European Journal of Human Genetics (2003) 11, 253–264

www.nature.com/ejhg

ARTICLE

Mitochondrial DNA analysis reveals diverse histories of tribal populations from India

Richard Cordaux*[,1], Nilmani Saha[2], Gillian R Bentley[3], Robert Aunger[4], SM Sirajuddin[5] and Mark Stoneking[1]

[1]*Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany;* [2]*Department of Pediatrics, National University of Singapore, Singapore;* [3]*Department of Anthropology, University College London, UK;* [4]*Department of Biological Anthropology, Cambridge, UK;* [5]*Anthropological Survey of India, Mysore, Karnataka State, India*

We analyzed 370 bp of the first hypervariable region of the mitochondrial DNA (mtDNA) control region in 752 individuals from 17 tribal and four nontribal groups from the Indian subcontinent, to address questions concerning the origins, genetic structure and relationships of these groups. Southern Indian tribes showed reduced diversity and large genetic distances, both among themselves and when compared with other groups, and no signal of prehistoric demographic expansions. These results probably reflect enhanced genetic drift because of small population sizes and/or bottlenecks in these groups. By contrast, northern groups exhibited more diversity and signals of prehistoric demographic expansions. Phylogenetic analyses revealed that southern and northern groups (except northeastern ones) have related mtDNA sequences albeit at different frequencies, further supporting the larger impact of drift on the genetic structure of southern groups. The Indian mtDNA gene pool appears to be more closely related to the east Eurasian gene pool (including central, east and southeast Asian populations) than the west Eurasian one (including European and Caucasian populations). Within India, northeastern tribes are quite distinct from other groups; they are more closely related to east Asians than to other Indians. This is consistent with linguistic evidence in that these populations speak Tibeto-Burman languages of east Asian origin. Otherwise, analyses of molecular variance suggested that caste and tribal groups are genetically similar with respect to mtDNA variation.
European Journal of Human Genetics (2003) **11,** 253–264. doi:10.1038/sj.ejhg.5200949

Keywords: mitochondrial DNA; hypervariable region 1; India; tribes

Introduction

Archeological, fossil and genetic evidence points to a major expansion of anatomically modern humans out of Africa some 100 000 years ago,[1,2] but the migration routes remain poorly understood. In this respect, the Indian subcontinent is considered to be a crucial geographic area for human migrations,[3,4] since it is located at the crossroads of Africa, the Pacific and west and east Eurasia.

More than one billion people with enormous morphological, genetic, cultural and linguistic diversity inhabit the Indian subcontinent.[5,6] At least four potential sources of genes contributing to the current Indian gene pool can be envisaged.[3] The first one is an old Paleolithic component, probably almost extinct nowadays. The second component would have witnessed early Neolithic migrations of farmers from the eastern horn of the Fertile Crescent, probably speaking proto-Dravidian languages. The third source is responsible for the arrival of Indo-European speakers ~3500 years ago, who most probably introduced the caste

*Correspondence: Dr R Cordaux, Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany. Tel: +49 341 9952 537; Fax: +49 341 9952 555; E-mail: cordaux@eva.mpg.de
Received 30 July 2002; revised 19 November 2002; accepted 3 December 2002

system that hierarchically organized the vast majority of Indian society. The fourth component is associated with Austro-Asiatic and Tibeto-Burman speakers inhabiting east and northeast India, with ties to east Asia.

The molecular genetic data generated so far concerning the people of the Indian subcontinent have largely focused on caste populations rather than on tribal groups.[7–13] According to the 1991 census, ~8% of the Indian population belong to tribal communities.[14] They represent minorities that have not been absorbed by the caste system.[3] They are generally thought to be the aboriginal inhabitants of the Indian subcontinent that were present in the region before the arrival of Indo-European speakers. There are currently about 400 tribes in India that vary in size from a few hundred to a few million; they speak languages belonging to all four of the major language families represented in India (Austro-Asiatic, Dravidian, Indo-European and Tibeto-Burman). Their origins and genetic affinities remain largely unknown, although such information is of primary importance in understanding the possible role of India in early migrations of modern humans, since any remnants of genetic contributions from pre-Indo-European migrants would presumably be present in tribal populations rather than in caste populations.

The molecular genetic evidence on Indian tribal origins and relationships is rather scanty. The mitochondrial DNA (mtDNA) intergenic COII/tRNA^{Lys} 9-bp deletion marker, present at high frequency in some Asian[15] and African[16] populations, is found at low frequencies in India and arose multiple times independently.[17–19] Based on a study of autosomal and mtDNA markers in eight Indian tribes speaking Austro-Asiatic, Dravidian or Tibeto-Burman languages, it was concluded that these different language groups represented distinct founding groups, with Austro-Asiatic speakers being the most ancient inhabitants of the region.[19]

With the aim of furthering our understanding of their origins and relationships, we report here a comprehensive study of mtDNA variation in Indian tribal groups. We analyzed sequences of the first hypervariable segment (HV1) of the mtDNA control region in more than 750 individuals belonging to different language families and compared these to available data from caste populations and from world populations, in order to obtain a global picture of the genetic structure and the relationships of populations inhabiting the Indian subcontinent.

Subjects and methods

Samples

DNA samples from 752 individuals belonging to 21 populations inhabiting the Indian subcontinent were analyzed (Table 1, Figure 1). Additional information on the 17 tribal populations and the four nontribal populations can be found elsewhere.[18,20] Published data were also included from 10 tribal populations,[12,19] one Indo-European-speaking caste from north India,[12] eight Dravidian-speaking castes from Andhra Pradesh,[10] and two castes from south India.[7] Although they were sampled in the southern state of Andhra Pradesh, the Andh, Pardhi, Thoti and Lambadi tribes as well as the eight Dravidian-speaking castes[10] will be referred to as central Indian populations in the following text, in order to differentiate them from the populations sampled in the states of Kerala, Karnataka and Tamil Nadu (Figure 1).

mtDNA sequencing

The mtDNA HV1 was amplified using primers L15996 (Vigilant *et al*[21]) and H16410 (Handt *et al*[22]). Purified PCR products were sequenced on both strands using primers L16001 (5′-ATTAGCACCCAAAGCTAAGA-3′) and H16401 (Vigilant *et al*[21]). Sequencing reactions were prepared with the BigDye Terminator Cycle Sequencing Kit (Applied Biosystems, Foster City, CA, USA) and purified by isopropanol precipitation, according to the supplier's recommendations. Sequencing reactions were resolved on either ABI 377 or 3700 automated DNA sequencers (Applied Biosystems). The HV1 sequences were deposited in the HvrBase[23] and are also available from the authors.

Data analysis

HV1 sequences were manually aligned with the published reference sequence.[24] After removal of sites with insertions and deletions, the software package ARLEQUIN version 2.000 (Schneider *et al*[25]) was used to calculate haplotype and nucleotide diversity and their standard deviations (SD), mismatch distributions, mean pairwise differences and their SD, Fu's[26] Fs statistic and associated *P*-values based on 1000 simulated samples, raggedness index r[27] and Fst distances between pairs of populations and associated *P*-values based on 1000 permutations. Analyses of molecular variance[28] (AMOVA) were performed using ARLEQUIN to evaluate the genetic structure of the populations, with the significance of variance components tested with 10 000 permutations.

Multidimensional scaling (MDS) was performed by means of STATISTICA, based on Fst distances. Populations from the Indian subcontinent were compared to worldwide populations from Africa (!Kung,[29] Dinka,[30] Turkana and Yoruba[31]), Europe (Germans,[32] Bulgarians and Turks[33]), Caucasus (Armenians, Azerbaijanians, Georgians and Kabardinians[34]), central Asia (Kazakh, Kirghiz and Kirghiz-Talas[35]), east Asia (Han Chinese-Changsha, Han Chinese-Xi'an, Tottori,[36] Japanese,[37] Koreans[38] and Mongolians[39]), southeast Asia (Akkha, White and Red Karen, Lahu, Lisu-Chiang Rai, Lisu-Mae Hong Son[40] and Vietnamese[36]), Australia (from Arnhem Land, Sandy Desert,[41] Desert and Riverine[42]), and island southeast Asia (Indonesians,[29,43] coastal and highland Papua New Guineans[41] (PNG)).

Table 1 Relevant information for 31 populations from the Indian subcontinent

Ethnic groups	*Sample size*	*Linguistic affiliation*	*Population size* (× 1000)	*Reference*
Tribals				
North India				
Tharu	12	Indo-European	95	Kivisild *et al*[12]
Buksa	18	Indo-European	43	Kivisild *et al*[12]
Northeast India				
Adi	45	Tibeto-Burman	110	Present study
Apatani	52	Tibeto-Burman	23	Present study
Nishi	52	Tibeto-Burman	261	Present study
Naga	43	Tibeto-Burman	1400	Present study
Tipperah	20	Tibeto-Burman	105	Roychoudhury *et al*[19]
East India				
Lodha	14	Austro-Asiatic	75	Roychoudhury *et al*[19]
Munda	6	Austro-Asiatic	2000	Roychoudhury *et al*[19]
Santal	14	Austro-Asiatic	6000	Roychoudhury *et al*[19]
Central India				
Andh	40	Indo-European	80	Present study
Pardhi	42	Indo-European	18	Present study
Thoti	39	Not available	Not available	Present study
Lambadi	86	Indo-European	2000	Kivisild *et al*[12]
South India				
Jenukurumba	6	Dravidian	35	Present study
Kattunaiken	16	Dravidian	26	Present study
Soligas	14	Dravidian	16	Present study
Koragas	53	Dravidian	16	Present study
Kuruchian	46	Dravidian	22	Present study
Mullukurunan	44	Dravidian	20	Present study
Mullukurumba	17	Dravidian	6	Present study
Bettakurumba	19	Dravidian	10	Present study
Paniya	17	Dravidian	6	Present study
Yerava	53	Dravidian	19	Present study
Irula	14	Dravidian	75	Roychoudhury *et al*[19]
Kurumba	10	Dravidian	150	Roychoudhury *et al*[19]
Kota	25	Dravidian	2	Roychoudhury *et al*[19]
Non-tribals				
Pushtoons	36	Indo-European	9000	Present study
Northern Indians	40	Indo-European	400 000	Present study
Bangladeshis	29	Indo-European	120 000	Present study
Southern Indians	49	Dravidian	300 000	Present study

A gene tree was computed using the neighbor-joining method, based on *p*-distances, as implemented in MEGA version 2.1 (Kumar *et al*[44]). The tree was rooted using a Neanderthal HV1 sequence[45] and bootstrap analysis was carried out with 500 replications. In addition to Indian sequences, worldwide HV1 sequences[9] were included for comparison, as well as PNG highlanders, central Asians and Australians described above. We inferred the frequencies of some major mtDNA haplogroups in our data from the resulting tree topology, from Indian HV1 sequences whose haplogroup information is available,[12] and from haplogroup-diagnostic substitutions in HV1.[46]

Results

Diversity indices and demographic parameters

Sequence data corresponding to nucleotide positions (np) 16 022–16 391 in the reference sequence[24] were obtained from 752 individuals. Nucleotide substitutions were observed at 153 sites, which defined 316 different HV1 sequences. Some individuals exhibited length variation between np 16 181 and 16 183 (these positions were removed from the subsequent analyses). Deletions were observed at five sites (np 16 166, 16 179, 16 194, 16 195 and 16 258) and insertions at two sites (a C between np 16 169 and 16 170 and an A between np 16 189 and 16 190).

Figure 1 Map of the Indian subcontinent indicating approximate locations of populations used in this study.

The 752 HV1 sequences from the present study were subsequently analyzed together with 219 previously published sequences from 10 Indian tribes, enabling a more comprehensive study of HV1 variation in Indian tribal populations. Since small sample sizes could potentially affect the reliability of the analyses, some populations were pooled. Pooling was performed according to several criteria including geographical proximity, linguistic affiliation, historical record and population relationships deduced from Fst distances. For example, the Mullukurunan and Mullukurumba are both south Indian tribes speaking a Kannada language (a Dravidian language-subfamily). Some scholars have argued that they are the same although they nowadays live in different areas, and it has been reported that the Mullukurunan are also known as 'Mullu Kurumba'.[47] Moreover, the two tribes were separated by an Fst distance of −0.030, which is not significantly different from zero ($P=0.84$). Therefore the data from these two groups were pooled. In summary, this approach resulted in 23 groups, each of which was composed of at least 20 individuals (Table 2).

Diversity indices and demographic parameters estimated for these groups are reported in Table 2. Overall, haplotype diversity in Indian tribals ranged from 0.671 to 0.995 and nucleotide diversity from 0.005 to 0.023. Haplotype diversity was significantly higher (Mann–Whitney *U*-test: $Z=3.24$, $P<0.01$) in north, east and northeast India (0.940–0.995) than in south India (0.671–0.939). Intermediate haplotype diversity values were observed in central India (0.884–0.985); they were not significantly different from north, east and northeast India ($Z=1.51$, $P=0.13$) or south India ($Z=1.70$, $P=0.09$). Similarly, nucleotide diversity in north, east and northeast India (0.014–0.023) was significantly higher ($Z=2.78$, $P<0.01$) than in south India (0.005–0.017). Again, central India exhibited intermediate values (0.012–0.017); they were not significantly different from north, east and northeast India ($Z=1.61$, $P=0.11$) or south India ($Z=1.78$, $P=0.07$). Therefore, north, east and northeast Indian tribes showed greater mtDNA diversity than south Indian tribes.

These patterns of genetic diversity in Indian tribes were further strengthened by the analysis of mean pairwise differences (MPD). MPD for south tribes (1.77–5.80) were significantly lower than MPD from north, east and northeast tribes (5.06–7.69; $Z=2.78$, $P<0.01$) or from central tribes (4.46–5.91; $Z=2.04$, $P=0.04$), whereas MPD from central and north, east and northeast tribes were not significantly different ($Z=1.70$, $P=0.09$). Mismatch distributions (Figure 2) were computed for the 14 tribal and four nontribal populations from Table 1 whose sample size is ≥ 20. Unimodal distributions were observed mostly in northeast tribes and some central tribes, whereas all of the south tribes exhibited multimodal distributions with a higher frequency of the low difference classes (0 and 1): 25–61% of the pairwise differences for the south tribes were in the 0/1 classes *vs* only 1–13% for the other tribes. Unimodal distributions are interpreted as signs of demographic expansions while multimodal distributions are interpreted as signs of constant population size over time.[27] Moreover, the peaks observed at 0/1 classes in the mismatch distributions indicated bottlenecks in these populations.[48] In parallel, the raggedness index *r* was generally less than 0.03 in north, east and northeast tribes but more than 0.07 in south tribes (Table 2). Values of *r* lower than 0.05 suggest demographic expansions while values of *r* greater than 0.05 are more consistent with constant population sizes.[27] Fu's Fs also support these patterns of demographic history in India (Table 2). Negative values of Fs that differ significantly from zero, indicative of population demographic expansions,[26] were obtained in 86% of north, east and northeast tribes, 50% of central tribes and only 25% of south tribes. Therefore, several approaches provided congruent evidence for different demographic histories in Indian tribes. In general, north, east and northeast tribes showed signs of expansion while south tribes, and to a lesser extent central tribes, were likely to have experienced bottlenecks and/or constant population sizes over time.

The four nontribal populations exhibited high gene and nucleotide diversities (Table 2). In addition, the Fs statistic

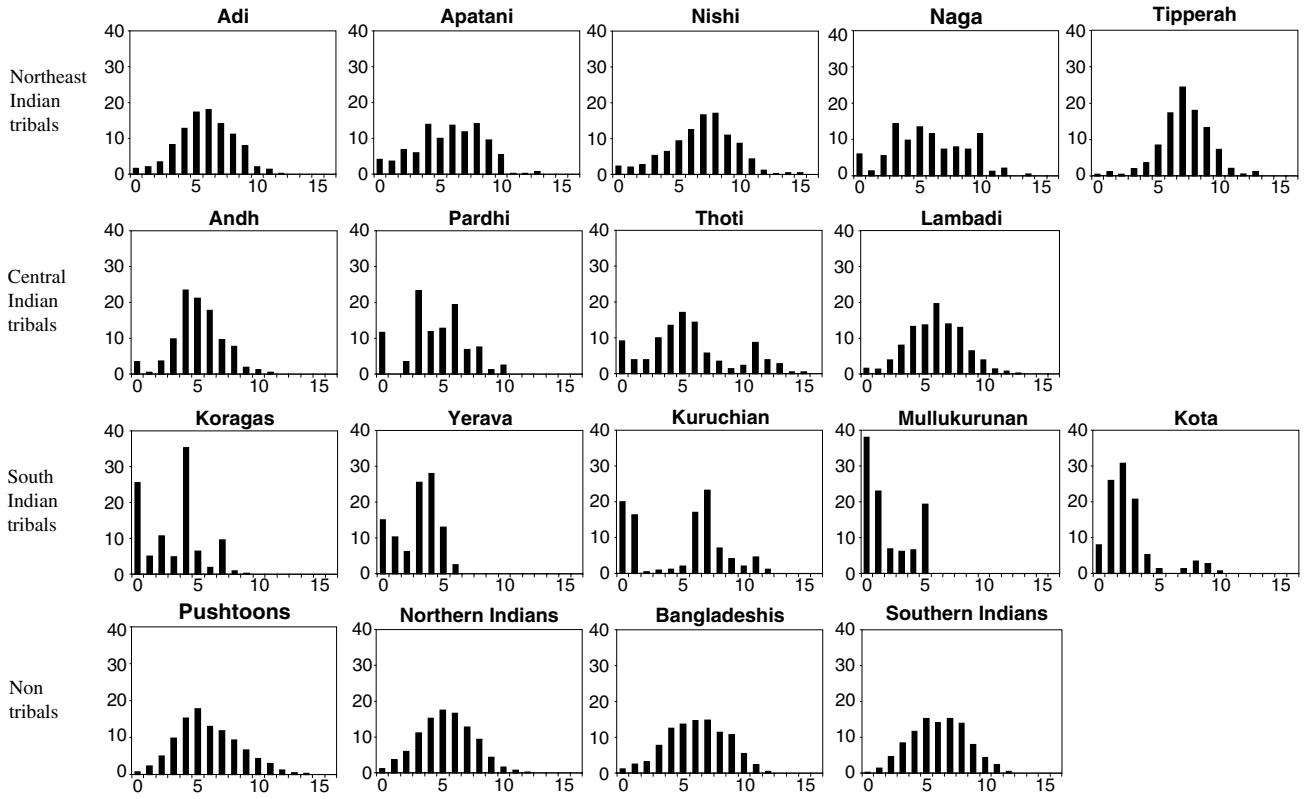


Figure 2 Mismatch distributions of mtDNA HV1 sequences from 18 populations from the Indian subcontinent. The number of nucleotide differences between pairs of sequences is indicated along the *x*-axis, and their frequency (%) is indicated along the *y*-axis.

as well as the raggedness index (Table 2) and mismatch distributions (Figure 2) suggested demographic expansions in these populations.

MDS analysis

MDS analysis based on Fst distances was performed to investigate relationships between Indian tribes, Indian castes and world populations. In the MDS plot (Figure 3), Indian tribal and caste populations both clustered together with other Eurasian and Australian populations. The Kung were a clear outlier, but removing them from the analysis did not significantly change the results, except as noted below (analysis not shown).

Among Indian populations, most of the south tribes were outliers within the Eurasian/Australian cluster. Further information on the affinities of Indian tribes was obtained by focussing on the Eurasian/Australian cluster in the MDS plot (Figure 4). Overall, Indian subcontinent populations were closer to east Eurasians (including central, east and southeast Asians) than to west Eurasians (including Europeans and Caucasians). Since the smaller number of European populations considered might have contributed to suggesting stronger affinities between India and east Eurasia, the analysis was repeated by adding six European populations (from Sweden, Austria, France, Britain, Italy and Spain). However, the trends in the MDS plot did not change (not shown). The smaller average Fst distance separating Indian populations from east Eurasians (0.081) than from west Eurasians (0.118) confirmed the aforementioned affinities. The removal of northeast tribes from the analysis led to similar average Fst values (0.087 for India–east Eurasia *vs* 0.116 for India–west Eurasia) and the removal of south tribes resulted in $\sim$25% smaller average Fst values (0.058 for India–east Eurasia *vs* 0.093 for India–west Eurasia). The removal of both northeast and south tribes was almost equivalent to removing only south tribes (0.061 for India–east Eurasia *vs* 0.083 for India–west Eurasia). In all cases, then, Indian populations showed closer affinities to east than to west Eurasians.

In contrast to other tribal groups, the five northeast Indian groups were closer to east Eurasian populations (average Fst distance: 0.049) than to other Indian tribes (average Fst distance: 0.118, dropping to 0.084 when south tribes were removed). The same trend was observed for east Indian tribes; however, when the !Kung were removed from the analysis, east Indian tribes were closer to other Indian groups than to east Asians in an MDS plot (not shown). Average Fst values also supported their closer

Table 2 Diversity and demographic parameters deduced from mtDNA HV1 sequences in India

Populations	N^a	n^b	*Haplotype diversity*[c]	*Nucleotide diversity*[c]	*MPD*[c,d]	*Fu's Fs*	r^e
Tribals							
North India							
North tribes[f]	30	28	0.995±0.010	0.014±0.008	5.06±2.53	−25.48*	0.014
Northeast India							
Adi	45	36	0.984±0.010	0.016±0.009	5.75±2.81	−25.33*	0.014
Apatani	52	28	0.958±0.012	0.016±0.008	5.73±2.79	−12.67*	0.018
Nishi	52	33	0.977±0.008	0.019±0.010	6.83±3.27	−17.74*	0.012
Naga	43	21	0.940±0.019	0.016±0.009	5.73±2.80	−5.93	0.031
Tipperah	20	19	0.995±0.018	0.021±0.012	7.17±3.51	−12.20*	0.028
East India							
East tribes[g]	34	29	0.989±0.010	0.023±0.012	7.69±3.67	−19.29*	0.013
Central India							
Andh	40	22	0.967±0.012	0.014±0.008	5.02±2.49	−9.13*	0.036
Pardhi	42	10	0.884±0.019	0.012±0.007	4.46±2.24	+0.88	0.091
Thoti	39	16	0.910±0.030	0.015±0.008	5.57±2.73	−2.29	0.025
Lambadi	86	57	0.985±0.005	0.017±0.009	5.91±2.85	−25.24*	0.018
South India							
South tribes-1[h]	22	6	0.671±0.077	0.007±0.004	2.54±1.42	+0.60	0.298
South tribes-2[i]	28	20	0.939±0.037	0.017±0.010	5.80±2.86	−9.19*	0.016
Koragas	53	7	0.746±0.039	0.008±0.005	3.00±1.59	+1.98	0.238
Kuruchian	46	11	0.800±0.034	0.013±0.007	4.76±2.37	+0.73	0.083
South tribes-3[j]	61	9	0.672±0.057	0.005±0.003	1.77±1.04	−1.31	0.073
South tribes-4[k]	46	19	0.937±0.019	0.013±0.007	4.44±2.23	−5.52	0.035
Yerava	53	12	0.775±0.049	0.007±0.004	2.64±1.43	−2.12	0.079
Kota	25	14	0.920±0.036	0.007±0.004	2.44±1.37	−8.12*	0.071
Nontribals							
Pushtoons	36	32	0.994±0.008	0.017±0.009	5.80±2.84	−25.32*	0.012
Northern Indians	40	33	0.990±0.008	0.015±0.008	5.27±2.60	−25.45*	0.012
Bangladeshis	29	25	0.988±0.013	0.017±0.009	6.11±2.99	−18.12*	0.010
Southern Indians	49	45	0.997±0.005	0.017±0.009	6.08±2.94	−25.26*	0.011

[a]Sample size. [b]Number of haplotypes. [c]±SD [d]Mean pairwise differences. [e]Raggedness index. [f]Includes Tharu and Buksa. [g]Includes Lodha, Munda and Santal. [h]Includes Jenukurumba and Kattunaiken. [i]Includes Soligas and Irula. [j]Includes Mullukurunan and Mullukurumba. [k]Includes Paniya, Kurumba and Bettakurumba. *Significant value after Bonferroni correction for multiple tests.

affinities to Indian tribes (0.048 with south and northeast tribes excluded; 0.084 if they are included) than to east Eurasians (0.068).

Gene tree

The outlier positions of south tribes in the MDS plot (Figure 3), coupled with demographic inferences indicating bottlenecks in these populations, mean that either south tribes have different HV1 sequences than other groups, or related sequences at very different frequencies. To distinguish between these two hypotheses, a neighbor-joining tree was constructed for 553 Indian and 420 worldwide HV1 sequences and rooted with a Neanderthal sequence (Figure 5). To facilitate the analysis, it was subdivided into 14 clusters (I–XIV). The deepest cluster (cluster I) was almost exclusively African-specific, with the exception of a single sequence from a south Indian tribe (Kuruchian). Sequences from the south tribes were found in all of the clusters, as were sequences from at least one other Indian population (with the exception of cluster I). This suggested that south tribes do not have different HV1 sequences than other groups, but rather related sequences at different frequencies.

The gene tree also suggested a close relationship between east Eurasians and northeast Indians in that 90% of the sequences in cluster XIII belonged either to east Eurasians or northeast Indians. In addition, 93% of the sequences from cluster XIV belonged either to Indian tribes (excluding northeast tribes) or castes, suggesting close relationships between these groups.

mtDNA haplogroup affiliation

mtDNA haplogroups are defined by RFLPs, but by using information from the gene tree, published data on Indian HV1 sequences for which the mtDNA haplogroup was known, and diagnostic HV1 mutations, we were able to infer the haplogroup affiliation for 90% of the Indian sequences (Table 3). Most Indian sequences belonged to

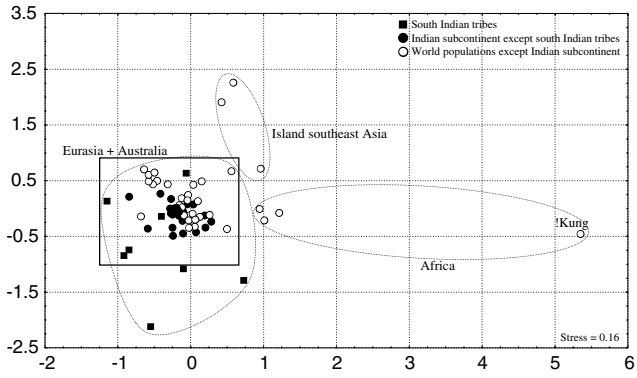


Figure 3 MDS plot of 66 world populations, based on Fst distances. The area enclosed by a solid line is magnified in Figure 4.

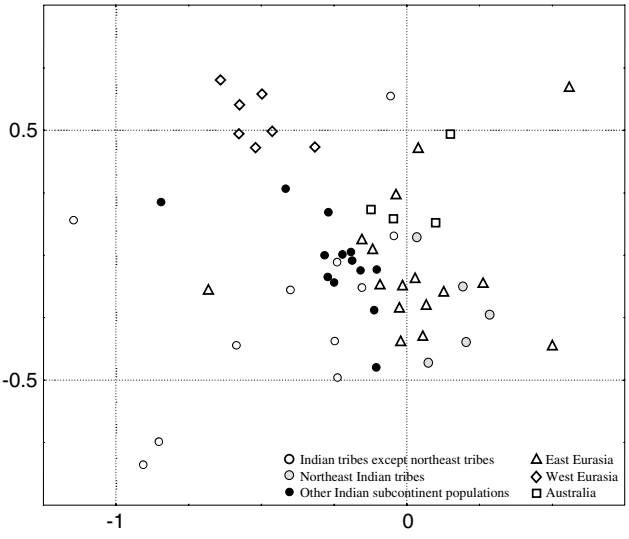


Figure 4 Magnified view of the MDS plot from Figure 3, focussing on Eurasian and Australian populations (area enclosed by solid line in Figure 3).

the Asian haplogroup M, as found previously.[9,12,19] South, central and east tribes exhibited very similar high frequencies of haplogroup M (~75%). North tribes showed a somewhat lower frequency (~67%), with a correspondingly higher frequency of the west Eurasian haplogroup JT (~7%). Haplogroup M frequency in northeast tribes was found to be lower (~56%) than in other Indian regions. The northeast tribes were also distinguished by a combined frequency of the east Asian haplogroups A and F of ~32%, while these two haplogroups were virtually absent elsewhere in the Indian subcontinent.

Genetic structure of Indian populations

AMOVA was used to investigate the genetic structure of Indian populations, focussing either on tribes only or on both tribes and castes (Table 4). In the total tribal sample (model 1), 88% of the variance was found within populations and 12% among populations. Indian tribes were then grouped according to geographic proximity (model 2), to linguistic affinities (model 3) and to the results suggested by the MDS analysis, namely two groups separating northeast tribes from all others (model 4). Under these models, 86–88% of the variance was found within populations, 10% among populations within groups and 2–4% among groups. A model that accurately reflects the genetic structure should maximize the variance among groups and minimize the variance among populations within groups; therefore, none of these models provides a good description of the genetic structure, although model 4 is the best.

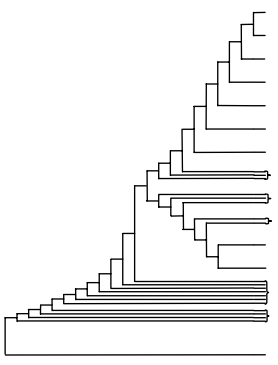
When the northeast, central and south groups of populations were analyzed separately, 22% of the variance was among populations in south tribes but only 3–5% in central and northeast tribes. These results emphasize the distinctiveness of south tribes from one another that was also evident in the MDS plots (Figures 3 and 4). Therefore, the analyses corresponding to models 1–4 were repeated with the south tribes excluded (not shown). Again, none of these models provided a good description of the genetic structure. The best model grouped populations on the basis of linguistic criteria, and was the only model for which the 'among groups' component was larger than the 'among populations within groups' component.

We also compared tribes to caste populations. East and northeast tribes were excluded from the analyses since: (i) no HV1 data from east and northeast castes are available, and (ii) these tribes speak Austro-Asiatic and Tibeto-Burman languages (respectively) which are spoken exclusively by tribal populations (preventing linguistically based analyses between tribes and castes). For the model based on social criteria (ie castes *vs* tribes; Table 4, model 5), 91% of the variance was within populations and 9% among populations within groups. The variance among groups did not differ significantly from zero ($P = 0.33$), suggesting that the social distinction of castes *vs* tribes does not accurately reflect the genetic structure of Indian populations. Removing south groups from the analysis did not change the picture (not shown), with the variance among groups still not differing significantly from zero ($P = 0.57$).

Discussion

Origins of tribal groups

Our analyses of mtDNA variation in tribal populations of India indicate that groups in different geographic regions have different demographic histories. In general, southern tribes have reduced mtDNA diversity and mismatch distributions strongly indicative of recent bottlenecks. The distinctiveness of southern groups is also emphasized by the MDS analyses and AMOVA. However, it is difficult to distinguish from these data between old and severe bottlenecks or more recent and less severe bottlenecks.



cluster	n	Africa	Australia	PNG	W Eurasia	E Eurasia	India			
							NE tribes	S tribes	other tribes	castes
XIV	59	1	0	0	0	2	1	8	18	29
XIII	30	0	0	0	0	9	18	3	0	0
XII	39	3	0	0	0	13	8	4	5	6
XI	79	9	8	6	3	17	11	5	8	12
X	134	12	10	1	0	37	14	7	20	33
IX	47	2	9	0	1	6	10	1	6	12
VIII	74	4	1	0	13	10	6	12	9	19
VII	80	0	4	0	0	7	8	12	13	36
VI	64	1	14	0	3	15	14	2	7	8
V	38	0	8	1	1	5	2	1	10	10
IV	68	0	8	0	6	10	3	10	9	22
III	111	0	9	6	23	15	9	5	11	33
II	104	20	11	0	5	26	4	1	12	25
I	46	45	0	0	0	0	0	1	0	0

Figure 5 Summary information of a gene tree of 973 mtDNA HV1 sequences, rooted with a Neanderthal sequence. The topology of the tree is represented on the left-hand side; *n* is the number of sequences in each cluster (I–XIV). The table also provides the number of sequences in each cluster sampled in the following geographic areas: Africa, Australia, Papua-New-Guinea (PNG), West Eurasia (W Eurasia), East Eurasia (E Eurasia) and India ('NE tribes': northeast tribes; 'S tribes': south tribes).

Present-day population sizes in southern tribes tend to be small (ie generally less than 30 000; Table 1), as compared to northern tribes (ie generally over 100 000; Table 1). Consequently, genetic drift could have generated large genetic distances both among southern groups and between southern and other groups, thereby masking their real affinities to other populations.[48] According to this scenario, southern populations have related mtDNA sequences (albeit at different frequencies) and hence a shared history with other Indian populations. An alternative hypothesis is that southern tribes have a specific mtDNA gene pool as compared to other Indian populations, indicating a long period of isolation and/or different history from other tribal groups. An example of the latter is PNG,[41,49] in which most sequences are found in two clusters (within clusters III and XI, Figure 5) characterized by long branches (not shown). However, southern Indian sequences are intermingled throughout the tree, clustering with sequences from multiple populations. In addition, south, central, and east Indian tribes all have similar mtDNA haplogroup compositions (Table 3). These results provide strong support for the hypothesis that southern tribes have mtDNA sequences closely related to those of other tribes, but with different frequencies, which would suggest fairly recent bottlenecks occurring in these populations.

A possible cause of these bottlenecks, put forth by Excoffier and Schneider,[48] involves Neolithic human expansions. According to this hypothesis, the recent settlement of Indo-European speakers in India some 3500 years ago[3] might have had a major impact on the demography of south tribes. West Eurasian mtDNA haplogroups H, JT and W represent 6–7% of north and central tribes (Table 3), which are located in the area where Indo-European languages are spoken. In contrast, these west Eurasian mtDNA types are virtually absent in south tribes, which are located where Dravidian languages are spoken. This might reflect different responses of local people to the Indo-European settlement of India. In the north and center, Indo-Europeans may have admixed with local people,[50] concomitant with the spread of Indo-European languages. In contrast, in the southern part of India, local populations may have challenged the arrival of Indo-European newcomers, resulting in limited admixture, reduction of tribal population sizes and retention of their original languages, thus explaining why Dravidian languages survived the spread of Indo-European languages in south India.

Tibeto-Burman speakers from northeast India show closer genetic affinities with east Asian groups than with other Indian groups. This is suggested by the MDS analyses and mtDNA haplogroup composition, in that northeast Indian tribes possess haplogroups A and F, which are frequent in east Asians but virtually absent from other regions of India. The mtDNA evidence (Clark *et al*[18], this study) thus agrees with Y chromosome evidence[51] as well as linguistic evidence,[52] indicating a probable east Asian origin of these particular tribes. Archeological, linguistic and genetic evidence[51] suggests that proto-Tibeto-Burman languages arose 5000–6000 years ago in east Asia. The fact that Tibeto-Burman speaking tribes from India have retained genetic traces of east Asian origins for such a long time suggests that, despite the more recent migrations to India, these populations remained relatively isolated, explaining the close correlation between genetic and linguistic results. This contrasts with the situation observed in other regions in the world, for example in Scandinavia[53] and the Caucasus,[34] where migrations led to language replacements and hence to incongruencies between genetic and linguistic results.

Table 3 Frequencies of some major mtDNA haplogroups in India inferred from HV1 sequences

	Indian castes			*Indian tribes*			
Haplogroups	*North*[a]	*Central*[b]	*North*[a]	*Central*	*South*	*East*[c]	*Northeast*
M (all)	54.8	65.7	66.7	73.3	74.6	76.3	55.7
M–C	—	—	3.3	—	2.9	—	7.5
R (all)	32.3	NA[d]	33.3	23.6	25.1	NA[d]	27.6
R–F	8.1	—	—	—	—	—[e]	15.5
R–JT	4.8	2.1	6.7	4.2	—	—[e]	—
R–H	—	1.2	—	0.5	0.6	—	—
R–U	4.8	10.3	10.0	6.3	14.9	13.6	1.7
Others	12.9	NA[d]	—	3.1	0.3	NA[d]	16.7
A	—	—	—	—	—	—	16.7
L	—	—	—	—	0.3	—	—
W	12.9	0.4	—	1.6	—	—[e]	—

[a]Deduced from Kivisild *et al*,[12] [b]Bamshad *et al*,[9] [c]Roychoudhury *et al*,[19] [d]Not available, [e]Deduced from sequence data of Roychoudhury *et al*.[19]

Table 4 Results of AMOVA

	Among groups		*Among populations within groups*		*Within populations*	
Model	*Var*[a]	*P*	*Var*[a]	*P*	*Var*[a]	*P*
(1) Total tribal sample	—	—	11.83	<0.001	88.17	—
(2) Geographic criteria	1.73	0.085	10.50	<0.001	87.77	<0.001
(3) Linguistic criteria	2.53	0.015	10.22	<0.001	87.25	<0.001
(4) Northeast vs. all others	3.96	<0.001	9.94	<0.001	86.10	<0.001
(5) Social criteria	0.09	0.326	9.37	<0.001	90.54	<0.001

[a]Variance (%).

Apart from northeast tribes, all other Indian tribes exhibited a similar and high frequency of mtDNA haplogroup M (ie 56% in northeast *vs* ~75% for others). These results are strikingly similar to a previous study based on RFLP analysis of a smaller set of Indian tribes,[19] according to which haplogroup M had a frequency of ~51% in northeast tribes and ~76% in east, central and south tribes. This homogeneity in frequency of haplogroup M further supports the view that south tribes have sequences closely related to those of their neighboring populations. Thus, our mtDNA data are compatible with at least three major sources for the present-day mtDNA gene pool of Indian tribes: (i) a major one associated with all non-northeast tribes (whatever their linguistic or geographic ties), with a high frequency of mtDNA haplogroup M; (ii) one associated with Tibeto-Burman speakers from northeast India, with affinities to east Asians; and (iii) a third one associated with the presence of west Eurasian-typical mtDNA haplogroups (ie haplogroups H, JT and W, which represent 6–7% of mtDNA types in north and central tribes), most probably attributable to admixture with recent Indo-European-speaking migrants to India.[50]

Relationships between castes and tribes

The comparison between Indian castes and tribes revealed no strong difference between them, as pointed out by the AMOVA. A possible explanation for the observed similarities in caste and tribal mtDNA gene pools is common ancestry, with a proto-Asian origin of Indian castes.[9] An alternative hypothesis involves a proto-west-Eurasian origin of castes, with the present-day similarities in caste and tribal mtDNA gene pools then being attributable to recent admixture with local Indian populations. The latter hypothesis would require extensive gene flow, which could seem *a priori* to be incompatible with the mating practices imposed by the caste system in India.[6] However, there is evidence for female gene flow between Hindu castes[10] and it has been suggested that male gene flow in south Indian populations may not be as negligible[54] as previously thought.[11] Our results show the presence of west-Eurasian typical mtDNA haplogroups in Indian tribes, presumably resulting from admixture with Indo-Europeans (ie who probably introduced the caste system in India). This interpretation would suggest that caste people initially possessed west-Eurasian mtDNAs rather than Asian mtDNAs. This view is reinforced by the fact that caste groups are more similar to west Eurasians (average Fst: 0.080) than are the tribals (average Fst: 0.149; and 0.117 if south tribes are excluded). Therefore, the similarities in caste and tribal mtDNA gene pools might reflect extensive maternal gene flow rather than common ancestry. However, caste and tribal populations are separated by an

average Fst distance of 0.049 (if south tribes are excluded; 0.082 if they are included), suggesting that overall, castes are closer to Indian tribes than to west Eurasians. This makes the hypothesis of proto-Asian ancestry of castes equally plausible. The mtDNA data alone do not support one hypothesis over the other. Moreover, caste populations from different regions of India may have different origins,[55] some of them being derived from west Eurasian ancestors with subsequent admixture with local populations, others being derived from local population ancestors via acculturation.

Relationships with other populations

mtDNA variation in India suggests that overall, Indian tribes show more affinities to east Eurasians than to west Eurasians. This means that migrations from the west (ie involving Indo-Europeans and Neolithic expansions of farmers) have not had a major distorting impact on the original gene pool. This view is consistent with the relatively small proportion of west Eurasian typical mtDNA haplogroups present in Indian tribes. On the other hand, three typical east-Asian mtDNA haplogroups (A, B and F) are absent or virtually absent from non-northeast India (Bamshad *et al*[9] Kivisild *et al*,[12] Roychoudhury *et al*,[19] this study). Furthermore, the fourth typical east Asian mtDNA haplogroup M has a different structure in India as compared to other Asian areas.[9] This suggests that, although they show close affinities, the east Asian and Indian mtDNA gene pools are fairly distinct. This result is consistent with the suggestion that the east Asian and Indian mtDNA pools have been separated from each other for about 30 000 years.[49]

It has been hypothesized that the peopling of Sahul (PNG and Australia) may have been the result of an early migration from east Africa through the Indian subcontinent following the 'southern route'.[1,3,56] Australian populations exhibited an average Fst distance of 0.067 with east Eurasians and of 0.089 with Indians (but only 0.062 if South tribes excluded), whereas the average Fst values separating Australians from PNG or African (!Kung excluded) populations were 0.194 and 0.145, respectively. These results suggested close genetic affinities between Australian populations and both Indian and east Eurasian populations. An India–Australia connection is consistent with other mtDNA[41] and Y chromosome[57] evidence. Taken together with other conclusions,[41,57] the present results give credence to the trihybrid model of peopling of Australia[58] involving 'Negrito', east Asian and Indian sources. The Indian influence on Australia may be recent (ie <5000 years),[41,57] thus much later than (and therefore independent from) the early migration that would have followed the southern route $\sim 60\,000$ years ago.

In addition, Forster *et al*[49] have proposed an mtDNA control region motif (16223C and 16357C) which could represent a signature of an early migration from Africa to Sahul through the southern route. This motif was not found in any Indian tribal mtDNA; 16357C had a frequency of only 2.1% and was always associated with 16223T, while 16223C had a frequency of 27.9%. Furthermore, Indians do not show particular affinities to Africans. A possible exception is the typical African HV1 sequence found in a Kuruchian from south India. However, there are communities in India such as the Siddis, who are known to be recent migrants from Africa.[6] The African-like sequence found in India could therefore originate from admixture between recent African migrants and Indian tribals, or it may represent a remnant of an ancient migration from Africa to India; it is difficult to draw conclusions from a single sequence.

In summary, although the data support a recent India–Australia connection, we could not find in Indian tribals any unquestionable genetic signature of the $\sim 60\,000$ year-old migration from Africa to Sahul following the postulated southern route. A possible explanation would be that such migration never occurred along that route. Alternatively, the early migrants from Africa may have made their way to Sahul following the southern route without settling in India. Another possibility, which is probably the most reasonable one, is that in India the genetic traces of early migrations along the southern route were erased by the subsequent migrations which shaped the present-day mtDNA gene pool of India.[3]

Acknowledgements

We are grateful to the original donors of samples. We thank Birgit Nickel and Carsten Schwarz for technical assistance. We also thank Partha Majumder, Michael Bamshad and Scott Watkins for kindly providing HV1 sequences, and Silke Brauer, Manfred Kayser, Vano Nasidze, Hiroki Oota, Brigitte Pakendorf, Anthony Ryan and Alice Salzat for fruitful discussions during the course of this study. This project was supported by funds from the Max Planck Society, Germany.

References

1 Lahr MM, Foley R: Multiple dispersals and modern human origins. *Evol Anthropol* 1994; **3**: 48–60.
2 Templeton AR: Out of Africa again and again. *Nature* 2002; **416**: 45–51.
3 Cavalli-Sforza LL, Menozzi P, Piazza A: *History and geography of human genes*. Princeton, Princeton University Press, 1994.
4 Cann RL: Genetic clues to dispersal of human populations: retracing the past from the present. *Science* 2001; **291**: 1742–1748.
5 Papiha SS: Genetic variation in India. *Hum Biol* 1996; **68**: 607–628.
6 Majumder PP: People of India: biological diversity and affinities. *Evol Anthropol* 1998; **6**: 100–113.
7 Mountain JL, Hebert JM, Bhattacharyya S *et al*: Demographic history of India and mtDNA-sequence diversity. *Am J Hum Genet* 1995; **56**: 979–992.
8 Bamshad M, Fraley AE, Crawford MH *et al*: MtDNA variation in caste populations of Andhra Pradesh. India. *Hum Biol* 1996; **68**: 1–28.

9 Bamshad M, Kivisild T, Watkins WS *et al*: Genetic evidence on the origins of Indian caste populations. *Genome Res* 2001; **11**: 994–1004.
10 Bamshad MJ, Watkins WS, Dixon ME *et al*: Female gene flow stratifies Hindu castes. *Nature* 1998; **395**: 651–652.
11 Bhattacharyya NP, Basu P, Das M *et al*: Negligible male gene flow across ethnic boundaries in India, revealed by analysis of Y-chromosomal DNA polymorphisms. *Genome Res* 1999; **9**: 711–719.
12 Kivisild T, Bamshad MJ, Kaldma K *et al*: Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. *Curr Biol* 1999; **9**: 1331–1334.
13 Majumder PP, Roy B, Banerjee S *et al*: Human-specific insertion/deletion polymorphisms in Indian populations and their possible evolutionary implications. *Eur J Hum Genet* 1999; **7**: 435–446.
14 Singh KS: *People of India: the scheduled tribes*. Delhi: Oxford University Press, 1994, (vol III.)
15 Merriwether DA, Friedlaender JS, Mediavilla J, Mgone C, Gentz F, Ferrel RE: Mitochondrial DNA variation is an indicator of Austronesian influence in Island Melanesia. *Am J Phys Anthropol* 1999; **110**: 243–270.
16 Soodyall H, Vigilant L, Hill AV, Stoneking M, Jenkins T: MtDNA control-region sequence variation suggests multiple independent origins of an 'Asian-specific' 9-bp deletion in Sub-Saharan Africans. *Am J Hum Genet* 1996; **58**: 595–608.
17 Watkins WS, Bamshad M, Dixon ME *et al*: Multiple origins of the mtDNA 9-bp deletion in populations of South India. *Am J Phys Anthropol* 1999; **109**: 147–158.
18 Clark VJ, Sivendren S, Saha N *et al*: The 9-bp deletion between the mitochondrial lysine tRNA and COII genes in tribal populations of India. *Hum Biol* 2000; **72**: 273–285.
19 Roychoudhury S, Roy S, Basu A *et al*: Genomic structures and population histories of linguistically distinct tribal groups of India. *Hum Genet* 2001; **109**: 339–350.
20 Melton T, Peterson R, Redd AJ *et al*: Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis. *Am J Hum Genet* 1995; **57**: 403–414.
21 Vigilant L, Pennington R, Harpending H, Kocher TD, Wilson AC: Mitochondrial DNA sequences in single hairs from a southern African population. *Proc Natl Acad Sci USA* 1989; **86**: 9350–9354.
22 Handt O, Krings M, Ward RL, Pääbo S: The retrieval of ancient human DNA sequences. *Am J Hum Genet* 1996; **59**: 368–376.
23 Burckhardt F, von Haeseler A, Meyer S: HvrBase: compilation of mtDNA control region sequences from primates. *Nucleic Acids Res* 1999; **27**: 138–142.
24 Anderson S, Bankier AT, Barrell BG *et al*: Sequence and organization of the human mitochondrial genome. *Nature* 1981; **290**: 457–465.
25 Schneider S, Roessli D, Excoffier L: *Arlequin ver. 2.000: a software for population genetics data analysis*. University of Geneva, Switzerland: Genetics and Biochemistry Laboratory, 2000.
26 Fu YX: Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 1997; **147**: 915–925.
27 Harpending HC, Sherry ST, Rogers AR, Stoneking M: The genetic structure of ancient human populations. *Curr Anthropol* 1993; **34**: 483–496.
28 Excoffier L, Smouse P, Quattro J: Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 1992; **131**: 479–491.
29 Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC: African populations and the evolution of human mitochondrial DNA. *Science* 1991; **253**: 1503–1507.
30 Krings M, Salem AH, Bauer K *et al*: MtDNA analysis of Nile River Valley populations: a genetic corridor or a barrier to migration? *Am J Hum Genet* 1999; **64**: 1166–1176.
31 Watson E, Bauer K, Aman R, Weiss G, von Haeseler A, Pääbo S: MtDNA sequence diversity in Africa. *Am J Hum Genet* 1996; **59**: 437–444.
32 Hofmann S, Jaksch M, Bezold R *et al*: Population genetics and disease susceptibility: characterization of central European haplogroups by mtDNA gene mutations, correlation with D-loop variants and association with disease. *Hum Mol Genet* 1997; **6**: 1835–1846.
33 Calafell F, Underhill P, Tolun A, Angelicheva D, Kalaydjieva L: From Asia to Europe: mitochondrial DNA sequence variability in Bulgarians and Turks. *Ann Hum Genet* 1996; **60**: 35–49.
34 Nasidze I, Stoneking M: Mitochondrial DNA variation and language replacements in the Caucasus. *Proc R Soc London B* 2001; **268**: 1197–1206.
35 Comas D, Calafell F, Mateu E *et al*: Trading genes along the Silk Road: mtDNA sequences and the origin of Central Asian populations. *Am J Hum Genet* 1998; **63**: 1824–1838.
36 Oota H, Kitano T, Jin F *et al*: Extreme mtDNA homogeneity in continental Asian populations. *Am J Phys Anthropol* 2002; **118**: 146–153.
37 Seo Y, Stradmann-Bellinghausen B, Rittner C, Takahama K, Schneider PM: Sequence polymorphism of mitochondrial DNA control region in Japanese. *Forensic Sci Int* 1998; **97**: 155–164.
38 Lee SD, Shin CH, Kim KB, Lee YS, Lee JB: Sequence variation of mitochondrial DNA control region in Koreans. *Forensic Sci Int* 1997; **87**: 99–116.
39 Kolman CJ, Sambuughin N, Bermingham E: Mitochondrial DNA analysis of mongolian populations and implications for the origin of New World founders. *Genetics* 1996; **142**: 1321–1334.
40 Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M: Human mtDNA and Y-chromosome variation is correlated with matrilocal *vs*. patrilocal residence. *Nat Genet* 2001; **29**: 20–21.
41 Redd AJ, Stoneking M: Peopling of Sahul: mtDNA variation in Aboriginal Australian and Papua New Guinean populations. *Am J Hum Genet* 1999; **65**: 808–828.
42 van Holst Pellekaan SM, Frommer M, Sved JA, Boettcher B: Mitochondrial control-region sequence variation in Aboriginal Australians. *Am J Hum Genet* 1998; **62**: 435–449.
43 Redd AJ, Takezaki N, Sherry S, McGarvey S, Sofro ASM, Stoneking M: Evolutionary history of the COII/tRNALys intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. *Mol Biol Evol* 1995; **12**: 604–615.
44 Kumar S, Tamura K, Jakobsen IB, Nei M: *MEGA2: Molecular evolutionary genetics analysis software*. Tempe, AZ, USA: Arizona State University, 2001.
45 Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Pääbo S: Neandertal DNA sequences and the origin of modern humans. *Cell* 1997; **90**: 19–30.
46 Macaulay V, Richards M, Hickey E *et al*: The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am J Hum Genet* 1999; **64**: 232–249.
47 Singh KS: *People of India: India's communities*. Delhi: Oxford University Press, 1998, Vol IV–VI.
48 Excoffier L, Schneider S: Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions? *Proc Natl Acad Sci USA* 1999; **96**: 10597–10602.
49 Forster P, Torroni A, Renfrew C, Röhl A: Phylogenetic star contraction applied to Asian and Papuan mtDNA evolution. *Mol Biol Evol* 2001; **18**: 1864–1881.
50 Passarino G, Semino O, Bernini LF, Santachiara-Benerecetti AS: Pre-Caucasoid and Caucasoid genetic features of the Indian population, revealed by mtDNA polymorphisms. *Am J Hum Genet* 1996; **59**: 927–934.
51 Su B, Xiao C, Deka R *et al*: Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. *Hum Genet* 2000; **107**: 582–590.
52 Matisoff JA: Sino-Tibetan linguistics: present state and future prospects. *Annu Rev Anthropol* 1991; **20**: 469–504.
53 Sajantila A, Pääbo S: Language replacement in Scandinavia. *Nat Genet* 1995; **11**: 359–360.
54 Ramana GV, Su B, Jin L *et al*: Y-chromosome SNP haplotypes suggest evidence of gene flow among caste, tribe, and the

migrant Siddi populations of Andhra Pradesh, South India. *Eur J Hum Genet* 2001; **9**: 695–700.
55 Majumder PP: Indian caste origins: genomic insights and future outlook. *Genome Res* 2001; **11**: 931–932.
56 Quintana-Murci L, Semino O, Bandelt HJ, Passarino G, McElreavey K, Santachiara-Benerecetti AS: Genetic evidence on an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. *Nat Genet* 1999; **23**: 437–441.
57 Redd AJ, Roberts-Thomson J, Karafet T *et al*: Gene flow from the Indian subcontinent to Australia: evidence from the Y chromosome. *Curr Biol* 2002; **12**: 673–677.
58 Birdsell JB: Preliminary data on the tri-hybrid origin of the Australian Aborigines. *Archeol Phys Anthropol Oceania* 1967; **2**: 100–155.