandrasekhar, A., Saheb, S. Y., Gangopadyaya, P., Gangopadyaya, S., Mukherjee, A., Basu, D. *et al.* YAP insertion signature in South Asia. *Ann. Hum. Biol.* **34,** 582–586 (2007).
38 Kumar, V., Reddy, A. N. S., Babu, J. P., Rao, T. N., Langstieh, B. T., Thangaraj, K. *et al.* Y-chromosome evidence suggests a common paternal heritage of Austro-Asiatic populations. *BMC Evol. Biol.* **7,** 47 (2007).
39 Hunter, W. W. *A Statistical Account of Bengal, Vol-X: District of Darjeeling, Jalpaiguri and State of Kuch Behar* (D. K. Publishing House, New Delhi, 1974).
40 Beekes, R. S. P. *Comparative Indo-European linguistics: An Introduction* (J Benjamins, Amsterdam/Philadelphia, 1995).
41 Mohyuddin, A., Ayub, Q., Underhill, P. A., Tyler-Smith, C. & Mehdi, S. Q. Detection of novel Y SNPs provides further insights into Y chromosomal variation in Pakistan. *J. Hum. Genet.* **51,** 375–378 (2006).

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