MBE Advance Access published October 26, 2010

Research Article

Population Genetic Structure in Indian Austroasiatic speakers: The Role of Landscape Barriers and Sex-specific Admixture

Gyaneshwer Chaubey[1], Mait Metspalu[1], Ying Choi[2], Reedik Mägi[3,4], Irene Gallego Romero[5], Pedro Soares[6], Mannis van Oven[2], Doron M. Behar[1,7], Siiri Rootsi[1], Georgi Hudjashov[1], Chandana Basu Mallick[1], Monika Karmin[1], Mari Nelis[8], Jüri Parik[1], Alla Goverdhana Reddy[9], Ene Metspalu[1], George van Driem[10], Yali Xue[11], Chris Tyler-Smith[11], Kumarasamy Thangaraj[9], Lalji Singh[9], Maido Remm[4], Martin B. Richards[6], Marta Mirazon Lahr[5], Manfred Kayser[2], Richard Villems[1], and Toomas Kivisild[1,5*]

1 – *Department of Evolutionary Biology, Institute of Molecular and Cell Biology, University of Tartu and Estonian Biocentre, Tartu, Estonia*

2 – *Department of Forensic Molecular Biology, Erasmus University Medical Center Rotterdam, The Netherlands*

3 – *Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK*

4 - *Department of Bioinformatics, Institute of Molecular and Cell Biology, University of Tartu and Estonian Biocentre, Tartu, Estonia*

5 –*Leverhulme Centre of Human Evolutionary Studies, The Henry Wellcome Building, University of Cambridge, Fitzwilliam Street, Cambridge, CB2 1QH, UK*

6 - *Institute of Integrative and Comparative Biology, Faculty of Biological Sciences, University of Leeds, Leeds, UK*

7 - *Molecular Medicine Laboratory, Rambam Health Care Campus, Haifa 31096, Israel*

8 -*Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, Estonia; Genotyping Core Facility, Estonian Biocentre, Tartu, Estonia and Estonian Genome Center, University of Tartu, Estonia*

9 - *Centre for Cellular and Molecular Biology, Hyderabad, India.*

10- *Himalayan Languages Project, Institut für Sprachwissenschaft, Universität Bern, Länggassstrasse 49, 3000 Bern 9, Switzerland*

11 - *The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, UK.*

Keyword: Austroasiatic, mtDNA, Y-Chromosome, Autosomes**,** Admixture

***Corresponding Author:**

Dr. Toomas Kivisild
Leverhulme Centre of Human Evolutionary Studies, The Henry Wellcome Building, University of Cambridge Fitzwilliam Street, Cambridge, CB2 1QH**,** UK
Email: tk331@cam.ac.uk

ABSTRACT

The geographic origin and time of dispersal of Austroasiatic (AA) speakers, presently settled in South and Southeast Asia, remains disputed. Two rival hypotheses, both assuming a demic component to the language dispersal, have been proposed. The first of these places the origin of Austroasiatic speakers in Southeast Asia with a later dispersal to South Asia during the Neolithic, whereas the second hypothesis advocates pre-Neolithic origins and dispersal of this language family from South Asia. To test the two alternative models this study combines the analysis of uniparentally inherited markers with 610,000 common SNP loci from the nuclear genome. Indian AA speakers have high frequencies of Y chromosome haplogroup O2a; our results show that this haplogroup has significantly higher diversity and coalescent time (17-28 KYA) in Southeast Asia, strongly supporting the first of the two hypotheses. Nevertheless, the results of principal component and "*structure*-like" analyses on autosomal loci also show that the population history of AA speakers in India is more complex, being characterised by two ancestral components - one represented in the pattern of Y chromosomal and EDAR results, the other by mtDNA diversity and genomic structure. We propose that AA speakers in India today are derived from dispersal from Southeast Asia, followed by extensive sex-specific admixture with local Indian populations.

INTRODUCTION

Austroasiatic is the eighth largest language family in the world in terms of the number of native speakers (104 million) (Lewis 2009). As its name implies, it is spoken in southern parts of Asia, - in Vietnam and Cambodia as the main official languages, and in India, Bangladesh, Nepal, Burma, Laos, Thailand and Malaysia as the first language of many minority groups that are isolated from each other by other language speakers. Two major extant branches of the Austroasiatic language tree are Munda in eastern, northeastern and central India and Khasi-Aslian, which stretches from the Meghalaya in the northeast of the subcontinent to the Nicobars, Malay peninsula and Mekong delta in Southeast Asia (Figure 1A). Since the birth of historical linguistics in the 1640s, attempts have been made to explain the wide and continuous geographic spread of some language families, such as the Indo European, Uralic and Bantu, in contrast to the more patchy or constrained distribution of others, *e.g.* the Basque and Khoi-San languages. Models proposed to explain the success of a few rather than many language families range from those stressing pure demic diffusion to pure cultural diffusion driven by some economic or technological advance as the key mechanism of the language spread. One of the prehistoric events that has been considered as a plausible device to fuel both demographic and cultural spread is the shift from a hunter-gatherer to an agricultural mode of subsistence thought to have occurred independently in only a few places in the world (Ammerman and Cavalli-Sforza 1984). However, the attempt at explaining the success of the ten most widely spoken language families of the world in terms of the Neolithic demic diffusion model (Diamond and Bellwood 2003) – that is, by linking the spread of languages, genes, and economy – has been challenged in almost every single case (Richards et al. 2000; Ehret et al. 2004; Fuller 2003). The hypothesis that the spread of the Austroasiatic language family can be traced back to rice cultivators of Southeast Asia

(Higham 2003; Bellwood 2005) is contested, but some relationship between early Austroasiatics and rice agriculture is a view which remains prevalent among linguists.

The Higham–Bellwood model (Higham 2003; Bellwood 2005) considers Indian Munda and Khasi-Aslian speaking hunter-gatherer populations, who regardless of their current lifestyle, share rice cultivation related cognates with Khasi-Aslian speaking populations of Southeast Asia, as Neolithic immigrants in India, because traditionally a single origin of rice cultivation in China has been assumed (Figure 1B). However, as argued by Fuller (Fuller 2007), the genetic evidence of independent domestications for the *Oryza indica* and *japonica* cultivars of *Oryza japonica* suggests a plausible alternative scenario (Figure 1C) by which the homeland of the Austroasiatic family lies in India. If *O. indica* rice was indeed domesticated first in India, then its spread to Southeast Asia may have been coupled with the spread of Austroasiatic speakers (Fuller 2007). However, the phylogenetic evidence from genes associated with rice domestication is not unequivocal – phylogenies of some functionally important genes continue to support the single origin model (*e.g.*, Tan et al. 2008; Jin et al. 2008). Opposing evidence from different genes may be reconciled by a model according to which the domestication was a lengthy process extending back to and even beyond the Last Glacial Maximum, as opposed to the earlier view of a rapid transition which placed the domestication of crops to the Pleistocene/Holocene boundary (Allaby et al. 2008). However, according to current archaeological evidence, the shift to a lifestyle where rice would be an essential staple food would be younger than 7 KYA (thousand years ago) in China and even more recent in India (Purugganan and Fuller 2009; Fuller et al. 2009). In the light of the archaeobotanical, linguistic and rice genomic evidence the differentiation of Austroasiatic languages into their major subgroups could therefore be placed either in South or Southeast Asia with their split or the latest date of contact probably being more recent than 7 KYA.

Genetic studies on human populations of South and Southeast Asia have, hitherto, proved to be inconclusive about the two opposing models of the geographic origins of the Austroasiatic speaking people and about the timing of the split between the two major branches in this language family. The mtDNA information available so far indicates a clear distinction of Indian Munda and Southeast Asian Khasi-Aslian speaking groups, as both share their mtDNA haplogroups with their regional neighbours who speak languages other than Austroasiatic (Figure 2 and Table 2). Consistent with this linguistic separation, the Khasi-Aslian speaking Nicobarese carry almost exclusively East Asian specific mtDNA (Thangaraj et al. 2005). Notably, Khasi (the only Khasi-Aslian group of mainland India) speakers residing in Meghalaya state in India show an admixed package of both Indian and East Asian mtDNA haplogroups (Figure 2 and Table 2). Overall, the mtDNA haplogroup distributions make a clear distinction between Indian and Southeast Asian Austroasiatic speakers; because of the lack of shared lineages this evidence is not informative about any shared phase of evolutionary history of Munda and Khasi-Aslian speaking populations. In contrast, Y chromosome haplogroup O2a occurs frequently both among Indian and Southeast Asian Austroasiatic speakers (Table 2) and thus appears as evidence for some degree of shared ancestry (Kivisild et al. 2003). Because all other branches of haplogroup O are largely restricted to East Asia, and given the recent time depth of Y-STR variation of Indian haplogroup O2a, its recent (<10 KYA) entry from Southeast Asia (Figure 1B) has been implied in some studies (Sahoo et al. 2006; Sengupta et al. 2006). On the one hand, the frequency of haplogroup O lineages in India is correlated with languages boundaries and cannot be explained only by isolation-by-distance (Figure 2A and Table 2). On the other, high levels of genetic diversity of mtDNA haplogroups in Munda speakers and an independent assessment of Y-STR diversity of haplogroup O2a in India, dating its origin to ~65 KYA, have been used to argue in

favour of a model that assumes direct descent of Austroasiatic speakers from the initial settlers of India (Figure 1C), and their subsequent dispersal to Southeast Asia, possibly before the Last Glacial Maximum (Basu et al. 2003; Kumar et al. 2007; Chakravarti 2009). Arguably, the more recent (<10 KYA) estimates of the age of O2a variation in India could have been deflated by limited regional sampling. It should be noted, however, that the 65 KYA dating of haplogroup O2a in India appears much older than the estimated age of its ancestral haplogroups K and NO (Rootsi et al. 2007; Karafet et al. 2008). Moreover, the Southeast Asian populations have been underrepresented in all previous studies, and, furthermore, no high resolution autosomal evidence has been considered in these debates. Therefore, the genetic origins of Austroasiatic speaking populations remain largely controversial.

In this paper, we sought to investigate the extent of population structure and admixture among the Indian and Southeast Asian AA speakers embedded in their autosomal genomes and to combine the results obtained with data from uniparental loci and from regional selection signatures, such as that of the EDAR gene. We used *Illumina* HumanHap *610K* genotyping chips on 45 diverse Indian samples covering three major language groups from India relevant to our study ((22 Austroasiatic (19 Munda and 3 Khasi-Aslian), 19 Dravidian (Behar et al. 2010), and 4 Tibeto-Burman speakers)) and 15 Burmese samples from Myanmar. These results were combined with the global data set (Li et al. 2008), generated with *Illumina* HumanHap *650K* chips, which, among others, included a set of Pakistani populations as proxy for the Indo-European speakers of South Asia and a sample of 10 individuals from Cambodia which is predominantly a Khmeric speaking country (for a full list of populations and sample sizes see supplementary Table 1).

MATERIALS AND METHODS

A detailed description of experimental procedures can be found in the Supplementary Experimental Procedures. The genotyping experiments for *Illumina* HumanHap *610K* on new 41 Indian and Burmese samples were carried out according to manufacturers' specifications. We combined our newly generated data with relevant reference datasets from Stanford HGDP SNP Genotyping Data (http://hagsc.org/hgdp/files.html) and 19 Dravidian from Behar et al. (2010) (Supplementary Table 1). The *EDAR* 1540T/C, a nonsynonymous SNP in exon12 was genotyped by PCR-direct sequencing using forward-GTAGGTCTTAGCCCCAC (Annealing T=54^0C) and reverse CATCCAGCCGCTCAATC (Annealing T=54^0C) primers. Altogether, 1077 Indian samples were assayed for this polymorphism. In total, 1563 Y chromosome samples were analyzed in this study. NRY specific multiplex (Indian Y-Plex) PCR was designed to characterize 589 Indian AA and TB samples. The ABI 3100 Genetic Analyzer (Applied Biosystems) was used for genetic typing. Fragment sizes were determined using the GeneMapper® Analysis Software v4.0 and allele designations were based on comparison with allelic ladders included in the Y-filer™ kit.

PC and Admixture analyses of genome wide SNP data

We used PLINK 1.05 (Purcell et al. 2007) to filter the combined dataset to include only SNPs on the 22 autosomal chromosomes with minor allele frequency >1% and genotyping success over 97%. Because background linkage disequilibrium (LD) can affect both PCA (Patterson et al. 2006) and "*structure*-like" analysis (Alexander et al. 2009) we thinned the dataset by excluding SNPs unique to either of the two Illumina platforms, SNPs from mtDNA, X and Y chromosomes and removing one SNP of a pair in a strong LD r^2>0.4 in a window of 2,000 SNPs (sliding the window by 25 SNPs at a time), the combined data set had data for 215,729 SNPs that were used in subsequent analyses. For PCA we generated an additional dataset with the same filters but excluding the African samples yielding a matrix of 631 samples by 189,512 SNPs.

We carried out PC analysis using smartpca program (with default settings) of the EIGENSOFT package (Patterson et al. 2006) to capture genetic variation described by the first 10 PCs. The fraction of total variation described by a PC is the ratio of its eigenvalue to the sum of all eigenvalues (Figure 3A).

Of the several "*structure*-like" (baptized by (Weiss and Long 2009) algorithms, we experimented with *Frappe* (Tang et al. 2005; Li et al. 2008) and ADMIXTURE 1.4 (Alexander et al. 2009), running the dataset with different settings several times. Although we settled on using the latter mostly due to faster computation time, we note that *Frappe* gave very similar results. In the final setting we ran ADMIXTURE with random seed number generator on the LD pruned dataset one hundred times at $K = 2$ to $K = 10$. Following an established procedure, we examined the Loglikelihood scores (LLs) of the individual runs and found that up to K = 9 (incl.) the maximum difference between LLs in the 10% fraction of the runs with the highest LLs was minimal (<1 LLs unit). Thus, we could, with some confidence, assume that these individual runs from K = 2 to K = 9 converged on the global maximum. The new version of ADMIXTURE (1.4) assists in choice of K with a cross validation (CV) procedure (we used hold-out fraction 0.1). The lowest CV scores we obtained at K =7 (Figure 3B). This choice of K was further bolstered by the observation that at higher Ks the new emerging clusters (ancestry components) were largely restricted to one population and thus of little interest in a population comparison study. However, plots of all converged Ks are given in Supplementary Figure 1. For plotting we took one run from the 10% fraction of runs with the highest LLs. We note however, that vast majority of the runs at each K (K = 2 to K = 7) yielded very similar LLs (on the same plateau of LLS distribution) indicating very similar (visually indistinguishable) cluster (ancestry components) distribution.

Using PLINK, we pruned our initial autosomal data set and excluded one from each pair of SNPs with LD $r^2 > 0.1$ in a 50 SNP window shifted at 10 SNP intervals to ensure complete data independence. This procedure resulted in a pruned data set containing 54,355 SNPs from which we calculated mean pairwise F_{ST} differences between linguistic and continental population groups using the method of Weir and Cockerham (Cockerham and Weir 1984). We also calculated Hs and Ho for all autosomal SNPs, in accordance to Nei (Nei 1987). Great circle distances were calculated as in Ramachandran et al.(Ramachandran et al. 2005).

Statistical Analysis (Y-STR)

Number of haplotypes and average number of pairwise difference (Supplementary Table 2) of Y-STR for studied populations were calculated using the Arlequin 3.01 software package (Excoffier et al. 2005). DYS 389I (DYS 389cd) was subtracted from DYS389II and re-named 389ab. A median-joining network, resolved with the MP algorithm, was constructed using the Network package (version 4.5.0.2) (www.fluxus-engineering.com); one Steiner tree is shown in figure 5B. The M95 (O2a) variance isofrequency map was generated using Surfer 8 (Golden Software Inc., Golden, Colorado), following the Kriging procedure. The age of M95 (O2a) was estimated from microsatellite variation within the haplogroup using the method described by (Zhivotovsky et al. 2004) and updated in (Sengupta et al. 2006). Moreover, different founders were identified based on Network analysis of Munda speakers. The age of these founders was estimated from the ρ statistic (the mean number of mutations from the assumed root of each and every founder), using a 25-year generation time and the *TD* statistic, assuming a mutation rate of 6.9×10^{-4} (Zhivotovsky et al. 2004), based on variation at 14 common Y-STR loci (Supplementary Table 3).

RESULTS AND DISCUSSION

Assessing autosomal population structure and admixture in Austroasiatic speakers

In Figure 3A, we present the PC analyses for Eurasian populations only. Principal component (PC) analysis (Figure 3A) resulted in a crude reflection of the geographic locations of the studied populations. We also performed PC analysis with the whole dataset. Naturally, the first component there differentiates Africans from all other populations and PC2 and PC3 correspond very closely (data not shown) to PC1 and PC2 of the Eurasian PC plot (Figure 3A). However, the Eurasian PC analysis shows better resolution on the east–west and north–south axes within Eurasia, thus being better suited to answering the questions we address in the present study. For example, the two Pakistani samples (HGDP00130 and HGDP00175), which show a high level of admixture with Africans, were positioned surprisingly close to the Khasi samples on the PC2/PC3 plot in the global PC analysis (plot not shown). In the Eurasian PC analysis, the first component (explaining 5.6% of the total variation) differentiates West from East Eurasia while the second component (1%) separates South Asians from the rest. None of the first 10 significant PC-s clustered the Munda speaking populations from the Indian subcontinent together with Khasi-Aslian speaking populations of Southeast Asia. In the first two principal components the Munda speakers from the eastern states of India cluster close to the Dravidian speakers while being slightly shifted towards the East Asian cluster by PC1 (Figure 3A). The Khasi-Aslian speaking Khasi, on the other hand, are closer to East Asians than to the Dravidian speakers. The position of the Garo (Tibeto-Burman speakers) overlaps with that of the Cambodians (Khasi-Aslian) who cluster with Tibeto-Burman speaking populations from Myanmar and China while being slightly drawn towards the Indian cluster. Mean genetic distances (F_{ST}) estimated over the whole genome recapitulate the pattern extracted by the first PCs, whereby Munda speakers are most closely related to Indian Dravidian speakers, whereas Khasi-Aslian and Tibeto-Burman

groups from India and Southeast Asia are more similar to each other, although the Indian Khasi-Aslian also have high affinity with Munda speakers (Supplementary Table 4). The PC plots and genetic distance estimates support the view of Southeast Asian origins of Indian Khasi-Aslian (and Tibeto-Burman) speaking populations, while, in contrast, Indian Munda speaking populations draw their genetic ancestry mainly from the source shared with Indian Dravidian speakers.

As another approach we used the "*structure*-like" algorithm ADMIXTURE (Alexander et al. 2009) which gives a maximum likelihood estimate for the population structure of sampled individuals. It assumes a specified number of discrete ancestral populations (K) and computes respective ancestry proportions for each studied individual. The approach should be considered with the caveat that the assumption of discrete ancestral populations is generally not a realistic model of population history (Weiss and Long 2009). Regardless of these conceptual difficulties, the results of the ADMIXTURE analyses may represent a robust picture of the similarities and dissimilarities between studied samples in terms of genetic patterning within the raw data. Thus, with these limitations in mind we note that irrespective of the number of assumed ancestral populations ($2 < K < 7$), the Munda speakers of India show consistently higher proportion of East Asian component than Dravidian or Indo-European speakers of the Indian subcontinent (Supplementary Figure 1).

At $K=7$, the Munda speakers are characterized by two ancestry components (Figure 3B). The predominant "dark green" component makes up approximately three quarters of the Munda ancestry palette. This component is most prominently apparent among the South Indian Dravidian speakers and is relatively rare among the Indo-European speaking Pakistani populations. On the other hand, the Munda speakers lack the "light green" component that is

prevalent among the Indo-European speaking Pakistani populations, and to a minor extent also present in South India, Near East and Europe. The East and Southeast Asian populations show the presence of two ancestry components: the pink component is most clearly pronounced in Oroqen and Hezhens from Northern China whereas the orange component is overwhelming among Cambodians, as well as Burmese of Myanmar and Dai and Lahu populations from Southwest China (Figure 3B). These two components reveal two contrasting patterns of East and Southeast Asian admixture among South Asian populations. Consistent with their Central Asian/Mongolian ancestry, Uygurs and Hazara carry predominantly the pink ancestry component, whilst the Munda speakers exhibit membership only in the orange cluster. Garo, Burmese (both Tibeto-Burman), and, notably, also Khasi (Khasi-Aslian), appear to have both East and Southeast Asian components, regardless of the absence of the pink component among the Khmer speaking Cambodians. While these results are thus consistent with notable (23%, SD 5%) Southeast Asian genetic admixture among Indian Munda speakers, in support of the model presented in Figure1B, there are also detectable traces of South Asian (dark green) admixture among the Cambodians (16%, SD 5%). This finding provides some quantitative support for the alternative model presented in Figure 1C that assumed an Indian origin for the Austroasiatic language family.

The observed patterns of genetic admixture on both sides of the Bay of Bengal suggest that models assuming only one episode of unidirectional gene flow are therefore likely to be oversimplifications in describing the historico-demographic processes underlying the origin and differentiation of the Austroasiatic speaking populations. These patterns could, however, also be understood as a result of long-term gene flow under isolation by distance (IBD) which would be the default model to explain geographic correlations in genetic patterning among populations

(Wright 1943). Arguably, a significant proportion of genetic variation in genome-wide STR and SNP diversity among world-wide populations can be explained by IBD (Handley et al. 2007). The IBD model would predict in our case that the ancestry components revealed by the ADMIXTURE analyses would apply to Indian and Southeast Asian populations regardless of their linguistic affiliation. Indeed, our Illumina whole-genome data for Dravidian speakers come only from populations of Karnataka and Kerala, which are geographically distant from the Austroasiatic groups concentrated in the eastern states of Orissa and Bihar. The whole-genome genotype data from a small number of populations are robust in terms of the number of loci considered while revealing the extent of East and Southeast Asian ancestry among the Indian Austroasiatic speakers. Yet, we would have to use data from a large number of populations, as for example in the case of the Y chromosome diversity patterns (Figure 2B), to address the question of whether the observed patterns in autosomal genes are due to IBD or dispersals.

When populations admix, alleles under positive selection are expected to proliferate more efficiently in the hybrid population than other alleles on average. A positively selected allele could therefore be used in a conservative approach to test the IBD versus dispersal hypotheses because, even in cases of limited gene flow between populations, the positively selected alleles would be expected to show higher than average penetrance, unless, of course, the selection is region specific. Scans of positive selection on genome-wide polymorphism data from global human populations have identified the *EDAR* (ectodysplasin-A receptor) gene as a candidate for the strongest positive selection in East Asians (Sabeti et al. 2007). *EDAR* is a major genetic determinant of hair thickness and with a nonsynonymous allele (Val370Ala) SNP rs3827760 (1540C allele), which shows high frequencies in populations of East Asian and Native American

origin but is essentially absent from European and African populations (Sabeti et al. 2007; Fujimoto et al. 2008).

Interestingly, in India, we observe the 1540C allele mainly in association with AA and TB populations (Figure 4). Tibeto-Burman speakers of India have the highest (~61%) 1540C allele frequency in South Asia, consistent with their predominantly East Asian ancestry inferred from autosomal and uniparental loci. Meanwhile, the Khasi population is characterized by a 40% frequency of the allele (Table 3). Munda speakers also show detectable presence, with a ~5% average, in contrast to its complete absence among Indo-European and Dravidian speakers (with a few exceptions *viz*, Tharu, Mushar, Hazara, and Burusho populations) (Figure 4). These results are in line with the models suggesting gene flow from Southeast Asia to India, albeit more significant among Khasi than Munda speaking populations. Given the evidence for strong positive selection on this allele in East Asia, our finding of only 5% frequency among Munda is surprisingly low, possibly reflecting the fact that the 1540C allele does not carry a significant biological advantage in India.

Overall, the genome-wide autosomal evidence is therefore consistent with bidirectional gene flow between India and Southeast Asia restricted mainly to Austroasiatic (and Tibeto-Burman) speaking populations. The analysis of geographic and linguistic patterns in the distribution of the 1540C allele of the *EDAR* gene in 49 Indian populations (Figure 4), shows that linguistic affiliation appears as a significant predictor of allele frequency and therefore, at least in case of this gene, the IBD model can be rejected. However, our analyses of autosomal variation did not inform us about the timing of the dispersal events.

Dating of the genetic variation in Y-chromosome haplogroup O2a

The autosomal genetic evidence above appears to support previous claims made on the basis of Y-chromosome evidence for the existence of a shared genetic component among Indian and

Southeast Asian Austroasiatic speakers. However, our analyses did not provide a date estimate for these shared elements of population history and furthermore suggested multidirectional gene flow. Genotyping of 12 SNP markers in 553 Y-chromosome samples representing 13 Indian Austroasiatic populations sampled from 15 locations revealed the presence of eight distinct haplogroups among Munda speakers, seven of which they share with other Indo-European and Dravidian speaking Indian populations (Supplementary Figure 2). Consistent with previous studies (Basu et al. 2003; Metspalu et al. 2004; Sengupta et al. 2006; Sahoo et al. 2006; Kumar et al. 2007), the eighth, O2a (M95), appears as the most frequent haplogroup among most Munda speaking populations (Figure 6A, Supplementary Figure 2 and Supplementary Table 5). Khasi (Khasi-Aslian) and Garo (Tibeto-Burman) populations of Northeastern India have two additional hg O subclades, i.e. O3 (M122) and O*, the latter found only in Garo (Supplementary Figure 2). The presence of M122 at moderate frequency in Khasi is consistent with the autosomal data considered above and can be explained by their close geographic proximity to, and likely admixture with, Tibeto-Burman speaking populations (*e.g.* Garo) among whom the O3 lineage is pre-dominant (Cordaux et al. 2004; Reddy et al. 2007).

Previous Y chromosome studies have provided controversial dates for the shared O2a lineage either because of different sampling or genotyping approaches. To avoid these issues we genotyped a wide range of samples both from India and Southeast Asia with the same widely used approach (AmpFℓSTR®*Y-filer*™ kit). Using data from fourteen Y chromosomal short tandem repeat (STR) loci (Supplementary Table 6) we estimate the age of all Y chromosomes from India and Southeast Asia with the M95 mutation as ~20 (±2.7) KY (Table 4). This estimate is significantly younger than the 65 KYA estimate of (Kumar et al. 2007), but similar to the estimates of other haplogroup O sub-clades (Shi et al. 2005). O2a coalescent times appear to be

significantly higher in Southeast Asian populations than in India, in contrast to genome-wide heterozygosity patterns (Supplementary Figure 3), suggesting that the long-term effective population size of Munda Y chromosomes in India has been lower than that of Khasi-Aslian speakers in Southeast Asia (Figure 5C and Table 4). However, the lack of clear regional clustering in the STR-based phylogenetic network (Figure 5B) makes a simple founder-effect scenario unlikely to explain the lower diversity in India – if Southeast Asia is the source of Indian O2a variation, more than one founding lineages would need to have been involved in the migration, and the differentiation time of Indian O2a lineages would have to be considered as the upper boundary of the migration rather than referring to the migration time itself (Table 4). The Shompen remain outliers and stay significantly equidistant from other populations, consistent with the view of their linguistic isolation (Figure 5B).

Our coalescent time estimate of 15.9±1.6 KY for Indian M95 carriers is more than two-fold greater than the age estimated by Sengupta et al. (Sengupta et al. 2006), while being more than four-fold smaller than the one reported by Kumar et al. (2007). All three studies used different sets of STR loci and different ranges of sampling but the same phylogenetic calibration of the molecular clock. The difference between our estimate from that of (Sengupta et al. 2006) can mainly be ascribed to the difference in geographic sampling: when applying the coalescent calculations to the subset of Ho and Santhal samples in our data we observe a value (7.3 ± 1.5 KY) which is not significantly different from the estimate (8.8 ± 2 KY) reported for these same populations by (Sengupta et al. 2006). It should be noted as well that all eight overlapping STR loci between our studies showed identical STR median haplotypes by this approach. Conversely, the age difference between our study and that of (Kumar et al. 2007) cannot be explained by differences in the range of geographic sampling, as both studies cover a wide assortment of

Austroasiatic speaking tribes from India (Supplementary Figure 4). Overall, due to the apparent lack of geographic clustering of Indian Austroasiatic O2a Y-STR haplotypes in the phylogenetic network, our 15.9+1.6 KY age estimate for the Indian subset should not be taken as a genetic estimate of dispersal time of Austroasiatic groups to India, but rather this date estimate can be considered as the upper boundary for any dispersal event(s) to India that involved the O2a lineage.

mtDNA evidence for sex-specific local admixture among Indian Austroasiatic speakers

Similarly to autosomal and Y chromosome data, the mtDNA evidence shows that Munda speakers of India have a substantial overlap with their local Dravidian and Indo-European speaking neighbours in their mtDNA haplogroup composition. However, in contrast to the inferences based on other loci, there is no detectable evidence in >700 DNA samples from the Munda speaking populations for a shared ancestry component with other Austroasiatic groups from Southeast Asia (Table 2).

The mtDNA haplogroup allocation of Munda speakers is similar to Dravidian and Indo-Europeans of the Indian subcontinent (Basu et al. 2003; Metspalu et al. 2004; Chaubey et al. 2007; Chaubey et al. 2008a,b; Thangaraj et al. 2009). We carried out a high resolution analysis of those haplogroups of Munda speakers which account for >4% of their maternal gene pool. All the seven maternal haplogroups found frequently in Munda speakers are autochthonous to India (Supplementary Figure 5) (Chandrasekar et al. 2009) and references therein), accounting altogether, for 57% of the maternal gene pool of present Munda speakers. The extensive analysis of these haplogroups revealed relatively recent sharing of most recent common ancestors within these groups between AA and non-AA speakers (MRCA), suggestive of admixture; a similar result was observed recently for hg R7, which is the most frequent among these in AA speakers (Chaubey et al. 2008b). The mtDNA lineages of Munda speakers do not cluster in basal parts of

the tree (to founder haplogroups M, N or R), but are spread among the derived branches that date to <10KYA (Supplementary Figure 5), suggests that the mtDNA diversity found in contemporary Munda speakers is the result of admixture from neighboring populations of India.

In sharp contrast, among the geographically proximate Khasi-Aslian speaking Khasi population, approximately one third of the mtDNA lineages have Southeast Asian ancestry (Figure 2 and Table 2). Notably, the Khasi are known historically to have been matrilocal. This pattern of sex-specific gene flow is perhaps not unexpected considering the patrilocality that most Munda speaking groups practice today. Previous studies, though, have noted that the genetic effect of patrilocal practice in India is significantly different from Southeast Asia due to different degrees of permeability in the marital boundaries (Kumar et al. 2006).

CONCLUSIONS

Thus, our analyses of genetic data from uniparentally and biparentally inherited loci provide a range of estimates of gene flow across geographic and linguistic borders. The analysis of autosomal data suggests bidirectional gene flow across the Bay-of-Bengal restricted to Austroasiatic and Tibeto-Burman speaking populations. The presence of a significant (approximately one quarter) Southeast Asian genetic component among Indian Munda speakers is consistent with this model, implying their recent dispersal from Southeast Asia followed by extensive admixture with local Indian populations. The strongest signal of Southeast Asian genetic ancestry among Indian Austroasiatic speakers is maintained in their Y chromosomes, with approximately two thirds falling into haplogroup O2a. Geographic patterns of genetic diversity of this haplogroup are consistent with its origin in Southeast Asia approximately 20 KYA, followed by more recent dispersal(s) to India. Comparison of mtDNA and Y chromosome data reveals that the “import of local genes”, at least in case of the Munda speakers of India, has

likely been biased towards the female sex resulting in a situation where the Southeast Asian ancestry signal in the mtDNA lineages of Indian Munda speakers has been entirely lost. Further sampling of Southeast Asian Austro-Asiatic speaking populations and genome-wide sequence data along with in-silico simulations would be required in the future to assess the demographic parameters of population dispersals between South and Southeast Asia in explicit time frames.

Supplementary Data

Supplemental Data include includes six figures, and ten tables and can be found with this article online.

Acknowledgements:

We thank Ille Hilpus, Tuuli Reisberg, Viljo Soo and Lauri Anton for technical assistance. We also thank to Professor Thomas L. Rost, to allow us to use the rice plant picture (for Figure 1) from http://www-plb.ucdavis.edu/labs/rost/Rice/RICEHOME.HTML. This research was supported by the EU European Regional Development Fund through the Centre of Excellence in Genomics, Estonian Biocentre and University of Tartu, Estonian Basic Research grant SF0182474 (to RV), Tartu University grant PBGMR06901 (to TK), Estonian Science Foundation Grants 7858 (To EM) and 7445 (to SR), UKIERI grant RG47772 (to TK, MML, KT and LS), British Academy BARDA-48208 (to MBR) and Estonian Ministry of Education and Research grant 0180142s08 and European Commission grant nr. 245536 (OPENGENE) (to MN). We also thank European Commission grant ECOGENE 205419 to EBC. D.M.B. thanks to the European Commission, Directorate-General for Research for FP7 Ecogene grant 205419. LS and KT were supported by CSIR, Government of India and and YX and CTS by The Wellcome Trust.

Figure Legends:

Figure 1. (A) Language tree of the major subgroups of the Austroasiatic (AA) language family according to (Diffloth 2009). The branching of the hypothetically extinct para-Munda languages Melluha and Kubha-Vipas is shown by a broken line. The branching pattern of the extant languages allows for both South and Southeast Asia to be considered equally as potential homelands for the initial spread of AA. According to Fuller (2007) the acceptance of the extinct para-Munda branch would support the origin of AA in the Indian subcontinent. The map depicts the geographic distribution of the AA family (adopted from Diffloth 2001 and Anderson 2007 covering Southeast Asia and India respectively) and the sampling locations (with the precision of district) for the Indian AA samples. Numbers correspond to populations as given it Table 1. Note, that for India only the concentrated AA regions are highlighted. Munda speakers can be found in low frequencies throughout East India, thus the few sampling locations outside the shown AA areas still represent AA populations. (B) Out of Southeast Asia and (C) Out of India dispersal models. These two models represent two alternative views to explain the spread of AA speaking populations, all sharing rice domestication related vocabulary, in South and Southeast Asia. According to model B the AA family originated in Southeast Asia. This model requires only one domestication event of rice in East Asia. In contrast, model C implies the origin of the AA family and its initial split in India. According to this model, *O. indica* and *O. japonica* rice were independently domesticated in what today are India and China. Recent gene flow between local Indian (Ind) non-AA groups and Munda speakers (Mun) in model B and between Khasi-Aslian (Kh-As) and local East Asian (EAs) derived populations is indicated by broken lines. Depending on the extent of the recent admixture, model B allows for preservation of some Southeast Asian genetic ancestry among Munda whereas no distinguishable Indian contribution is expected among Khasi-Aslian groups of Southeast Asia. Conversely, model C assumes continuity of Munda groups in India with no specific East Asian contribution to their genes (apart from secondary gene flow from local Tibeto-Burman groups of India), while Khasi-Aslian would be expected to represent admixture between populations derived from the Indian subcontinent and Southeast Asia.

Figure 2. Scatter plot, showing Southeast Asian specific lineages among different linguistic groups of India. The geographical distribution of Munda languages in India is mainly governed by longitudinal distances, therefore, frequencies of Y chromosome (left panel) and mtDNA (right panel) haplogroups are plotted against longitudinal distances (X-axis). Mushar and Tharu (who now speaks Indo European language and showing exceptional levels of

East Asian haplogroups in contrast to their linguistic affiliation) are arrow marked. South Asian haplogroups:- mtDNA- M2-6, N5, M33-65, R5-8 and R31-32; Y-Chromosome- C5, F, H, L and R2. Southeast Asian haplogroups:- mtDNA: A-G, M7-12, R22 and N9; Y-Chromosome: C2, C3, D and M-O. Unresolved haplogroups:- mtDNA: M*, R*, N* including other lineages, *e.g.* M31 and West Eurasian specific; Y-Chromosome: C*, G, I-K*, P*, Q and R1. Haplogroup frequencies and associated references are given in detail in supplementary information (Supplementary Tables 9 and 10).

Figure 3. (A) Principal component analysis of Indian Austroasiatic, Dravidian and Tibeto-Burman groups in the context of other Eurasian populations. PC analysis was carried out using smartpca program (with default settings) of the EIGENSOFT package. After filtering SNPs (see Methods for detail), the combined data set yielded a matrix of 615 samples with 189,533 SNPs. **(B)** Bar plot displays individual ancestry estimates for studied populations from a *structure* analysis by using ADMIXTURE with $K = 7$.

Figure 4. (A) Geographic distribution of the *EDAR* 1540C allele frequency world-wide. The map was generated using Surfer8 of Golden Software (Golden Software Inc., Golden, Colorado), following the Kriging procedure. Red dots indicate sampling location. (B) Geographic distribution of the *EDAR* 1540C allele frequency in different groups of South and Southeast Asia. The frequency is shown in proportion to the bubble size.

Figure 5. Surfer maps showing (A) the frequency and (B) the mean microsatellite variance distributions of haplogroup O2a (M95) in South and Southeast Asia. Surfer maps were generated using Surfer8 of Golden Software (Golden Software Inc., Golden, Colorado), following the Kriging procedure. (C) Phylogenetic network relating Y-STR haplotypes within haplogroup O2a (M95). The network was constructed using a median-joining with MP (maximum parsimony) algorithm as implemented in the Network 4.5.0.2 program. The size of the circles is proportional to the number of samples.

Table 1. Detailed description of AA and TB samples typed for O2a (M95), EDAR and *Illumina* HumanHap *610K* (WGA) in present study. ‘nr’ is the code of population shown in Figure 1 and Supplementary Figure 4.

nr	Population	Indian State	District	Language group	n for M95	n for EDAR	n for WGA
1	Bonda	Orissa	Koraput	South Munda	42	38	4
2	Savara	Orissa	Koraput	South Munda	21	38	2
3	Gadaba	Orissa	Koraput	South Munda	27	28	1
4	Birhor	Chattishgarh	Raipur	North Munda	27	35	-
5	Birhor	Maharastra	Chandrapur	North Munda	35	15	-
6	Juang	Orissa	Sambhalpur	South Munda	54	20	2
7	Baiga	Orissa	Kendujhar	South Munda	42	21	-
8	Mahli	Jharkhand	Bokaro	North Munda	32	20	-
9	Mawasi	Jharkhand	Gumla	North Munda	27	29	-
10	Santhal	Jharkhand	Gumla	North Munda	20	19	1
11	Kharia	Chattishgarh	Raigarha	South Munda	37	20	2
12	Baiga	Madhya-Pradesh	Guna	South Munda	23	19	-
13	Mawasi	Madhya-Pradesh	Bhopal	North Munda	12	10	-
14	Ho	Bihar	Begusarai	North Munda	45	32	5
15	Khasi	Meghalaya	East Garo hills	Khasi-Aslian	21	20	3
16	Garo	Meghalaya	East Garo hills	Tibeto-Burman	25	20	4
17	Asur	Jharkhand	Dhanbad	North Munda	13	-	-
18	Asur	Jharkhand	Ranchi	North Munda	48	35	1
19	Asur	Jharkhand	Palamau	North Munda	27	-	1
20	Burmese	-	-	Tibeto-Burman	-	-	15
21	Cambodians	(Li et al. 2008; Xue et al. 2009)		Khasi-Aslian	3	10	10

Table 2. mtDNA and Y chromosome haplogroup profiles in South (S) and Southeast (SE) Asia by population, n= number of samples, AA=Austroasiatic, KA=Khasi-Aslian, MK-Mon-Khmer. * Tharu and Mushar populations, who have frequent East Asian haplogroups, are included in Indo-European speakers. ** The Southeast Asian Y-chromosomal frequency of Munda and Tibeto-Burman speakers of India is due to the presence of haplogroup O2a (M95) and O3 (M122) respectively. See Supplementary Tables 9 and 10 for detailed information on the populationwise frequency.

		mtDNA				Y chromosome			
		n	S-Asian	SE-Asian	unresolved	*n*	S-Asian	SE-Asian	unresolved
Nicorbarese (MK/KA/AA) speakers		46	2,18	91,3	6,52	11	0	100	0
Khasi (KA/AA) speakers	South Asia	363	39,67	38,57	21,76	465	10,11	74,62	15,27
Munda/AA speakers	South Asia	742	75,2	0	24,8	1572	26,78	60,56**	12,66
Indo-European speakers*	South Asia	838	59,07	12,65	28,28	1593	43,69	14,12	42,19
Dravidic speakers	South Asia	665	59,55	0,3	40,15	1445	62,63	2,49	34,88
Tibeto-Burman speakers	South Asia	139	2,16	66,91	30,94	242	7,44	85,95	6,61
KA/AA speakers	SEA	138	0	88,41	11,59	395	1,27	89,11	9,62
Tibeto-Burman speakers	SEA	523	0,57	75,72	23,71	387	1,55	66,93	31,52

Table 3. The frequency of 1540C allele of EDAR gene in India by language family. Global frequencies of the 1540C allele are given in Supplementary Table 2.

Language group	*n* (number of samples)	1540C
Tibeto-Burman	57	0.61
Austroasiatic (Khasi-Aslian)	20	0.40
Austroasiatic (Munda)	379	0.05
Indo-European	338	0.01
Dravidian	283	0.00

Table 4. Y chromosome age estimates for population groups of India and Southeast Asia.

Group	Sample size (*n*)	Age (kya)
India (overall)	**178**	**15.9 ± 1.6**
North Munda	87	12.4 + 1.3
South Munda	79	18.4 ± 2.4
Garo	6	5.8 ± 2.7
Khasi	6	10.6 ± 2.6
Southeast Asia (overall)	**142**	**22.4 ± 4.9**
Islands (Southeast Asia)	120	20.8 ± 4.9
Mainland (Southeast Asia)	22	23.8 ± 4.2
Nicobarese	**11**	**16.9 ± 5.9**
Shompen	**10**	**15.3 ± 10.8**
O2a (M95) overall	**331**	**19.5 ± 2.7**

References:

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19:1655-64.

Allaby RG, Fuller DQ, Brown TA. 2008. The genetic expectations of a protracted model for the origins of domesticated crops. *Proc Natl Acad Sci U S A.* 105:13982-86.

Ammerman AJ, Cavalli-Sforza LL. 1984. The Neolithic transition and the genetics of populations in Europe. Princeton: Princeton University Press

Anderson GDS. 2007. The Munda verb: typological perspectives. Mouton De Gruyter

Basu A, Mukherjee N, Roy S, et al. (11 co-authors). 2003. Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Res.* 13:2277-90.

Behar DM, Yunusbayev B, Metspalu M, et al. 2010. The genome-wide structure of the Jewish people. Nature. 466:238-42.

Bellwood PS, ed. 2005. First farmers. Wiley-Blackwell.

Chakravarti A. 2009. Human genetics: Tracing India's invisible threads. *Nature.* 461:487-88.

Chandrasekar A, Kumar S, Sreenath J, et al. (20 co-authors). 2009. Updating phylogeny of mitochondrial DNA macrohaplogroup m in India: dispersal of modern human in South Asian corridor. *PloS one.* 4:e7447.

Chaubey G, Metspalu M, Kivisild T, Villems R. 2007. Peopling of South Asia: investigating the caste-tribe continuum in India. *Bioessays.* 29:91-100.

Chaubey G, Metspalu M, Karmin M et al. (14 co-authors). 2008a. Language shift by indigenous population: a model genetic study in South Asia. *International Journal of Human Genetics*. 8:41.

Chaubey G, Karmin M, Metspalu E, et al. (31 co-authors). 2008b. Phylogeography of mtDNA haplogroup R7 in the Indian peninsula. *BMC Evol Biol.* 8:227.

Cockerham CC, Weir BS. 1984. Covariances of relatives stemming from a population undergoing mixed self and random mating. *Biometrics.* 40:157-64.

Cordaux R, Weiss G, Saha N, Stoneking M. 2004. The northeast Indian passageway: a barrier or corridor for human migrations? *Mol Biol Evol*. 21:1525-33.

Diamond J, Bellwood P. 2003. Farmers and their languages: the first expansions. *Science.* 300:597-603.

Diffloth, G. 2009. More on Dvaravati Old Mon. Paper presented at the Fourth International Conference on Austroasiatic Linguistics, Mahidol University at Salaya, 29 October 2009.

Ehret C, Keita SOY, Newman P. 2004. The origins of Afroasiatic. Science. 306:1680; author reply 1680.

Excoffier L, Laval G, Schneider S. 2005. Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evolutionary bioinformatics online*. 1:47-50.

Fornarino S, Pala M, Battaglia V, Maranta R, Achilli A, Modiano G, Torroni A, Semino O, Santachiara-Benerecetti SA. 2009. Mitochondrial and Y-chromosome diversity of the Tharus (Nepal): a reservoir of genetic variation. *BMC Evol Biol*. 9:154.

Fujimoto A, Kimura R, Ohashi J, et al. 2008. (14 co-authors). A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum Mol Genet.* 17:835-43.

Fuller D. 2003. An Agricultural Perspective on Dravidian Historical Linguistics: Archaeological Crop Packages, Livestock and Dravidian Crop Vocabulary. In Examining the farming/language dispersal hypothesis, ed. P Bellwood, C Renfrew, Cambridge: The McDonald Institute for Archaeological Research

Fuller D. 2007. Non-Human Genetics, Agricultural Origins and Historical Linguistics in South Asia. In Vertebrate Paleobiology and Paleoanthropology, ed. M Petraglia, B Allchin, pp. 393-443. : Springer

Fuller DQ, Qin L, Zheng Y, Zhao Z, Chen X, Hosoya LA, Sun GP. 2009. The domestication process and domestication rate in rice: spikelet bases from the Lower Yangtze. *Science.* 323:1607-10.

Handley LJL, Manica A, Goudet J, Balloux F. 2007. Going the distance: human population genetics in a clinal world. *Trends Genet.* 23:432-39.

Higham C. 2003. Languages and Farming Dispersals: Austroasiatic Languages and Rice Cultivation. In Examining the farming/language dispersal hypothesis, ed. P Bellwood, C Renfrew, Cambridge: The McDonald Institute for Archaeological Research

Jin J, Huang W, Gao JP, Yang J, Shi M, Zhu MZ, Luo D, Lin HX. 2008. Genetic control of rice plant architecture under domestication. *Nat Genet.* 40:1365-69.

Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res*. 18:830-38.

Kayser M, Brauer S, Cordaux R, et al. (14 co-authors). 2006. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Mol Biol Evol*. 23:2234-44.

Kivisild T, Rootsi S, Metspalu M, Metspalu E, Parik J, Kaldama K, Usanga E, Mastana S, Papiha SS, Villems R. 2003. The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am J Hum Genet*. 72:313-32.

Kumar V, Reddy AN, Babu JP, Rao TN, Langstieh BT, Thangaraj K, Reddy AG, Singh L, Reddy BM. 2007. Y-chromosome evidence suggests a common paternal heritage of Austro-Asiatic populations. *BMC Evol Biol*. 7:47.

Kumar V, Langstieh BT, Madhavi KV, Naidu VM, Singh HP, Biswas S, Thangaraj K, Singh L, Reddy BM. 2006. Global patterns in human mitochondrial DNA and Y-chromosome variation caused by spatial instability of the local cultural processes. *PLoS Genet*. 2:e53.

Kumar V, Reddy ANS, Babu JP, Rao TN, Langstieh BT, Thangaraj K, Reddy AG, Singh L, Reddy BM. 2007. Y-chromosome evidence suggests a common paternal heritage of Austro-Asiatic populations. *BMC Evol Biol*. 7:47.

Lewis MP, ed. 2009. Ethnologue: Languages of the World, Online version: http://www.ethnologue.com/. Dallas, Tex: SIL International

Li JZ, Absher DM, Tang H, et al. (10 co-authors). 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*. 319:1100-04.

Metspalu M, Kivisild T, Metspalu E, et al. (16 co-authors). 2004. Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet*. 5:26.

Nei M. 1987. Molecular Evolutionary Genetics. New York: Columbia University Press

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet*. 2:e190.

Purcell S, Neale B, Todd-Brown K, et al. (11 co-authors). 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 81:559-75.

Purugganan MD, Fuller DQ. 2009. The nature of selection during plant domestication. *Nature*. 457:843-48.

Ramachandran S, Deshpande O, Roseman C, Rosenberg N, Feldman M, Cavalli-Sforza L. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A.* 102:15942.

Reddy BM, Langstieh BT, Kumar V, Nagaraja T, Reddy ANS, Meka A, Reddy AG, Thangaraj K, Singh L. 2007. Austro-Asiatic tribes of Northeast India provide hitherto missing genetic link between South and Southeast Asia. *PLoS ONE*. 2:e1141.

Richards M, Macaulay V, Hickey E, et al. (36 co-authors). 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet*. 67:1251-76.

Rootsi S, Zhivotovsky LA, Baldovic M, et al. 2007. A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe. *Eur J Hum Genet.* 15:204-11.

Sabeti PC, Varilly P, Fry B, et al. (99 co-authors). 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature*. 449:913-18.

Sahoo S, Singh A, Himabindu G, et al. (11 co-authors). 2006. A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. *Proc Natl Acad Sci U S A.* 103:843-48.

Sengupta S, Zhivotovsky LA, King R, et al. (14 co-authors). 2006. 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*. 78:202-21.

Shi H, Dong YL, Wen B, Xiao CJ, Underhill PA, Shen PD, Chakraborty R, Jin L, Su B. 2005. Y-chromosome evidence of southern origin of the East Asian-specific haplogroup O3-M122. *Am J Hum Genet*. 77:408-19.

Tan L, Li X, Liu F, et al. (10 co-authors). 2008. Control of a key transition from prostrate to erect growth in rice domestication. *Nat Genet*. 40:1360-64.

Tang H, Peng J, Wang P, Risch NJ. 2005. Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol*. 28:289-301.

Thangaraj K, Sridhar V, Kivisild T, et al. 2005. Different population histories of the Mundari- and Mon-Khmer-speaking Austro-Asiatic tribes inferred from the mtDNA 9-bp deletion/insertion polymorphism in Indian populations. *Hum Genet.* 116:507-17.

Thangaraj K, Chaubey G, Kivisild T, et al. (36 co-authors). 2008. Maternal footprints of Southeast Asians in North India. *Hum Hered.* 66:1-9.

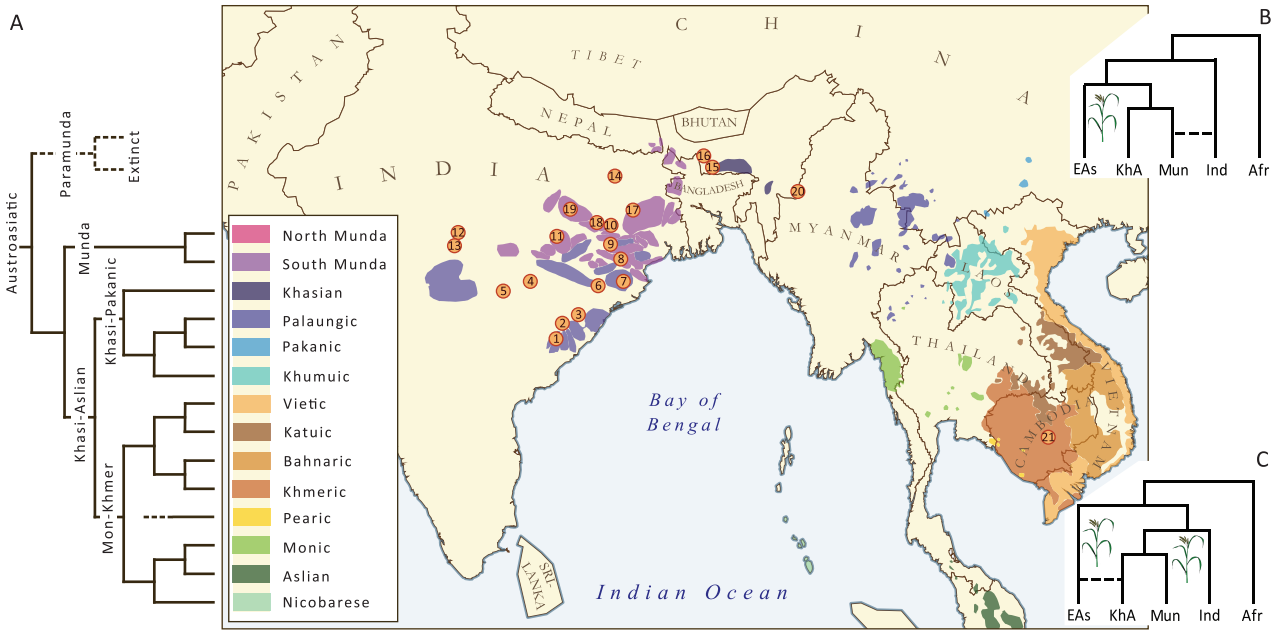
Thangaraj K, Nandan A, Sharma V, et al. (11 co-authors). 2009. Deep rooting in-situ expansion of mtDNA Haplogroup R8 in South Asia. *PloS one*. 4:e6545.

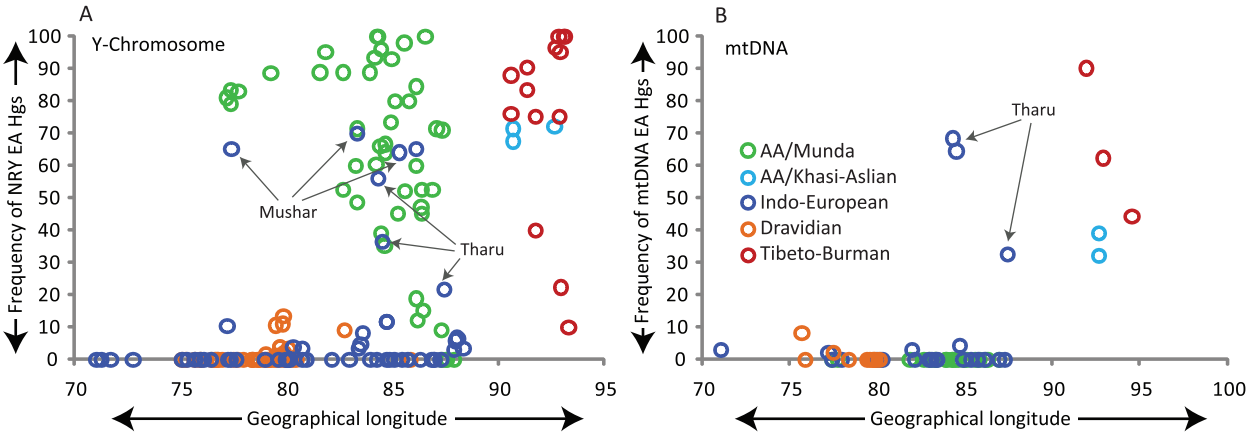
Weiss KM, Long JC. 2009. Non-Darwinian estimation: my ancestors, my genes' ancestors. *Genome Res*. 19:703-10.

Wright S. 1943. Isolation by Distance. *Genetics.* 28:114-38.

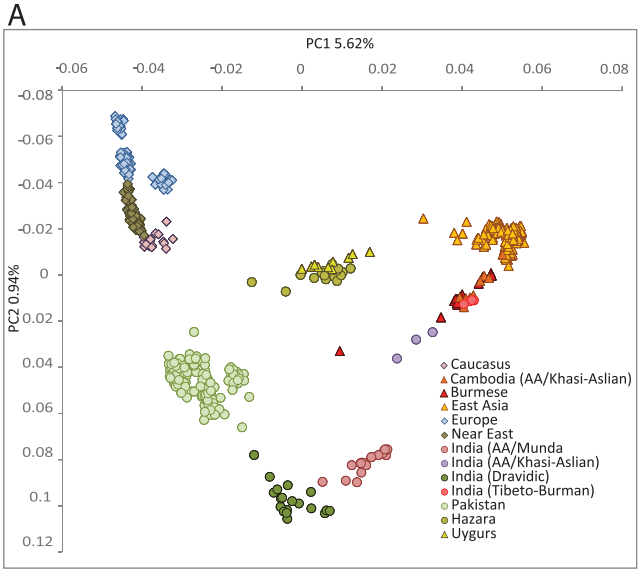
Xue Y, Zhang X, Huang N, et al. (14 co-authors). 2009. Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation. *Genetics.* 183:1065-77.

Zhivotovsky LA, Underhill PA, Cinnioğlu C, et al. (16 co-authors). 2004. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am J Hum Genet.* 74:50-61.

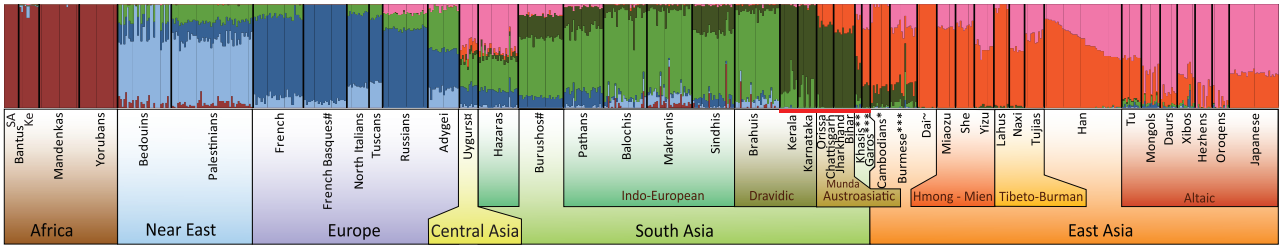

A
Austroasiatic
Paramunda
Extinct
Munda
Khasi-Pakanic
Khasi-Aslian
Mon-Khmer
North Munda
South Munda
Khasian
Palaungic
Pakanic
Khumuic
Vietic
Katuic
Bahnaric
Khmeric
Pearic
Monic
Aslian
Nicobarese
PAKISTAN
INDIA
CHINA
TIBET
NEPAL
BHUTAN
BANGLADESH
MYANMAR
LAOS
THAILAND
CAMBODIA
VIETNAM
Bay of Bengal
Indian Ocean
SRI LANKA
B
EAs KhA Mun Ind Afr
C
EAs KhA Mun Ind Afr


A
Y-Chromosome
Frequency of NRY EA Hgs
Mushar
Tharu
Geographical longitude
B
mtDNA
Frequency of mtDNA EA Hgs
Tharu
AA/Munda
AA/Khasi-Aslian
Indo-European
Dravidian
Tibeto-Burman
Geographical longitude

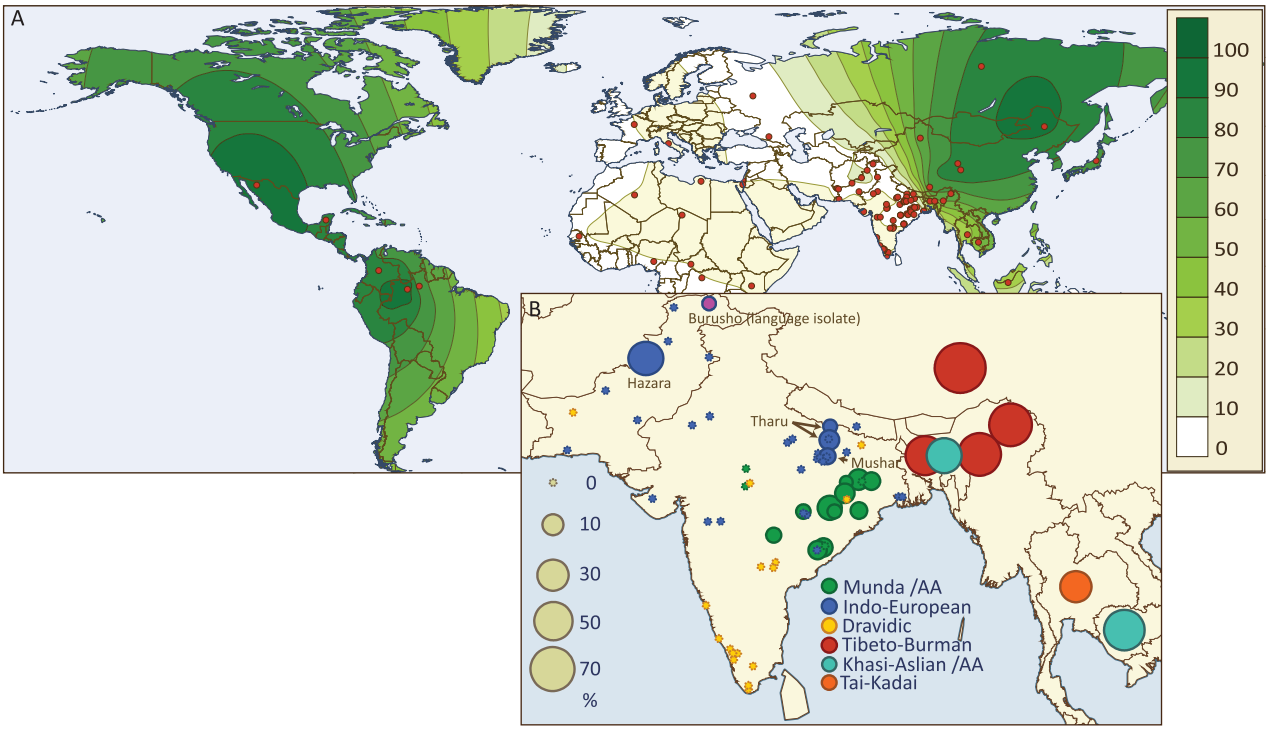
A

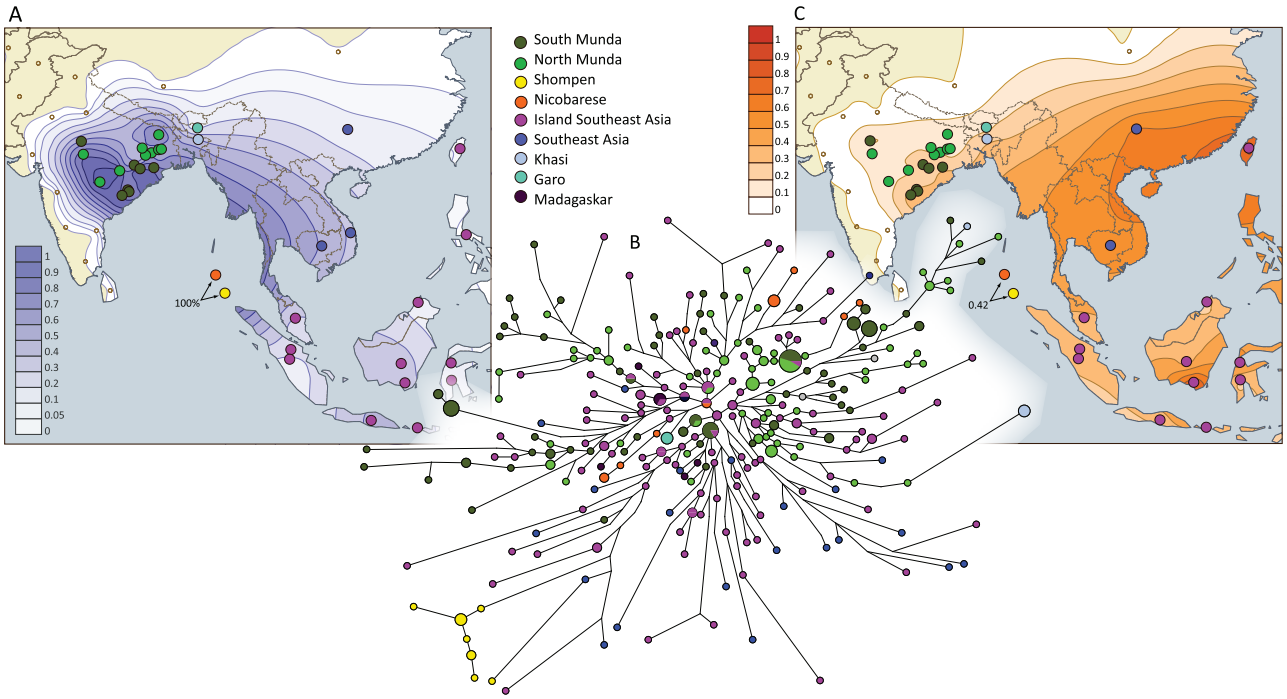


B



Language isolate ¤ Altaic * Mon-Khmer / Austroasiatic ** Khasi-Pakanic / Austroasiatic *** Tibeto-Burman ~ Tai - Kadai


A
100
90
80
70
60
50
40
30
20
10
0
B
Burusho (language isolate)
Hazara
Tharu
Mushar
0
10
30
50
70
%
Munda /AA
Indo-European
Dravidic
Tibeto-Burman
Khasi-Aslian /AA
Tai-Kadai


A
B
C
South Munda
North Munda
Shompen
Nicobarese
Island Southeast Asia
Southeast Asia
Khasi
Garo
Madagaskar
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.05
0
100%
0.42