# COMMON MATERNAL LEGACY OF INDIAN TRIBAL AND CASTE POPULATIONS

Metspalu M.[1], Kivisild T.[1] , Papiha S.S.[2], Villems R.[1]

[1]Department of Evolutional Biology, Institute of Molecular and Cell Biology, Tartu University and Estonian Biocentre. Riia 23, Tartu 51010 Estonia.
[2]Department of Human Genetics, University of Newcastle-upon-Tyne, UK

## Introduction

The origins of Indian tribals, who presently constitute ~7% of the total population of India, have been subject to numerous studies in different fields of science. The resulting hypotheses range from referring to some tribals as the descendants of the original Palaeolithic inhabitants of India to conclusions that yet some others are recent immigrants. Given the immense diversity and number of tribal communities, there seems to be no single answer.

MtDNA haplogroup (Hg) M appears at the highest frequency among both tribal and caste populations of India. Hg M is also the major component of the mtDNA genepool to the east and to the north of India while a sharp cline exists to the west: in Iran, Hg M frequency is a mere 5%. Phylogeny of haplogroup M in Indian populations differs profoundly from that observed in east and central Asian populations, where Hg M sub-haplogroups D, E, G, C, Z constitute the bulk of Hg M lineages. The coalescence times of both, the eastern Asian and the Indian haplogroup M have been estimated to be over 50 000 BP (Wallace 1995, Chen et al. 1995, Mountain et al. 1995, Kivisild et al. 1999b). Note that the term coalescence time refers to the time since the start of expansion of a lineage not to the age of a lineage. The given coalescence times suggest that the two macro-populations started to expand separately but simultaneously and since then, there has been only very limited gene flow between India and eastern Asia. The lack of any signs for extensive re-migrations of eastern Asians to India is further supported by the scarcity of mtDNA lineages belonging to haplogroups A, B and F in India (see Fig 2a for the spread of East-Asian specific mtDNA lineages).

Geographically, the distribution of Hg U is a mirror image of that for haplogroup M: U is not present in eastern Asia, but is frequent in European populations and among Indians. This reverse analogy goes further: Indian U lineages differ substantially from those observed in Europe and their coalescence to a common ancestor, like that for the haplogroup M lineages, dates back to about 50,000 years (Kivisild et al. 1999a) (see Fig.2b for the spread of West-Eurasian specific mtDNA lineages)

In sum, the (characterised so far) maternal lineages present in India are largely Indian specific and show expansion time in the Palaeolithic.

To push the understanding on the origins of Indian tribals further, we have studied maternal lineages of 4 tribal and 2 caste populations by analysing mtDNA HVS I and II sequence variation accompanied by RFLP typing of characteristic coding area sites in these populations. In the analyses we rely on these and published data. See Table 1 for details on the studied populations and data included into the analysis.

## Table 1. Observed mtDNA haplogroup frequences

| | | Bhoksa n=23 | | Tharu n=39 | | Kanet n=37 | | Lodha n=56 | | Kurmi n=55 | | Bengali n=51 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Hg | | n | % | n | % | n | % | n | % | n | % | n | % |
| A | | | | | | 1 | 3(1-14) | | | | | | |
| B | | | | | | 2 | 5(2-18) | | | | | | |
| F | | | | | | 4 | 11(4-25) | | | | | | |
| H | | | | | | 1 | 3(1-14) | | | | | 3 | 6(2-16) |
| I | | | | | | 1 | 3(1-14) | | | | | | |
| M | | 17 | 74(53-87) | 22 | 56(41-71) | 19 | 51(36-67) | 56 | 100(95-100) | 46 | 84(72-91) | 30 | 59(45-71) |
| | M2 | 1 | 4(1-21) | | | | | | | 4 | 7(3-17) | | |
| | M3a | 2 | 9(3-27) | 1 | 3(1-13) | 1 | 3(1-14) | | | | | 2 | 4(1-13) |
| | M4a | | | 2 | 5(2-17) | | | | | | | | |
| | M6 | | | | | | | | | 4 | 7(3-17) | 2 | 4(1-13) |
| | M18 | 1 | 4(1-21) | | | 3 | 8(3-21) | 22 | 39(28-52) | | | | |
| | M25 | | | 1 | 3(1-13) | 3 | 8(3-21) | | | | | 1 | 2(1-10) |
| | MC | | | 1 | 3(1-13) | 1 | 3(1-14) | | | | | | |
| | MD | | | 3 | 8(3-20) | 2 | 5(2-18) | | | | | | |
| | ME | | | | | 1 | 3(1-14) | | | | | | |
| | MZ | | | | | 1 | 3(1-14) | | | | | | |
| | M* | 13 | 57(37-74) | 14 | 36(23-52) | 7 | 19(1-34) | 34 | 61(48-72) | 38 | 69(56-80) | 25 | 49(36-62) |
| Q | | | | 2 | 5(2-17) | | | | | 2 | 4(1-12) | 5 | 10(4-21) |
| R | | 3 | 13(5-32) | 4 | 10(4-24) | 2 | 5(2-18) | | | 1 | 2(0-1) | 4 | 8(3-19) |
| T | | | | 2 | 5(2-17) | | | | | | | 2 | 4(1-13) |
| U | | 3 | 13 | 7 | 18(9-33) | 5 | 14(6-28) | | | 6 | 11(5-22) | 5 | 10(4-21) |
| | U2 | 2 | 9(3-27) | 5 | 13(6-27) | 1 | 3(1-14) | | | 5 | 9(4-20) | 4 | 8(3-19) |
| | U4 | | | 1 | 3(1-13) | | | | | | | 1 | 2(1-10) |
| | U5 | | | | | 1 | 3(1-14) | | | | | | |
| | U7 | 1 | 4(1-21) | | | 3 | 8(3-21) | | | 1 | 2(0-1) | | |
| W | | | | 1 | 3(1-13) | 2 | 5(2-18) | | | | | | |
| X | | | | | | | | | | | | 1 | 2(1-10) |

Table 1 presents the mtDNA haplogroup frequencies among the studied Indian populations. The 95% credible region for proportion is calculated as in Berger 1985.

Note that only the Kanet from Himachal Pradesh and to a lesser extent the Tharu from Uttar Pradesh harbor East-Asian specific (EA) mtDNA haplogroups A, B, F, MC, MD, ME and MZ. Social and trade connections between the Kanet and Tibetans have been well documented. As has the genetic admixture between these populations been shown before with extensive “classical genetic markers” frequency studies (Papiha et al. 1996).

In general, the West-Eurasian specific (WE) H, I U4, U5, U7, W and X are scattered more evenly. Still, among the Bhoksas from Himachal Pradesh and the Kurmis from West Bengal the occurrence of WE haplogroups is restricted to U7 only, while the Austro-Asiatic speaking Lodha from West-Bengal lack any WE or EA haplogroups whatsoever. From the relatively low level of diversity, 100% of our sample belongs to mtDNA Hg M (see also Fig 1a), it is evident that the Lodha have gone through

(A) bottleneck(s) and/or founder-effect(s) in path of their demographic history.

## Table 2. Populations

| State | population | n | social status | Language group | |
|---|---|---|---|---|---|
| Andra Pradesh | - | 246 | mixed cast | Dravidic | Bamshad et al. 1998 |
| Andra Pradesh | - | 9 | mixed cast | Dravidic | Quintana-Murci et al. 1999 |
| Andra Pradesh | Chenchu | 96 | tribal | Dravidic | Kivisild et al. (in prep) |
| Andra Pradesh | Koya | 81 | tribal | Dravidic | Kivisild et al. (in prep) |
| Andra Pradesh | Lambadi | 86 | tribal | Indo-European | Kivisild et al. 1999 |
| Delhi* | - | 10 | | Indo-European | Quintana-Murci et al. 1999 |
| Gujarat | Cobra Brahmin | 58 | cast | Indo-European | this laboratory |
| Gujarat | - | 55 | mixed cast | Indo-European | this laboratory |
| Himachal | Bhoksa | 23 | tribal | Indo-European | this study |
| Himachal | Kanet | 37 | tribal | Tibeto-Burman | this study |
| Iran | - | 437 | - | Indo-European | this laboratory |
| Karnataka | Havik | 47 | cast | Dravidic | Mountain et al. 1995 |
| Karnataka | Mukri | 42 | cast | Dravidic | Mountain et al. 1995 |
| Kashmir | - | 19 | mixed cast | Indo-European | Kivisild et al. 1999 |
| Kerala | Kadar | 7 | tribal | Dravidic | Mountain et al. 1995 |
| Madhya Pradesh | Muria | 12 | tribal | Dravidic | Roychoudhury et al. 2001 |
| Maharashtra | Parsi | 55 | cast | Indo-European | this laboratory |
| Pakistan | - | 8 | - | Indo-European | Kivisild et al. 1999 |
| Punjab | Lobana | 62 | tribal | Indo-European | Kivisild et al. 1999 |
| Punjab | - | 8 | mixed cast | Indo-European | Quintana-Murci et al. 1999 |
| Punjab | Punjab | 113 | mixed cast | Indo-European | this laboratory |
| Rajastan | Rajput | 36 | cast | Indo-European | this laboratory |
| Sri Lanka | Moor | 50 | cast | Dravidic | this laboratory |
| Sri Lanka | Sinhalese | 81 | cast | Indo-European | this laboratory |
| Tamil Nadu | Irula | 14 | tribal | Dravidic | Roychoudhury et al. 2001 |
| Tamil Nadu | Kota | 25 | tribal | Dravidic | Roychoudhury et al. 2001 |
| Tamil Nadu | Kurumba | 10 | tribal | Dravidic | Roychoudhury et al. 2001 |
| Tripura | Tipperah | 20 | tribal | Tibeto-Burman | Roychoudhury et al. 2001 |
| Uttar Pradesh | Tharu | 38 | tribal | Indo-European | Kivisild et al. 1999 |
| Uttar Pradesh | - | 111 | mixed cast | Indo-European | Kivisild et al. 1999 |
| West bengal | - | 51 | mixed cast | Indo-European | this study |
| West Bengal | Kurmi | 55 | scheduled cast | Indo-European | this study |
| West Bengal | Lodha | 56 | tribal | Austro-Asiatic | this study |
| West Bengal | Lodha | 14 | tribal | Austro-Asiatic | Roychoudhury et al. 2001 |
| West Bengal | Munda | 6 | tribal | Austro-Asiatic | Roychoudhury et al. 2001 |
| West Bengal | Santal | 14 | tribal | Austro-Asiatic | Roychoudhury et al. 2001 |
| | | 2092 | | | |

* Indians from Delhi and Nepal

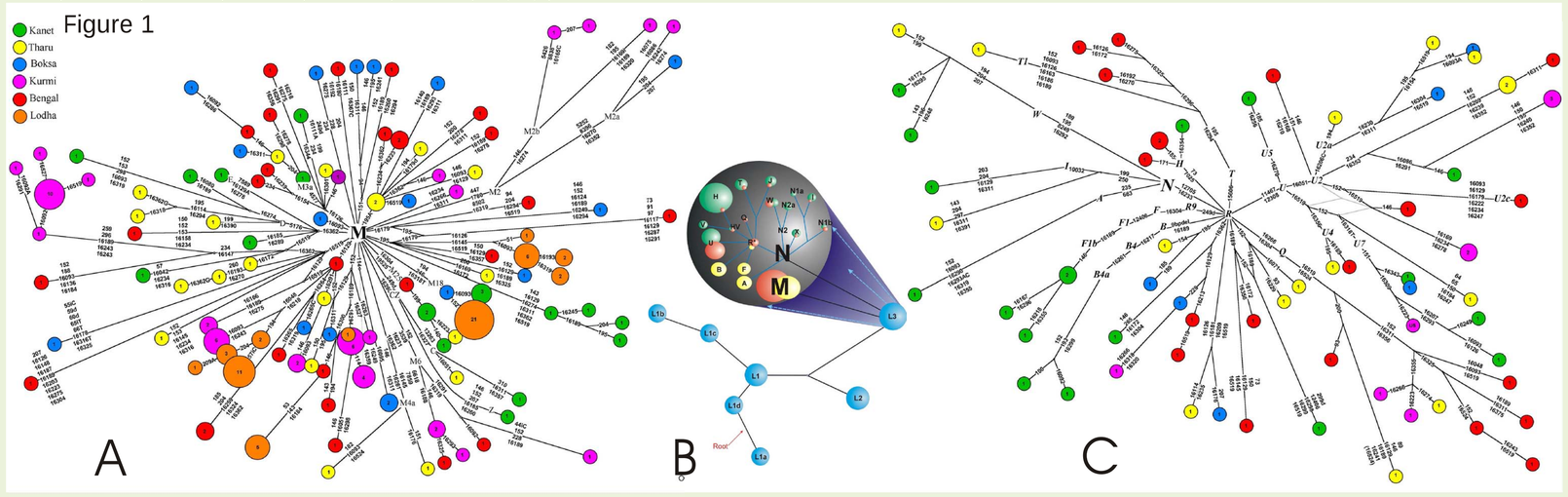



Figure 1

Panel B (adapted from Kivisild et al. 1999) depicts the general backbone of the global human mtDNA tree. Colours of spheres indicate population groups as follows: blue - Africans; yellow - east Asians and native Americans; red - Indians and green - western Eurasians. The diameter of the sphere depicts the relative frequency of the haplogroup. Note that all non-African lineages arise from two lineages M and N which in turn branch from a single African mtDNA cluster L3.

Panels A and C present reconstructions of the mtDNA macro-lineages M and N among the studied 6 populations. Reconstruction is based on HVS1 and HVS2 sequence variation data accompanied by data on some characteristic coding area polymorphisms.

Panel A: HVS1 and HVS2 sequence variation based greedy network (reduced median algorithm followed by median joining algorithm as in Bandelt et al.2000; r=2, e=0) of the macro haplogroup M lineages spotted among the studied 6 populations. The network was calculated using the Network 3110 software by Fluxus-Engineering (www.fluxus-engineering.com). Star contraction (Forster et al. 2001) resulting in 64 nodes, was implemented prior to network calculation. HVS1 sites were weighted into 4 classes (adapted from Hasegawa et al. 1993) in addition HVS2 sites 146 and 152 were given low weights (as in Helgason et al. 2000). As a second stage all the observed Hg M subclades (defined by coding area mutations) e.g. East Asian specific MZ, MC etc. and Indian specific M2, M3a etc. were manually added to the network. Node areas correspond to haplotype frequencies, colour denotes population (see legend), numbers inside circles indicate number of individuals and numbers on lines connecting haplotypes denote substitution positions. Deletions are shown as position followed by “d” and transversions as position followed by a letter indicating the resulting nucleotide.

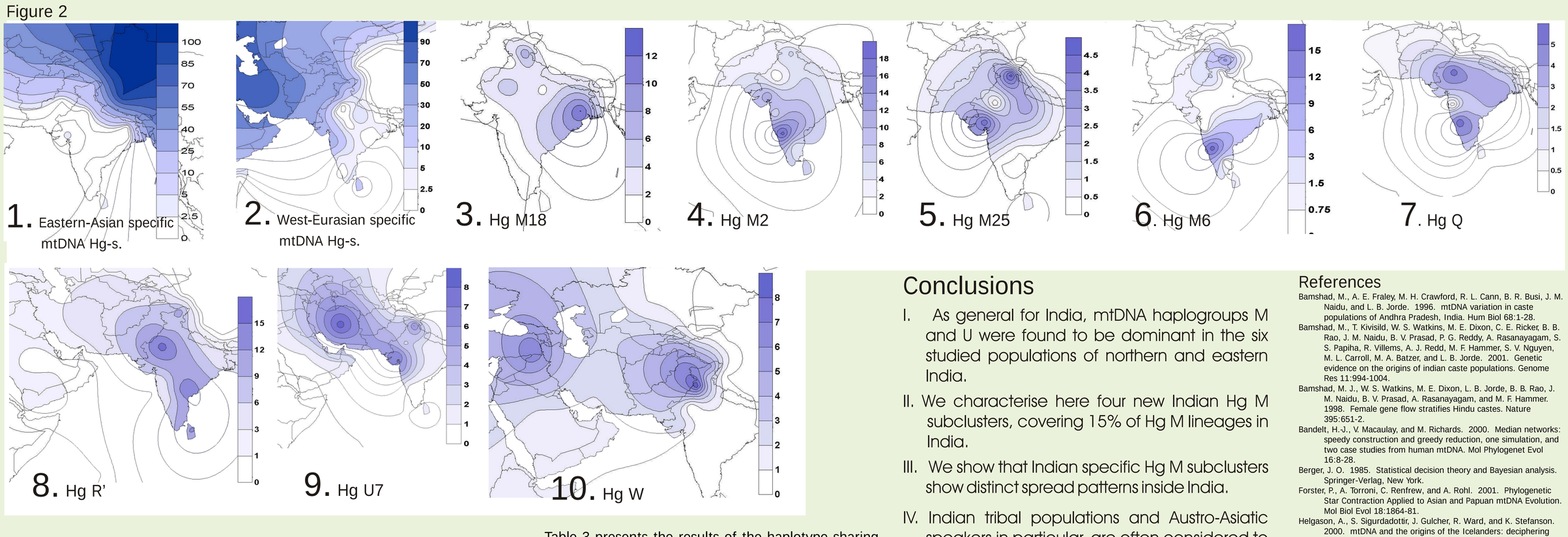
We characterise here four new Indian M subclusters covering 21% of Hg M lineages in studied populations. M4a, M6, M18, M25 are in addition to gain of AluI restriction site at np10400 defined by gain of HaeIII site at np6618 and loss of MboI site at np 7859, loss of AluI site at np3539, an A-T transversions at np 16318 and loss of MspI site at np15925, respectively.

Note the relatively diminished diversity of the Lodha, the analysed 56 samples produced only 12 haplotypes, all belonging to Hg M. We note, however, that other studies have shown greater variability among the Lodha (Roychoudhury et al. 2001). The somewhat small global sample size (n=70) for the Lodha seems to be still too small to draw decisive conclusions.

Panel C: Reconstruction of mtDNA lineages arising from the macro lineage N. Construction methods described for panel A apply also here.

## Figure 2

We have used Surfer 7.0 package (Goldensoftware) to present geographical variation of haplogroup frequencies. Grid files were generated using the Kriging method with default settings. The East-Asian specific Hg-s on panel 1 are A, B, F, MD, ME, MG and MZ. Hg-s H, HV, I, J, K, T, U1, U3, U4, U5, U7, W and X constitute the West-Eurasian specific group depicted on panel 2. Panels 3-7 illustrate the spreads of Hg M subclades, which are restricted to India only. The somewhat north-western and north-eastern distribution patterns of Hg-s M18 and M25, respectively, are illustrated on panels 3 and 5. Hg M2, on the contrary seems to be more concentrated in the south (panel 4). The bipolar arrangement of Hg M6 (panel 6) is at this point hard to interpret. The spread of Hg Q (panel 7) seems to be quite uniform except for the far north and the Parsi in Maharashtra. R* lineages are generally present all over India. Quite extensive overlap between north-western India and Iran is provided by Hg U7 (panel 9). Hg U7 together with Hg W (panel 10) constitute much of the West-Eurasian (WE) specific Hg-s spread in India depicted on panel 2, however, these two seem not to account for recent admixture. Hg W lineages in India and WE overlap only at their ancestral node and both coalescence at ca 25 000 BP (Kivisild et al. 1999). The diversity of U7 lineages in India is comparable to that in Iran but again, the overlap is somewhat restricted to central nodes. Given the well in Palaeolithic coalescence time of U7 (24 000 - 54 000 BP Richards et al. 2000), recent admixture seems highly unlikely.




## Table 3. Haplotype sharing analysis

| | populations belonging into Hindu cast system | Indian tribals | Indo-Aryan speakers | Dravidic speakers | Southern populations | Northern populations | Andhra Pradesh | North-West states of India ¤ | Iran | random control group 1 | | random control group 2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample size n | 1012 | 643 | 867 | 641 | 812 | 773 | 519 | 468 | 437 | 815 | 95 | 840 | 95 |
| Haplotypes n * | 600 | 295 | 507 | 309 | 415 | 472 | 250 | 299 | 300 | 436 | 61 | 466 | 58 |
| Shared haplotypes between groups | 9,00% *0,544* [1] | | 8,80% *0,584* | | 7,79% *0,491* | | 5,2% *0,530* | | | 9,84% *0,595* | | 0,45% *0,027* | |
| | | | | | | | | 3,47% *0,384* | | | | | |
| | | | | | | | | | 1,67% *0,175* | | | | |

*[1]relation of shared haplotypes to pooled sample size*

indicates 95% confidence level

* total number of unique haplotypes is 750 for language coloumn, 539 and 576 for Andra Pradesh / Iran and Andra Pradesh / N-W Indian states respectively.

¤ Punjab, Uttar pradesh, Gujarat, Rajastan

Table 3 presents the results of the haplotype sharing analysis. Haplotypes were defined as HVS1 sequence variation together with haplogroup designation (to be able to tell apart the alike HVS1 haplotypes of different haplogroups). As a control test, we divided all Indian populations (see Table 2) randomly into two groups and determined the proportion of shared haplotypes between the groups. We repeated this analysis for ten times to be able to calculate mean and 95% credible region. We found that ca 10% of the observed haplotypes were shared between the two groups. When comparing the proportions of shared haplotypes of two two-group sets, different sample sizes of the sets would induce a bias. To somewhat diminish this bias we present the proportion of shared haplotypes also in relation to the sample size of the set (printed in italic). We found that neither grouping the Indian populations into tribals and the rest, nor grouping them by the two main language groups (Indo-Aryan and Dravidic) induced a noteworthy decline as far as proportion of shared haplotypes was concerned. A slight additional differentiation was observed in case of northern and southern grouping. Comparing Iranian sample and with samples from Andhra Pradesh and N-W states of India resulted in somewhat expected 3,5 fold decrease in the proportion of shared haplotypes in the first case and 1,5 fold decrease in the latter case.

## Conclusions

I. As general for India, mtDNA haplogroups M and U were found to be dominant in the six studied populations of northern and eastern India.

II. We characterise here four new Indian Hg M subclusters, covering 15% of Hg M lineages in India.

III. We show that Indian specific Hg M subclusters show distinct spread patterns inside India.

IV. Indian tribal populations and Austro-Asiatic speakers in particular, are often considered to be the otherwise lost genetic relics of the indigenous (Palaeolithic) inhabitants of India. Maternal lineages of Austro-Asiatic speaking tribals (studied by us and published before, see Table 2) fit well into the Hg M and U dominated framework of Indian maternal gene pool, which coalescences around 50 000 BP. Moreover, on mtDNA haplotype level (proportion of exact matches) Indian tribals and cast populations do not differ more than any two random groups composed of Indian populations. As much as 20% of the haplotypes detected among the Austro-Asiatic speakers are shared with the Indo-Aryan speakers. Taken together, it strongly suggests a common, Indian-specific origin of the maternal gene pool of the Indian tribal and caste groups.

## References

Bamshad, M., A. E. Fraley, M. H. Crawford, R. L. Cann, B. R. Busi, J. M. Naidu, and L. B. Jorde. 1996. mtDNA variation in caste populations of Andhra Pradesh, India. Hum Biol 68:1-28.

Bamshad, M., T. Kivisild, W. S. Watkins, M. E. Dixon, C. E. Ricker, B. B. Rao, J. M. Naidu, B. V. Prasad, P. G. Reddy, A. Rasanayagam, S. S. Papiha, R. Villems, A. J. Redd, M. F. Hammer, S. V. Nguyen, M. L. Carroll, M. A. Batzer, and L. B. Jorde. 2001. Genetic evidence on the origins of indian caste populations. Genome Res 11:994-1004.

Bamshad, M. J., W. S. Watkins, M. E. Dixon, L. B. Jorde, B. B. Rao, J. M. Naidu, B. V. Prasad, A. Rasanayagam, and M. F. Hammer. 1998. Female gene flow stratifies Hindu castes. Nature 395:651-2.

Bandelt, H.-J., V. Macaulay, and M. Richards. 2000. Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. Mol Phylogenet Evol 16:8-28.

Berger, J. O. 1985. Statistical decision theory and Bayesian analysis. Springer-Verlag, New York.

Forster, P., A. Torroni, C. Renfrew, and A. Rohl. 2001. Phylogenetic Star Contraction Applied to Asian and Papuan mtDNA Evolution. Mol Biol Evol 18:1864-81.

Helgason, A., S. Sigurdadottir, J. Gulcher, R. Ward, and K. Stefanson. 2000. mtDNA and the origins of the Icelanders: deciphering signals of recent population history. Am J Hum Genet 66

Kivisild, T., M. J. Bamshad, K. Kaldma, M. Metspalu, E. Metspalu, M. Reidla, S. Laos, J. Parik, W. S. Watkins, M. E. Dixon, S. S. Papiha, S. S. Mastana, M. R. Mir, V. Ferak, and R. Villems. 1999a. Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. Curr Biol 9:1331-1334.

Kivisild, T., K. Kaldma, M. Metspalu, J. Parik, S. S. Papiha, and R. Villems. 1999b. The place of the Indian mitochondrial DNA variants in the global network of maternal lineages and the peopling of the Old World. Pp. 135-152 in R. Deka and S. S. Papiha, eds. Genomic Diversity. Kluwer/Academic/Plenum Publishers.

Kivisild, T., S. S. Papiha, S. Rootsi, J. Parik, K. Kaldma, M. Reidla, S. Laos, M. Metspalu, G. Pielberg, M. Adojaan, E. Metspalu, S. S. Mastana, Y. Wang, M. Gölge, H. Demirtas, E. Schnekenberg, G. F. Stefano, T. Geberhiwot, M. Claustres, and R. Villems. 2000. An Indian Ancestry: a key for understanding human diversity in Europe and beyond. Pp. 267-279 in C. Renfrew and K. Boyle, eds. Archaeogenetics: DNA and the population prehistory of Europe. McDonald Institute for Archaeological Research University of Cambridge, Cambridge.

Mountain, J. L., J. M. Hebert, S. Bhattacharyya, P. A. Underhill, C. Ottolenghi, M. Gadgil, and L. L. Cavalli-Sforza. 1995. Demographic history of India and mtDNA-sequence diversity. Am J Hum Genet 56:979-92.

Papiha, S. S., S. M. Chahal, and S. S. Mastana. 1996. Variability of genetic markers in Himachal Pradesh, India: variation among the subpopulations. Hum Biol 68:629-54.

Quintana-Murci, L., O. Semino, H.-J. Bandelt, G. Passarino, K. McElreavey, and A. S. Santachiara-Benerecetti. 1999. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat Genet 23:437-41.

Richards, M., V. Macaulay, E. Hickey, E. Vega, B. Sykes, V. Guida, C. Rengo, D. Sellitto, F. Cruciani, T. Kivisild, R. Villems, M. Thomas, S. Rychkov, O. Rychkov, Y. Rychkov, M. Golge, D. Dimitrov, E. Hill, D. Bradley, V. Romano, F. Cali, G. Vona, A. Demaine, S. Papiha, C. Triantaphyllidis, and G. Stefanescu. 2000. Tracing european founder lineages in the near eastern mtDNA pool. Am J Hum Genet 67:1251-76.

Roychoudhury, S., S. Roy, A. Basu, R. Banerjee, H. Vishwanathan, M. V. Usha Rani, S. K. Sil, M. Mitra, and P. P. Majumder. 2001. Genomic structures and population histories of linguistically distinct tribal groups of India. Hum Genet 109:339-50.