Heliyon 10 (2024) e34354

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Investigating linguistic and genetic shifts in East Indian tribal groups

Bhavna Ahlawat[a,b,1], Hemlata Dewangan[c,1], Nagarjuna Pasupuleti[d,1], Aparna Dwivedi[a,e], Richa Rajpal[a,e], Saurabh Pandey[a], Lomous Kumar[a,*], Kumarasamy Thangaraj[d,**], Niraj Rai[a,e,***]

[a] *Birbal Sahni Institute of Palaeosciences, Lucknow, 226007, India*
[b] *Department of Anthropology, Panjab University, Chandigarh, 160014, India*
[c] *Shreyanshi Health Care Private Limited, Raipur, Chattisgarh, 492001, India*
[d] *CSIR—Centre for Cellular and Molecular Biology, Uppal Road, Hyderabad, 500007, India*
[e] *Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, 201002, India*

ARTICLE INFO

Keywords:
East India
Austro-Asiatic languages
Tribes
Indo-European languages
Mitochondrial DNA
Autosomal DNA

ABSTRACT

South Asia is home to almost a quarter of the world's total population and is home to significant ethnolinguistic diversity. Previous studies of linguistic and genetic affiliations of Indian populations suggest that the formation of these distinct groups was a protracted and complex phenomenon involving multiple waves of migration, cultural assimilation, and genetic admixture. The evolutionary processes of migration, mixing and merging of populations thus impact the culture and linguistic diversity of different groups, some of which may retain their linguistic affinities despite genetic admixture with other groups, or vice versa. Our study examines the relationship of genetic and linguistic affinities between Austroasiatic and Indo-European speakers in adjacent geographical regions of Eastern India. We analyzed 224 mitogenomes and 0.65 million SNP genotypes from 40 unrelated individuals belonging to the Bathudi, Bhumij, Ho, and Mahali ethnic groups from the Eastern Indian state of Odisha. These four groups are speakers of Austroasiatic languages who have adopted elements from Indo-European languages spoken in neighbouring regions. Our results suggest that these groups have the greatest maternal genetic affinity with other Austroasiatic-speaking groups in India. Allele frequency-based analyses, genome-wide SNPs, haplotype-based methods and IBD sharing further support the genetic similarity of these East Indian groups to Austroasiatic speakers of South Asia rather than regional populations speaking Indo-European and Dravidian languages. Our study shows that these populations experienced linguistic mixing, likely due to industrialization and modernization that brought them into close cultural contact with neighbouring Indo-European-speaking groups. However, linguistic change in these groups is not reflected in genetic mixing in these populations, as they appear to maintain strict genetic boundaries while simultaneously experiencing cultural mixing.

[*] Corresponding author. Birbal Sahni Institute of Palaeosciences, 53, University Road, Lucknow, Uttar Pradesh 226007, India.
[**] Corresponding author.
[***] Corresponding author. Birbal Sahni Institute of Palaeosciences, 53, University Road, Lucknow, Uttar Pradesh 226007, India.
E-mail addresses: lomousmishra@gmail.com (L. Kumar), thangs@ccmb.res.in (K. Thangaraj), nirajrai@bsip.res.in (N. Rai).
[1] Bhavna Ahlawat, Hemlata Dewangan and Nagarjuna Pasupuleti contributed equally.

https://doi.org/10.1016/j.heliyon.2024.e34354
Received 29 February 2024; Received in revised form 6 July 2024; Accepted 8 July 2024
Available online 9 July 2024

1. Introduction

India is a culturally and geographically highly heterogeneous country and has long been a melting pot of socio-cultural exchanges and ethnolinguistic diversity. This diversity stems from several demographic events, starting about 65 KYA with the first appearance of anatomically modern humans [1,2]. Indian populations are structured further by their linguistic and religious affiliations [3–6]. There are four major language families spoken across India: Indo-European, Dravidian, Austro-Asiatic, and Tibeto-Burman, out of which Austro-Asiatic speakers primarily reside in the central and Eastern parts of the country [2,3,7–10]. Studies have shown that a further linguistic sub-division of Austro-Asiatic speakers into North and South groups is consistent with genome-wide data, indicating sub-structure among speakers of this language family [11]. The origin and dispersal of these language families in South Asia is still an enigma [8]. The Austro-Asiatic language family includes about 150 different languages spoken in India, Vietnam, and the Malay Peninsula in the South. Mainland Southeast Asia (MSEA) has more than 100 million individuals who speak Austro-Asiatic languages, while more than 10 million speakers of the Austro-Asiatic Munda language live in East and Central India bounded by Indo-European, Dravidian and Trans-Himalayan languages speakers [7,12]. This huge geographical range of Austro-Asiatic speaking communities and the presence of the oldest India-specific mitochondrial DNA (mtDNA) haplogroup M2 among Austro-Asiatic speaking tribes of India have been used to propose that Austro-Asiatic speakers may have been the initial settlers of India [7].

The widespread distribution of Austro-Asiatic speakers is also speculated to be associated with the agricultural expansion in India and South Asia. Agricultural expansion has been speculated to be a result of either demographic or cultural spread in global contexts and has, hence, received attention from geneticists and paleo-geneticists. The expansion of Austro-Asiatic speakers can be traced back to rice cultivation in South Asia. Regardless of their current lifestyle, Munda-speaking and Khasi-Aslian–speaking hunter-gatherer populations of India share considerable rice cultivation allies with Khasi-Aslian–speaking populations of Southeast Asia [8]. Traditionally, it has been assumed that rice has a singular origin in China. However, genetic evidence suggests the multiple centres of domestication in Asia. Genetic evidence [13] shows independent domestication of *Oryza indica* and *Oryza japonica* cultivars, which further suggests that the homeland of the Austro-Asiatic family lies in India. Finally, in light of archaeobotanical, linguistic, and genetic evidence, the bifurcation of Austro-Asiatic languages into their major subgroups in South and Southeast Asia is reported to have happened around 7 KYA [7]. These findings suggest a migratory route of Austro-Asiatic languages from India to Southeast Asia, and not the other way around [14].

Genetic studies conducted so far have proven to be inconclusive about the geographical origin of Austro-Asiatic speakers and the timing of the split into subgroups in this language family. mtDNA data from Indian Munda and Southeast Asian Khasi–Aslian–speaking groups indicate a clear distinction in their maternal ancestries, with both sharing more mtDNA lineages with their respective regional neighbours, who speak languages other than Austro-Asiatic [7]. The Mundari-speaking group reflects a dominant presence of Indian-specific mtDNA haplogroups that are also reported in speakers of Dravidian and Indo-European languages [9,15–18], whereas the Mon-Khmer speakers have more East Asian-specific mtDNA haplogroups [7]. The presence of different haplogroups among Mundari and Mon-Khmer speakers points towards independent migration and origin of both these groups, these findings have also been confirmed by the independent occurrence of a 9 bp deletion polymorphism in them [4,16,19]. Overall, Indian and Southeast Asian Austro-Asiatic speakers can be distinguished based on differences in their respective mtDNA haplogroup distributions. Y chromosomes present the strongest signals of shared Southeast Asian genetic ancestry among the Indian Austro-Asiatic speakers, since approximately two-thirds of the group's population falls under Y haplogroup O2a-M95. This finding suggests that the migration of Austro-Asiatic speakers to South Asia was male mediated [7,15,20].

While the Y chromosomal analysis mainly suggest the common origin of different Austro-Asiatic groups, mtDNA studies have revealed different demographic histories for maternal lineages of the Mundari and Mon-Khmer groups [21]. The geographical pattern of mtDNA haplogroup is consistent with its origin in South Asia dating back approximately 20 KYA [7]. Studying the time depth of Y STR (short tandem repeat) variations of haplogroup O2a, it can be observed that the entry of O2a in Southeast Asia is quite recent <10 KYA, and all other branches of haplogroup O are majorly restricted to East Asia [20,22]. On the contrary, the genetic diversity of mtDNA haplogroups and Y STR O2a haplogroup diversity in Austroasiatic speakers favour the direct descent model of Austroasiatic speakers from the initial settlers of India ~65 KYA [15,23]. What must be taken into account is that this 65 KYA dating of haplogroup O2a appears much older than the ancestral haplogroups K and NO [24], and this contradiction needs to be resolved.

Some Austro-Asiatic speaking communities in India have also undergone a language shift, a cultural process which involves changes in the spoken language [25,26]. For example, a large proportion of Mundari-speaking Austro-Asiatic tribes have adopted Indo-European languages as a mode of communication in the recent past [25], likely facilitated by exchange of ideas, goods and food. However, these cases of cultural changes have been shown to be not accompanied by genetic shifts, indicating impregnable genetic boundaries in these groups [7,8,18]. Genetic studies conducted till date also lack high-resolution autosomal evidence and, hence, the genetic origins of Austro-Asiatic-speaking populations remain largely controversial [7]. Here, we present whole mitochondrial genomes from blood samples collected from 224 unrelated individuals from East Indian states, the collected samples represent four Austroasiatic groups including Bathudi (N = 58), Bhumij (N = 58), Ho (55), and Mahali (N = 53). Additionally, we genotyped 40 individuals (10 individuals from each group) for genome-wide autosomal SNP markers. The primary objective of this study is to investigate the maternal phylogeny and the autosomal SNPs based genetic affinities of these four groups within the context of the genetic affinities that characterize other linguistically-diverse South Asian groups.

2. Materials and methods

2.1. Sampling

Blood samples were collected from 224 unrelated individuals inhabited in East Indian state of Odisha (Fig. 2b), which include Bathudi (N = 58), Bhumij (N = 58), Ho (55) and Mahali (N = 53) tribes. Informed written consent was obtained from the participants.

2.2. DNA extraction and genotyping

DNA was extracted using a phenol chloroform method. Complete mitochondrial genome was amplified using 24 sets of primers designed to cover complete mitochondrial genome [27](Supplementary Table S2). Amplified PCR fragments were sequenced using Sanger sequencing technique (ABI 3130XL Genetic Analyzer, Applied Biosystems, USA). Mutations were identified by aligning the mtDNA sequences against the revised Cambridge Reference Sequence (rCRS) [28] using the AutoAssembler [29] tool. mtDNA haplogroups were assigned based on Phylotree 17 FU1 [30] using the Haplogrep 2.4.0 software [31]. Details of mtDNA haplogroups determined for each population are listed in the supplementary file (Supplementary Table 1). Genome-wide SNP genotype data was generated for 40 individuals from the four tribal populations (Bathudi = 10, Bhumij = 10, Ho = 10 and Mahali = 10) using the Illumina GSA genotyping panel. The genotyping was performed according to manufacturer's protocol for Infinium Global screening Array-24 v-1.0 from Illumina, Inc, CA, USA.

2.3. Population diversity statistics and Maximum-Likelihood phylogeny

DnaSP6 [32] software was used to calculate nucleotide diversity, haplotype diversity, Fu and Li statistics, and Tajima's D statistics. A maximum likelihood tree was constructed using Mega 11 [33] using merged fasta sequences of mitochondrial haplogroups M2 and M5. Combined fasta sequences (including Bathudi, Bhumij, Ho, Mahali, published mtDNA sequences from Phylotree from India, 1000 Genomes and HGDP) were first aligned using MUSCLE algorithm in MEGA 11 [33]. Gaps and redundant sites (309.1, 315.1, 515–522, 16182, 16183, 16193.1 and 16519) were excluded in further analyses. Maximum Parsimony (MP) tree was used as the initial tree and 500 bootstrap replicates were done and a strong branch swap filter was applied.

2.4. Bayesian evolutionary analysis and haplotype networks

Bayesian evolutionary analysis was performed on the combined mitochondrial genome fasta using BEAST v2.7.4 [34]. For constructing the phylogenetic tree, we aligned the merged fasta sequences of the four tribal populations with reference sequences from published Indian, 1000 Genomes and HGDP populations, affiliated under the clades M2 and M5, with haplogroup L0a1 from a Bantu individual used as an outgroup. Sequences were aligned using MUSCLE [35] alignment method implemented in MEGA 11 [33]. Best partitioning scheme was determined using Bayesian Information Criterion (BIC) implemented in PartitionFinder v2.1.1 [36] and using the greedy algorithm. Best fit nucleotide substitution model was inferred using jModelTest v2.1.10 [37] for each of the partitions (HVS and coding regions) of mitogenome. BEAST [38] runs were performed using linked tree model and a strict molecular clock was chosen. As tree prior Coalescent Constant population was used for all partitions. Three individual Markov Chain Monte Carlo (MCMC) chain were run with 50,000,000 steps with every 10,000 steps used for parameter sampling and 10 % steps were discarded as burn-in. Convergence of MCMC chains and Effective Sample Size (ESS >200) were evaluated by visualizing in Tracer v1.7 [39].

All the Log files and Tree files of three MCMC runs were combined using LogCombiner tool of BEAST v2.7.4 [34] package. Consensus tree was generated from combined tree file using TreeAnnotator [34] by discarding 10 % samples as burn-in and using median heights. Consensus tree was visualized in FigTree v1.4.4. Median Joining networks for mitochondrial haplogroups M2 and M5 of four tribal groups were constructed using PopART [40] tool in R.

2.5. Extended Bayesian Skyline Plot

For the estimation of population size (Ne) history through time, individual aligned fasta sequences of Bathudi, Bhumij, Ho and Mahali populations were used in BEAST v2.7.4 [34]. PartitionFinder [36] was used to generate best partitioning scheme of each population aligned fasta files. Nucleotide Substitution model was derived for each partition using jModelTest [37]. Coalescent Extended Bayesian Skyline was used as tree prior and 50,000,000 iterations were run, discarding 10 % as burn-in. Three individual runs were performed at random seed numbers and log files from individual runs were combined in Logcombiner [34]. Extended Bayesian Skyline Plots were done using R script plotEBSP.R.

2.6. Population genetic analyses on autosomal SNP data

Genotype data was merged with data from the Simons Genome Diversity Panel as well as modern genotype data from our lab also generated on the Illumina platform. Quality filtering was performed in Plink v1.9 [41] with quality cutoff values for geno = 0.03, mind = 0.05 and maf = 0.01. Kinship based filtering was done by Plink2 [41] King-cutoff function and individuals up to second degree of relationship were excluded (one individual from each pair). For PCA and ADMIXTURE, the merged SNP dataset was pruned for linkage disequilibrium (LD pruning with 200 25 0.04 setting).

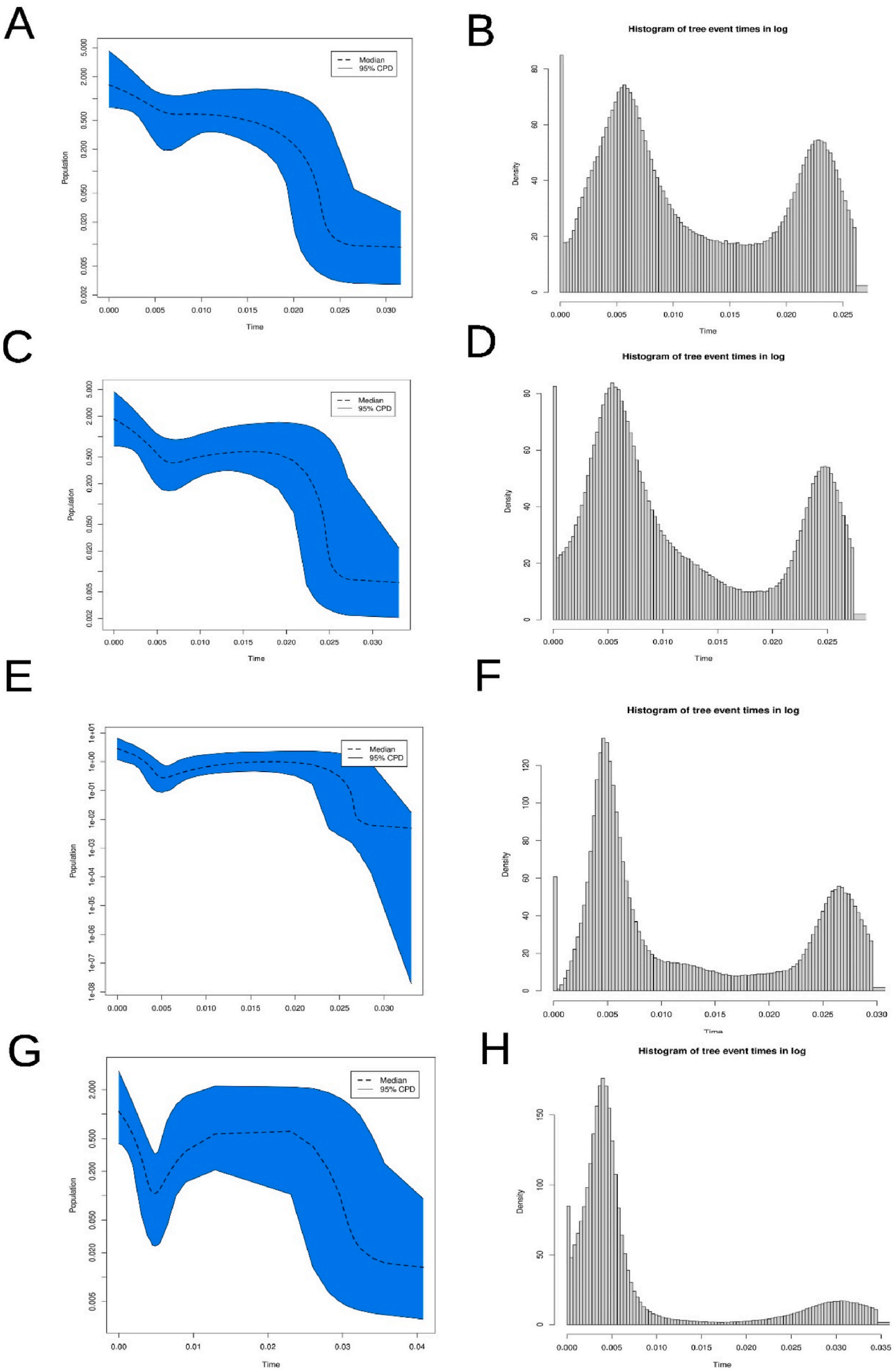

Fig. 1. Historical maternal effective population size history and corresponding histograms of tree event for Ho population **A. & B.** for Bathudi population **C. & D.** for Bhumij **E. & F.** and for Mahali **G & H.**

PCA was run on the LD pruned genome-wide SNP data using *smartpca* tool in EIGENSOFT [42] package. Model based unsupervised clustering was performed using ADMIXTURE [43]. Outgroup-f3 statistics was performed using Admixtools [44]. Weir and Cockerham Fst was calculated in R package *assigner* (Gosselin et al., 2020). All the results were plotted in R [45].

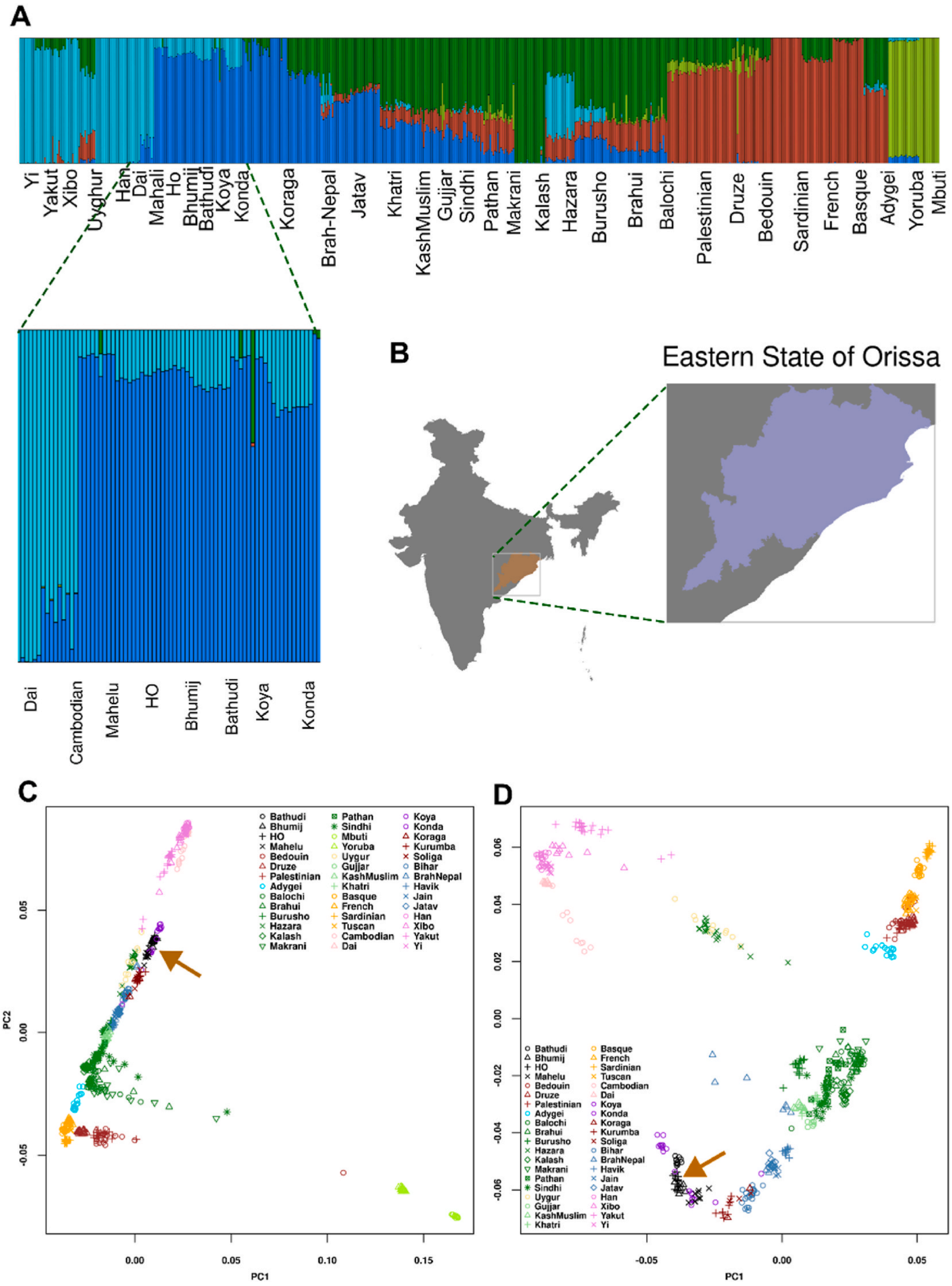

Fig. 2. A. Admixture barplot with modern references, distinct colors represent putative ancestral sources and populations are arranged from right to left, upper panel is continental or linguistic groups (AFR: Africa, NWI: Northwest India, IEU: India Indo-European, DRA: Dravidian, AAS: Austro-asiatic); **B.** Sample location map, inset show the eastern state of Odisha in India; **C.** PCA biplot using first two principal components with African populations and D. without African populations.

We further used haplotype-based approaches to infer more recent and fine-scale haplotype sharing by implementing ChromoPainter [46] and fineSTRUCTURE [46]. We first phased our data using 1000 genome reference panel with SHAPEIT4 [47] with default parameters. ChromoPainter [46] was first run on phased data by performing 10 Expectation-Maximization (EM) iteration with 5 randomly selected chromosomes on a subset of individuals to infer the global mutation rate (μ) and switch rate parameters (Ne). The main ChromoPainter algorithm was run with all chromosomes and all the individuals to get co-ancestry chunks. This co-ancestry chunk counts matrix was fed into fineSTRUCTURE [46] to infer the tree using a probability model by applying Markov chain Monte Carlo (MCMC). For the run, we used 500,000 burn-in iterations and 1,000,000 subsequent iterations and stored the results from every 10,000th iteration. IBD sharing pattern was inferred using the phased data and Refined IBD [48] method. All the plots were created using R statistical packages [45].

3. Results

3.1. Mitochondrial haplogroup distribution

Major mitochondrial haplogroups observed among the four tribal populations (Bathudi, Bhumij, Ho and Mahali) are listed in the supplementary file (Supplementary Table 1). There was high genetic heterogeneity observed in the haplogroup distribution among the four groups. Ho and Mahali have the highest frequency of mtDNA haplogroup M40 (0.15 and 0.17, respectively), while Bathudi has the highest prevalence of haplogroup M2a (0.16) and Bhumij has M52 (0.14) as most frequent mtDNA haplogroup. Subgroups of M2 and M5 haplogroups were common among all four tribal groups, which are among the most ancient mitochondrial lineages in Indian subcontinent [17,49]. Further downstream analyses to infer evolutionary analysis were mainly focused on lineages of mtDNA haplogroups M2 and M5.

3.2. Maximum likelihood tree of mtDNA haplogroups M2 and M5

Maximum-Likelihood tree in MEGA [33] was constructed using combined fasta sequences of haplogroups M2 and M5 affiliated samples with 500 bootstrap replicates. Placement of the four tribal groups in the ML tree of mitochondrial clade M2 was highly heterogenous. ML tree was divided into three major clades and an outgroup (Bantu13) (Supplementary Fig. S1). Most of the Bhumij and Bathudi individuals clustered together with Austro-Asiatic and Dravidian speaking individuals, while the Katkari (an Indo-European speaking tribe) form a separate sub-cluster within this major clade. All M2 Mahali individuals clustered in a heterogenous clade that primarily consisted of Austro-Asiatic speaking individuals (Korku, Munda, Malpahariya), along with a few 1000 Genomes Indo-European and Dravidian speaking individuals.

In the ML tree of haplogroup M5, most of the Bhumij, Bathudi, Ho and Mahali individuals were present in a single major clade with a minor presence of Indo-European, Dravidian and Tibeto-Burman speaking individuals (Supplementary Fig. S2). Some of the other Austro-Asiatic speaking groups in this clade were Munda, Korku, Malpaharia, and Paudi Bhuiya. Two of the Bhumij (Bhumij53 and Bhumij54) and one Bathudi (BTD08) individuals were placed in a separate clade, shared with Indo-European and Dravidian speaking individuals. One of the Ho individuals (HO31; HG M5a) was an outlier with respect to all major clusters.

3.3. Extended Bayesian Skyline Plot

Extended Bayesian Skyline Plot (EBSP) analysis was done for the four tribal groups to trace population size (Ne) change history through time. Best portioning scheme divided the mitochondrial genome into two compartments (HVS and Coding region). Best nucleotide substitution model for HVS region was HKY + I + G and for coding region TrN + I + G. A maternal population bottleneck was observed in case of Mahali, whereas we observed a maternal population size expansion for Bhumij, Bathudi, and Ho (Fig. 1 a-h).

3.4. Median-joining networks of mtDNA haplogroups M2 and M5

Median joining network of mtDNA haplogroup M2 differentiated into two nodes defining the major sub clades M2a and M2b (Supplementary Figs. S3a–b). Sub clade M2a was having uniform sharing of haplotypes among the three major linguistic groups (Austroasiatic, Dravidian and Indo-European), but diversity was high among Austroasiatic groups indicating their highly divergent (founder) lineage in terms of M2a. On the other hand, sub haplogroup M2b had Dravidian founders and among four East Indian tribes, Mahali shared haplotypes in this cluster both with Austroasiatic and Dravidian groups (Supplementary Figs. S3a–b). This clearly indicates mainly derived ancestry in Mahali related to this sub haplogroup (M2b).

A similar apparent differentiation was observed in case of haplotype sharing patterns among sub haplogroups of mtDNA haplogroup M5 (Supplementary Figs. S4a–b). The sub clade M5a is shared among haplotypes from Pakistan, Indian Indo-Europeans, linguistic isolate Nihali, Austroasiatic groups and three of the East Indian tribes (Bathudi, Bhumij and Mahali). Whereas, remaining three subclades namely, M5b, M5c and M5d had haplotypes mainly from Indian Austroasiatic speakers, Bathudi, Bhumij and Mahali (Supplementary Figs. S4a–b). All four sub clades are highly diversified among Austroasiatic speakers, clearly indicating their founder lineage in Austroasiatic speakers from India.

3.5. Population diversity statistics

Population diversity statistics were calculated for the four tribal groups using DnaSP6. The number of mitochondrial haplotypes for Bathudi, Bhumij, Ho, and Mahali were 57, 55, 55 and 50, respectively, while haplotype diversity was 0.9994, 0.9975, 1, and 0.9978, respectively. Nucleotide diversity (Pi) value was highest for Bhumij (0.00184) and lowest in case of Mahali (0.00164). The value of Fu and Li D test statistics for Bathudi, Bhumij, Ho, and Mahali were −2.46605, −3.28539, −2.25893 and −1.06731, respectively. Tajima's D statistics values were −2.02970, −2.06801, −2.08711 and −1.25634 for Bathudi, Bhumij, Ho, and Mahali, respectively.

3.6. PCA and admixture analysis on autosomal SNPs

Genome-wide SNP data from the four tribal populations (Bathudi, Bhumij, Ho and Mahali) were merged with modern Eurasian reference populations from published sources and unpublished genotype dataset from our lab. We have analyzed by including African populations (Fig. 2c) and also by excluding the Africans (Fig. 2d). Principal Component Analysis (PCA) on LD pruned data placed individuals from all four tribal populations in a separate cluster along with other Austro-Asiatic speaking groups in the main South Asian cline (starting with Pakistan groups in forest green and ending with Austro-Asiatic speaking groups in khaki) (Fig. 2c&d). Black dots shown by arrows in this cluster represents the four tribal groups sequenced in this study (Bathudi with filled circle, Bhumij with filled diamond, Ho with filled square and Mahali with circled plus). All other Eurasian populations including modern Indians are arranged in a single Eurasian cline extending from the Caucasus, Europe and the Middle East to East Asians and Southeast Asians.

In the model based unsupervised clustering, all four tribal groups maximized the blue component (corresponding to a South Asian ancestral component), with traces of the cyan component (maximized in Southeast Asians and East Asians). Mahali differed from Bathudi, Bhumij and Ho due to the substantially smaller proportion of the cyan component in the former (Fig. 2a). Two Bhumij and one Mahali Individuals additionally showed the forest green component (Indo-European specific).

In the clustered heatmap of pairwise wcFst, all the four tribal groups (Bathudi, Bhumij, Ho, and Mahali) are clustered together and share a clade with Southeast Asian and East Asian groups (Supplementary Fig. 5). Cladogram based on Nei's genetic distance also placed the four tribal groups together in a clade not shared by any other group, but Ho and Bhumij were additionally placed together in a subclade within this larger clade (Supplementary Fig. 6).

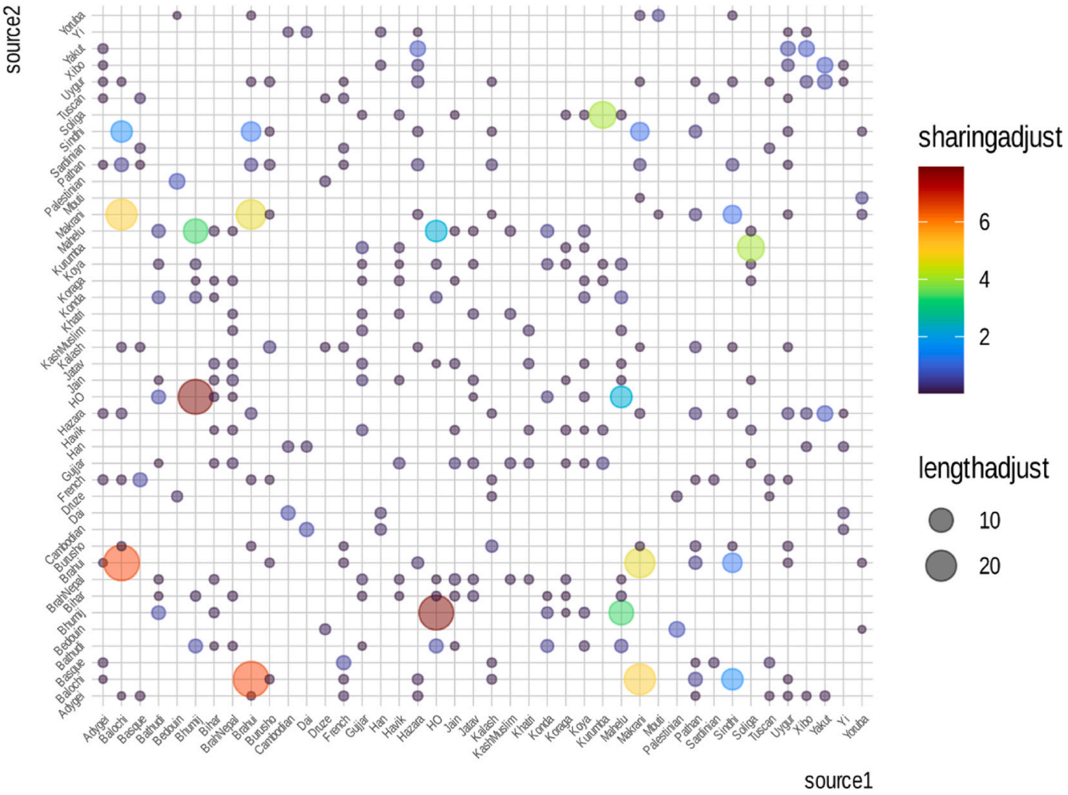

Fig. 3. IBD sharing matrix generated by adjusting for sample numbers. X-axis represent source1 and Y-axis represents source2 populations for IBD sharing. Colour scale represents proportion of IBD sharing while size of the circle represents length of IBD sharing (both adjusted for sample numbers).

4. Haplotype based analysis and IBD sharing

The FineStructure dendrogram divided all populations into three major clades, one including the African outgroups (Mbuti and Yoruba), second including European, Middle Eastern, Pakistani, and 1 KG South Asian populations, and third including East Asian, South East Asian, admixed Pakistani (Hazara), and Indian populations. In this third major clade, the four study groups (Bathudi, Bhumij, Ho, and Mahali) clustered together in a sub-clade distinct from other linguistic groups from India and with two other Austro-Asiatic speaking groups, Koya and Konda (Supplementary Fig. 7). Only a single Bhumij Individual was observed to cluster with Indian Dravidian and Indo-European speaking cluster, which may reflect recent gene flow between Bhumij and the latter populations. In the phylogenetic tree based on Nei's genetic distance, all four study groups again formed a separate clade with the Ho and Bhumij sharing a sub-clade with each other.

In the IBD sharing matrix normalized for sample size, Bhumij and Ho were observed to have high IBD sharing, while Mahali only shared moderately with Ho and Bhumij. Bathudi were observed to not have significant IBD sharing with Ho, Bhumij, and Mahali (Fig. 3).

5. Discussion

The general consensus among geneticists is that Austroasiatic speaking people were among the first settlers in India; today there are around 100 million people in India and Southeast Asia [7,11]. It has been suggested that the expansion of the Austroasiatic speaking population was related to agriculture and rice cultivation in South Asia [13,50], but this hypothesis has been constantly disputed. There are mainly three groups of Austroasiatic speakers: Mundari, Mon-Khmer and Khasi [51]. Overall, these speakers are distributed across central India in Chotanagpur and Odisha, Eastern India and the Meghalaya regions of India. Recently, these communities have experienced drastic cultural and linguistic changes [25,26]. Many Austroasiatic speaking groups have lost their traditional hunter-gatherer lifestyle and shifted to a more urban way of life characterized by interaction with the surrounding Indo-European-speaking population [25,26]. Indo-European languages have significantly expanded the strict socio-cultural boundaries that previously existed in the Indian subcontinent [52,53]. Although the genetic effects of cultural integration of one of the Austroasiatic speakers (Khasi) in Northeast India was investigated [54], these changes have mostly socioeconomic impacts on other Austroasiatic speakers in Eastern India [25]. It has been observed that despite this language adaptation by communities, the genetic makeup of other Austrian speakers remains largely unchanged [7,14,55].

We extensively investigated the plausible impact of linguistic and cultural changes among East Indian tribes (Ho, Bathudi, Bhumij and Mahali/Mahelu) on their genetic make-up using mitochondrial and autosomal SNP markers. As observed in their mtDNA haplogroup distribution, the maternal genetic composition of these four tribes is primarily indigenous and deeply rooted. Higher prevalence and divergent sub lineages of mtDNA haplogroups M2 (oldest subgroup of M) and M5 in Bathudi, Bhumij and Ho indicate their deeper maternal lineage in the Indian subcontinent. As shown by the maximum likelihood tree and phylogenetic networks of haplogroups M2 and M5, they share clades and haplotypes with almost all language groups in India. Furthermore, unlike Dravidian speakers from South India who have evidence of West Eurasian influence on their maternal genome [56,57], which may be as a consequence of trade connections [58]or multiple migrations [59], East Indian tribes did not show a major presence of West Eurasian maternal lineages. This further confirms the lesser influence of cultural and linguistic changes on their maternal genetic composition. However, the minor effects of recent demographic changes due to urbanization and industrialization are well imprinted in the maternal genome of the Mahali group, as evidenced by the historical maternal bottleneck (Fig. 1g).

We further examined the autosomal genome of East Indian tribes and found a good concordance between the indigenous linguistic affiliation and their genetics in our frequency and haplotype-based analysis. In PCA using autosomal markers, all four groups cluster in separate sub-clusters along with other Austroasiatic speaking groups in the main South Asian lineage. However, the haplotype-based approach provided clear evidence that two of these four groups, Ho and Bhumij, shared a common genetic history. This is further reinforced by the average number of shared IBD patterns, which shows a greater number of shared segments between Ho and Bhumij compared to the other two groups. Mahali showed only moderate commonality with Ho and Bhumij, reflecting recent gene flow from Ho or Bhumij (Fig. 3).

In summary, our study shows a close genetic relationship between the studied Austroasiatic groups despite the geographical distance and linguistic transitions. Our analyses of genetic data from mitochondrial and autosomal markers provide a range of estimates of gene flow across geographical and linguistic boundaries. However, more finer details and inferences can be drawn by covering more individuals from the populations covered, as some of the methods such as haplotype-based analysis will then clear picture of recent gene flow pattern. Still, with the current evidences we argue that language shifts in the extant East Indian tribes of Odisha are a consequence of cultural adaptation to the ever-changing environment due to urbanization and industrialization and are not due to larger demographic or genetic integrations. This study expands our understanding of the genetic history and admixture history of Austroasiatic populations in South Asia and may provide insights into the dynamics between genes and languages in these populations.

Data availability statement

Data is publicly available with restricted access at following link: https://zenodo.org/records/11071002. An email can be sent to get access to the data.

Ethics approval

The study was approved by Institute Ethical Committee of Birbal Sahni Institute of Palaeosciences, Lucknow, India under institutional ethical no. BSIP/Ethical/2021. All the procedure has been followed according to the recommendations of the Helsinki Declaration.

Informed consent

Informed written consent was obtained from all the participants involved in the study.

CRediT authorship contribution statement

Bhavna Ahlawat: Writing – original draft, Formal analysis, Data curation. **Hemlata Dewangan:** Writing – original draft, Data curation. **Nagarjuna Pasupuleti:** Writing – original draft, Formal analysis, Data curation. **Aparna Dwivedi:** Writing – original draft. **Richa Rajpal:** Writing – original draft. **Saurabh Pandey:** Writing – original draft, Data curation. **Lomous Kumar:** Writing – review & editing, Validation, Supervision, Software, Methodology, Formal analysis, Conceptualization. **Kumarasamy Thangaraj:** Validation, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Niraj Rai:** Writing – review & editing, Validation, Supervision, Project administration, Investigation, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank all the study participants who volunteered in this study. NR thanks the Director, Birbal Sahni Institute of Palaeosciences for scientific support and the Department of Science and Technology for CRG grant (CRG/2021/006762) from SERB-DST, India. KT was supported by the JC Bose Fellowship (JCB/2019/000027), SERB, Department of Science and Technology, India and Council of Scientific and Industrial Research, Ministry of Science and Technology, Government of India.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e34354.

References

[1] H. McColl, et al., The prehistoric peopling of Southeast Asia, Science 361 (6397) (2018) 88–92.

[2] K. Thangaraj, et al., Reconstructing the origin of andaman islanders, Science 308 (5724) (2005) 996.

[3] K. Thangaraj, et al., Maternal footprints of Southeast Asians in North India, Hum. Hered. 66 (1) (2008) 1–9.

[4] B.M. Reddy, et al., Austro-Asiatic tribes of Northeast India provide hitherto missing genetic link between South and Southeast Asia 2 (11) (2007) e1141.

[5] D. Reich, et al., Reconstructing Indian population history, Nature 461 (7263) (2009) 489–494.

[6] R. Roberts, et al., Population Increase and Environmental Deterioration Correspond with Microlithic Innovations in South Asia Ca, in: 000 Years Ago, vol. 35, 2009.

[7] G. Chaubey, et al., Population genetic structure in Indian Austroasiatic speakers: the role of landscape barriers and sex-specific admixture, Mol. Biol. Evol. 28 (2) (2011) 1013–1024.

[8] G. Chaubey, et al., Language shift by indigenous population: a model genetic study in South Asia 8 (1–2) (2008) 41–50.

[9] G. Chaubey, et al., Peopling of South Asia: investigating the caste–tribe continuum in India 29 (1) (2007) 91–100.

[10] K. Thangaraj, et al., In situ origin of deep rooting lineages of mitochondrial Macrohaplogroup’M’in India 7 (1) (2006) 1–6.

[11] K. Tatte, et al., The genetic legacy of continental scale admixture in Indian Austroasiatic speakers, Sci. Rep. 9 (1) (2019) 3818.

[12] K. Tätte, et al., The genetic legacy of continental scale admixture in Indian Austroasiatic speakers 9 (1) (2019) 3818.

[13] D.Q. Fuller, Contrasting patterns in crop domestication and domestication rates: recent archaeobotanical insights from the Old World, Ann. Bot. 100 (5) (2007) 903–924.

[14] G. Chaubey, The Demographic History of India: A Perspective Based on Genetic Evidence, 2010.

[15] A. Basu, et al., Ethnic India: a genomic view, with special reference to peopling and structure, Genome Res. 13 (10) (2003) 2277–2290.

[16] G. Chaubey, et al., Phylogeography of mtDNA haplogroup R7 in the Indian peninsula, BMC Evol. Biol. 8 (1) (2008) 227.

[17] M. Metspalu, et al., Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans 5 (1) (2004) 1–25.

[18] K. Thangaraj, et al., The influence of natural barriers in shaping the genetic structure of Maharashtra populations, PLoS One 5 (12) (2010) e15283.

[19] A. Chandrasekar, et al., Updating phylogeny of mitochondrial DNA macrohaplogroup m in India: dispersal of modern human in South Asian corridor, PLoS One 4 (10) (2009) e7447.

[20] S. Sahoo, et al., A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios, Proc. Natl. Acad. Sci. U.S.A. 103 (4) (2006) 843–848.

[21] T. Kivisild, et al., The Genetics of Language and Farming Spread in India, 2003, pp. 215–222.

[22] S. Sengupta, et al., Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists 78 (2) (2006) 202–221.

[23] V. Kumar, et al., Y-chromosome evidence suggests a common paternal heritage of Austro-Asiatic populations, BMC Evol. Biol. 7 (1) (2007) 47.

[24] S. Rootsi, et al., A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe, Eur. J. Hum. Genet. 15 (2) (2007) 204–211.

[25] M. Ishtiaq, Determinants and correlates of language shift among the tribals of Central India, Geojournal 45 (3) (1998) 189–200.

[26] M. Ishtiaq, Language shifts among the scheduled tribes in India: a geographical study, in: MLBD Series in Linguistics, Motilal Banarsidass Publishers, 1999.

[27] M.J. Rieder, et al., Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome, Nucleic Acids Res. 26 (4) (1998) 967–973.

[28] R.M. Andrews, et al., Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA, Nat. Genet. 23 (2) (1999) 147.

[29] S.R. Parker, AutoAssembler sequence assembly software, in: S.R. Swindell (Ed.), Sequence Data Analysis Guidebook, Springer New York, Totowa, NJ, 1997, pp. 107–117.

[30] A. Dür, N. Huber, W. Parson, Fine-tuning phylogenetic alignment and haplogrouping of mtDNA sequences, J.I.J.o.M.S. 22 (11) (2021) 5747.

[31] H. Weissensteiner, et al., HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing, Nucleic Acids Res. 44 (W1) (2016) W58–W63.

[32] J. Rozas, et al., DnaSP 6: DNA sequence polymorphism analysis of large data sets, Mol. Biol. Evol. 34 (12) (2017) 3299–3302.

[33] K. Tamura, G. Stecher, S. Kumar, MEGA11: molecular evolutionary genetics analysis version 11, Mol. Biol. Evol. 38 (7) (2021) 3022–3027.

[34] R. Bouckaert, et al., Beast 2.5: an advanced software platform for Bayesian evolutionary analysis, PLoS Comput. Biol. 15 (4) (2019) e1006650.

[35] R.C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinf. 5 (1) (2004) 113.

[36] R. Lanfear, et al., PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses, Mol. Biol. Evol. 34 (3) (2017) 772–773.

[37] D. Darriba, et al., jModelTest 2: more models, new heuristics and parallel computing, Nat. Methods 9 (8) (2012) 772.

[38] A.J. Drummond, A. Rambaut, BEAST: Bayesian evolutionary analysis by sampling trees, BMC Evol. Biol. 7 (1) (2007) 214.

[39] A. Rambaut, et al., Posterior summarization in bayesian phylogenetics using tracer 1.7, Syst. Biol. 67 (5) (2018) 901–904.

[40] J.W. Leigh, D. Bryant, S. Nakagawa, popart: full-feature software for haplotype network construction, Methods Ecol. Evol. 6 (9) (2015) 1110–1116.

[41] C.C. Chang, et al., Second-generation PLINK: rising to the challenge of larger and richer datasets, GigaScience 4 (2015) 7, 2047-217X (Electronic).

[42] N. Patterson, A.L. Price, D. Reich, Population structure and eigenanalysis, PLoS Genet. 2 (12) (2006) e190.

[43] D.H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals, Genome Res. 19 (9) (2009) 1655–1664.

[44] N. Patterson, et al., Ancient admixture in human history, Genetics 192 (3) (2012) 1065–1093.

[45] R.C. Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2021.

[46] D.J. Lawson, et al., Inference of population structure using dense haplotype data, PLoS Genet. 8 (1) (2012) e1002453.

[47] O. Delaneau, et al., Accurate, scalable and integrative haplotype estimation, Nat. Commun. 10 (1) (2019) 5436.

[48] B.L. Browning, S.R. Browning, Improving the accuracy and efficiency of identity-by-descent detection in population data, Genetics 194 (2) (2013) 459–471.

[49] C. Sun, et al., The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes, Mol. Biol. Evol. 23 (3) (2006) 683–690.

[50] P. Bellwood, Examining the farming language dispersal hypothesis in the East Asian context, in: The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics, Routledge Curzon, 2005, pp. 17–30.

[51] G. Diffloth, The peopling of east asia: putting together archaeology, The contribution of linguistic paleontology to the homeland of Austro-asiaticlinguistics Genet 1 (2005) 79–82.

[52] A. Basu, N. Sarkar-Roy, P.P. Majumder, Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure, Proc. Natl. Acad. Sci. U.S.A. 113 (6) (2016) 1594–1599.

[53] T. Gayden, et al., Genetic insights into the origins of Tibeto-Burman populations in the Himalayas, J. Hum. Genet. 54 (4) (2009) 216–223.

[54] D. Tagore, et al., Multiple migrations from East Asia led to linguistic transformation in Northeast India and mainland Southeast Asia, Front. Genet. 13 (2022) 1023870.

[55] B.M. Reddy, et al., Austro-Asiatic tribes of Northeast India provide hitherto missing genetic link between South and Southeast Asia, PLoS One 2 (11) (2007) e1141.

[56] M. Metspalu, et al., 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 (1) (2004) 26.

[57] M.G. Palanichamy, et al., Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia, Am. J. Hum. Genet. 75 (6) (2004) 966–978.

[58] B. Ahlawat, et al., Deciphering the West Eurasian genetic footprints in ancient South India, Genes 14 (2023), https://doi.org/10.3390/genes14050963.

[59] L. Kumar, et al., Genetic affinities and adaptation of the south-west coast populations of India, Genome Biol Evol 15 (12) (2023) evad225.