frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 11 October 2022
DOI 10.3389/fgene.2022.1023870

OPEN ACCESS

EDITED BY
Sibte Hadi,
Naif Arab University for Security Sciences (NAUSS), Saudi Arabia

REVIEWED BY
Hui Li,
Fudan University, China
Randy J. LaPolla 羅仁地,
Nanyang Technological University, Singapore

*CORRESPONDENCE
Analabha Basu,
ab1@nibmg.ac.in

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

RECEIVED 20 August 2022
ACCEPTED 27 September 2022
PUBLISHED 11 October 2022

CITATION
Tagore D, Majumder PP, Chatterjee A and Basu A (2022), Multiple migrations from East Asia led to linguistic transformation in NorthEast India and mainland Southeast Asia.
Front. Genet. 13:1023870.
doi: 10.3389/fgene.2022.1023870

Multiple migrations from East Asia led to linguistic transformation in NorthEast India and mainland Southeast Asia

Debashree Tagore[1], Partha P. Majumder[1,2], Anupam Chatterjee[3,4] and Analabha Basu[1]*

[1]National Institute of Biomedical Genomics, Kalyani, India, [2]Indian Statistical Institute, Kolkata, India, [3]Department of Biotechnology, North-Eastern Hill University, Shillong, India, [4]School of Biosciences, Royal Global University, Guwahati, India

NorthEast India, with its unique geographic location in the midst of the Himalayas and Bay of Bengal, has served as a passage for the movement of modern humans across the Indian subcontinent and East/Southeast Asia. In this study we look into the population genetics of a unique population called the Khasi, speaking a language (also known as the Khasi language) belonging to the Austroasiatic language family and residing amidst the Tibeto-Burman speakers as an isolated population. The Khasi language belongs to one of the three major broad classifications or phyla of the Austroasiatic language and the speakers of the three sub-groups are separated from each other by large geographical distances. The Khasi speakers are separated from their nearest Austroasiatic language-speaking sub-groups: the "Mundari" sub-family from East and peninsular India and the "Mon-Khmers" in Mainland Southeast Asia. We found the Khasi population to be genetically distinct from other Austroasiatic speakers, i.e. Mundaris and Mon-Khmers, but relatively similar to the geographically proximal Tibeto Burmans. The possible reasons for this genetic-linguistic discordance lie in the admixture history of different migration events that originated from East Asia and proceeded possibly towards Southeast Asia. We found at least two distinct migration events from East Asia. While the ancestors of today's Tibeto-Burman speakers were affected by both, the ancestors of Khasis were insulated from the second migration event. Correlating the linguistic similarity of Tibeto-Burman and Sino-Tibetan languages of today's East Asians, we infer that the second wave of migration resulted in a linguistic transition while the Khasis could preserve their linguistic identity.

KEYWORDS

Austroasiatic, Khasi, Tibeto Burman, admixture, migration, linguistic transformation

Introduction

The Indian subcontinent is genetically one of the most diverse regions of the world harboring over 1.25 billion people (2011 census). The region has been part of the earliest waves of Anatomically Modern Human (AMH) migrations which peopled South and Southeast Asia, including Australia, beginning around 60,000 years ago (Basu et al., 2003; Kivisild et al., 2003; Macaulay et al., 2005; Thangaraj et al., 2005). Over time, the region has also witnessed multiple waves of migration (Basu et al., 2003; Endicott, Metspalu, and Kivisild 2007; Majumder 2008; Basu et al., 2016) that has contributed to its huge genetic, linguistic, and cultural diversity. The Indian subcontinent is bounded in the North and Northeast by the Himalayas. NorthEast India (NEI) is a unique region that is bordered in the north by the high ranges of eastern Himalayas and two-thirds of it is intermediate hilly terrain, interspersed by fertile riverbeds and flat valleys. The population density also varies accordingly; while the river valleys are densely populated and cosmopolitan, the highlands are sparsely populated by small isolated ethno-lingual groups. Major population groups that reside here speak Tibeto-Burman languages, which belong to the non-Sinitic phylum of the Sino-Tibetan language family. Now restricted by political boundaries, this region is likely to have been a land bridge between peninsular India (PI) and Mainland Southeast Asia (MSEA) and has been an active corridor of migration and admixture of different ethnolinguistic populations in the past (Gadgil et al., 1993; Cavalli-Sforza et al., 1994; Reddy et al., 2007; Tagore et al., 2021; Liu et al., 2022) and hence should be considered in continuum with the population demographic history of East and Southeast Asia. Individuals from mainly five language families reside in NEI and the neighborhood of MSEA: namely Sino-Tibetan, Tai-Kadai, Hmong-Mein, Austronesian, and Austroasiatic (AA). However, more recent migrations of the ancestors of Indo-European language speakers of India, who possibly entered India through the northwestern corridor also had a large impact on the populations of NEI (Gayden et al., 2009; Basu et al., 2016). The Austroasiatic language family comprises three major subfamilies: Munda, Mon-Khmer, and Khasi-Khmuic (Diffloth Gerard 2005a). Within the Austroasiatic family, the Khasi language (the sole language of the Khasi-Khmuic branch of the Austroasiatic language family in India) is spoken in NEI mainly in parts of the north-eastern state of Meghalaya.

In this study we look into the population genetics of the Khasi, residing amidst the Tibeto-Burman speakers as an isolated population. These Khasi speakers are separated by large physical distance from their nearest Austroasiatic language-speaking sub-groups: the Munda sub-family from East and peninsular India and the Mon-Khmer sub-family in Mainland Southeast Asia. Here, we dissect the genetic relationship of the Khasis with the other Austroasiatic subgroups and in an attempt to do so, reconstruct the population history of NorthEast India, and the neighboring East and Southeast Asia, in the context of the Khasi Austroasiatics.

Despite a strategic location, most genetic studies on NEI populations have been done using either uniparental markers (Cordaux et al., 2003; Borkar et al., 2011) or a small number of autosomal markers (Maity, Nunga, and Kashyap 2003; Krithika et al., 2005; Krithika et al., 2006; Mastana et al., 2007; Gayden et al., 2009). Cordaux et al. (2003) mitochondrial DNA (mtDNA) and Y chromosome-based study suggest two possibilities regarding the peopling of NEI: either TBs were the earliest inhabitants, or the TB replaced the Austroasiatic (AA) inhabitants of NEI. Using microsatellite data, and comparing the Khasi-Khmuic speakers with their neighboring Tibeto-Burman speakers, showed the populations to be extremely homogeneous (Langstieh et al., 2004) a fact further supported by later studies with mtDNA and Y-chromosome (Cordaux et al., 2004). Initially, researchers came up with opposing views on the origin of TB populations. One theory, based on Y-chromosome analyses, suggests that the TB ancestors originated in the upper and middle Yellow River basin (Su et al., 2000). Another theory suggests the Yangtze River as their ancestral source followed by the northward movement to the Yellow River basin (Van Driem 2005). In our previous study (Tagore et al., 2021) we also proposed a theory where we suggested that present-day Tibeto Burmans were likely Austroasiatics in the past, who were part of the earliest settlers of the region (Hill et al., 2006). Y-chromosome based study by Wang et al. (2018) suggested that the peopling of the Tibetan plateau by Tibeto Burman ancestors happened some 40KYA (40 thousand years ago). This coincides with the presence of hunter gatherers in this region. However, it was during the Neolithic period, ~6KYA, when the expansion of different Y chromosome lineages was observed leading to the present day distribution of the TBs. This time coincides with the migration of East Asians in MSEA (Tagore et al., 2021). Yu et al. (2021) have suggested that migration of both Yellow river basin millet farmers and Yangtze river basin rice farmers contributed to different linguistic and genetic groups in MSEA. They have also proposed that around 6KYA, people from the middle Yellow River Basin migrated south-westward and mixed with the local population to give rise to the initial TBs. Basu et al. (2003) has shown that the TB and AA speakers of India are similar in their mtDNA profile but harbor very distinct Y-chromosomes. Our previous study (Tagore et al., 2021) observed the genetic relatedness between the Tibeto Burmans and Austroasiatic speakers (Mon Khmers) of Malaysia, because of ancient shared ancestry as well as owing to gene flow in both these populations from East Asia. Another study (Guo et al., 2022) found the present day TBs to cluster between the millet cultivators of Yellow River basin as well as the Austroasiatic speakers of Southeast Asia in a Principal Components Analysis. In further analyses they found southern Tibeto Burmans were genetically closest to the AAs.

The Northeast Indian populations were clustered with populations of East and Southeast Asia than with mainland Indians (Langstieh et al., 2004; Basu et al., 2016; Tagore et al., 2021). Other studies on the mtDNA hypervariable region and autosomal microsatellite markers found that despite the present political boundary, the Tibeto-Burman speakers from NEI showed a closer genetic affinity with East Asian populations than with other mainland Indian populations (Cordaux et al., 2003; Krithika, Maji, and Vasulu 2008; Basu et al., 2016). This further supports the fact that ancient migration events occurred through the NEI corridor before the political boundaries were drawn (Basu et al., 2003; Basu et al., 2016). Such genetic studies are in agreement with the linguistics of Northeast India: the Tibeto Burman language group is closely related to the languages of East Asia. Apart from the genetic similarity with the East Asians, the TB also shows some complex admixture with other Indian populations belonging to different ancestries. The TBs harbor genetic ancestry predominantly in Indo-European speakers (henceforth referred to as ANI or Ancestral North Indian) who mainly reside in the northern part of India, and also genetic ancestry predominantly in Dravidian language speakers (henceforth referred to as ASI or Ancestral South Indian) who are almost exclusively confined to the southern part of India (Basu, Sarkar-Roy, and Majumder 2016).

The Khasis are one of the few populations in the world that follow a matrilineal system of inheritance. Besides the linguistic similarity, anthropologists and archaeologists have established that the Khasis have cultural similarities with Mundaris and Mon-Khmer populations. It has been shown that they share similar stone tools and have similar death rituals of erecting memorial stones for the deceased (Gurdon 1914). Linguistically, the Khasi language is more similar to languages of the Mon-Khmer branch than those of the Mundari branch and linguists have often assigned Khasi and Mon-Khmer languages to the same group (Pinnow et al., 1942; Chazée 1999). Khasi language also bears lexical and morphological similarities to some Tibeto Burman languages (Longmailai 2015). Peiros suggested a significant number of words were similar between Proto Austroasiatics and Proto-Sino-Tibetans (Peiros 2011).

Nevertheless, the presence of ancient Austroasiatics (AA) speakers across NEI still remains a possibility. Our previous study (Tagore et al., 2021) on autosomal data of the Mundari and Mon-Khmer Austroasiatics indicated that in pre-Neolithic times, the ancestors of today's Austroasiatic speakers had a widespread distribution possibly extending from Central India to Southeast Asia (SEA), further supported by Lipson et al. (2018). They were later in time fragmented and isolated to small pockets resulting in their present-day disjoint geographic distribution. What is intriguing is that given the widespread distribution of Austroasiatic speakers from Central India to SEA across NEI and the central location of the Khasis, it is possible that the Khasis will serve as a genetic link between the two Austroasiatic groups on either side of NEI.

There have been very few studies on the genetics of Khasi, so as to reach any plausible conclusions. One previous study on uniparental markers has proposed a genetic continuity between the Mundari Austroasiatics of Central India, Khasi-Khmuic, and Mon-Khmer (Reddy et al., 2007). Using multidimensional scaling of the pairwise F_{ST} distances calculated on Y-haplogroups of Austroasiatics and neighboring populations, they found the three Austroasiatic groups (Mundari, Khasi-Khmuic, and Mon-Khmer) to cluster together. They also found the Y haplogroup O-M95, restricted within the Austroasiatics and postulated to have originated in the Mundaris, is present in the Khasis at a frequency intermediate to that of Mundaris and Mon-Khmers. They suggested an initial presence of Austroasiatics in Central India with rapid migration to Southeast Asia *via* the Northeast corridor carrying the O-M95 haplogroup.

The cultural and linguistic similarities of the Khasis with other Austraoasiatic groups as mentioned earlier prompt us to investigate their genetic affinities. The geographic location of the Khasis also makes it imperative to investigate the impact of East Asian migrations on the genetic make-up of the Khasis.

Materials and methods

Dataset preparation and quality control

DNA samples from 22 individuals speaking the Khasi language were sequenced and merged with the genotype dataset of 1,451 individuals that were used in our previous study (Tagore et al., 2021) using PLINK(Shaun et al., 2007). The details of the datasets are provided in Supplementary Table S1A–C. Only biallelic loci were included in our analysis. We removed all monomorphic variants and SNPs with alleles A/T and G/C from our analysis. We also removed SNPs with missingness of more than 5% in the entire dataset, or SNPs that were missing in more than 25% of individuals in any of the 15 subpopulations (second column of Supplementary Table S1A). We also excluded SNPs that were out of Hardy Weinberg equilibrium ($p < 10^{-6}$) in any of the 15 subpopulations. This combined dataset had 310110 SNPs.

Principal components analysis

In order to understand the overall population structure and the genetic affinities of the individuals in our dataset, we performed Principal Components Analyses (PCAs) using the smartpca program of the EIGENSOFT package (Patterson, Price, and Reich 2006). We performed an initial PCA on all the mainland Indians (all Indian populations excluding those belonging to the "Island" group as in Supplementary Table

S1A) and Malaysian populations. We considered the first two Principal Components (PCs) to visualize the data.

A second PCA was run on a subset of populations used in the first PCA. This subset was chosen based on linguistic similarity and geographic proximity to the Khasis. Thus, we included the Austroasiatics from Central India (AACI), Austroasiatics of Malaysia (AAM), Khasi, and Tibeto Burmans (TBs).

TreeMix

In order to understand how populations were related to each other through a common ancestor and the impact of genetic drift, we built ancestry graphs using TreeMix (Pickrell et al., 2012) version 1.12. Such graphs were created with AACI, AAM, TB, and the Khasi populations using the Mbuti Pygmies from Africa as an outgroup.

F_{st} estimates

Using PLINK (Shaun et al., 2007) version 1.9, the weighted F_{st} between each subpopulation of AACI, AAM, TB, and the Khasi was estimated. These values were rounded to the third decimal place.

Outgroup *f3* statistics

Outgroup *f3* statistics measures the shared drift between two populations relative to an extremely diverged population outgroup. Using ADMIXTOOLS (v5.1) (Patterson et al., 2006), we calculated outgroup *f3* statistics of the form *f3* (Mbuti Pygmy; Khasi, Y) where Mbuti Pygmy was the outgroup. Y was AACI, AAM, and TB subpopulations.

ADMIXTURE analysis

To infer the different ancestral components present in the admixed populations and the proportions of each such component in an individual's genome, we performed unsupervised clustering as implemented in ADMIXTURE (Alexander, Novembre, and Lange 2009) (v1.3.0). We ran ADMIXTURE using all Indian populations (AACI, ANI, ASI, ATB, and Khasi), Malaysian populations (AAM and ANS), and all East Asians. We ran ADMIXTURE by sequentially increasing the number of clusters, which corresponds to the number of identified ancestries (k), in each run of the analysis on a given dataset. ADMIXTURE estimates the proportion of each of the k ancestries in the genome of each individual of the dataset and also computes a cross-validation error (CVE) for that particular run. Standard error was estimated for the ancestry proportion estimates at the minimum CVE using the moving block bootstrap approach as implemented in ADMIXTURE.

Admixed segment length calculation

Dataset was phased using SHAPEIT v2.r790 (Delaneau, Marchini, and Zagury 2012). From the phased dataset we extracted the phased genomes for Khasi, Jamatia, Miazou (a Southern East Asian-like ancestry population), Yakut (a Northern East Asian-like ancestry population), and Kshatriya (an ANI-like ancestry population). This was followed by local ancestry estimation using RFMix (Maples et al., 2013) version 1.5.4, to identify regions of genomes of Khasi and a TB population (in this case Jamatia) corresponding to different ancestries. The different ancestries which we considered for the local ancestry estimation were inferred from the ADMIXTURE run where the CVE was minimum, i.e. at $k = 8$. Ancestries for which tract lengths were estimated in Khasi included: "Southern EA-like", "Jehai-like", "MahMeri-like", "Birhor-like (AACI-like)" and "ANI-like". In addition to these ancestries, tract lengths corresponding to "Northern EA-like" ancestry were also estimated for Jamatia. It is to be noted here that the Northern EA-like ancestry was practically absent in the Khasis. We plotted the cumulative distribution of these tract lengths to compare the sizes of these tract lengths corresponding to the different ancestries in both Khasi and Jamatia.

Estimating admixture time

Gene flow events between genetically distinct populations create linkage disequilibrium between all loci that are highly differentiated between the two ancestral populations. Segments resulting from admixture follow an exponential distribution, where as a result of recombination, this linkage disequilibrium pattern declines exponentially over time and from which the number of generations since admixture can be estimated (Racimo et al., 2015). To date the 'time since the last admixture event' between different populations, we generated "co-ancestry curves" using MOSAIC (Salter-Townshend and Myers 2019) (v1.2). Here the closest surrogate populations were chosen as "donors". Coancestry curves measure how often, in an admixed ("recipient") population; a pair of haplotypes has been inherited from each respective donor population. Given a single admixture event, ancestry chunks inherited from each source, reduce in size because of recombination, resulting in an exponential decay of these coancestry curves. The time (in generations) since admixture is calculated from the rate of decay in the curves.

To detect 2-way admixture events in Jamatia and Khasi, we used Yakut as a surrogate donor of the "Northern EA-like"

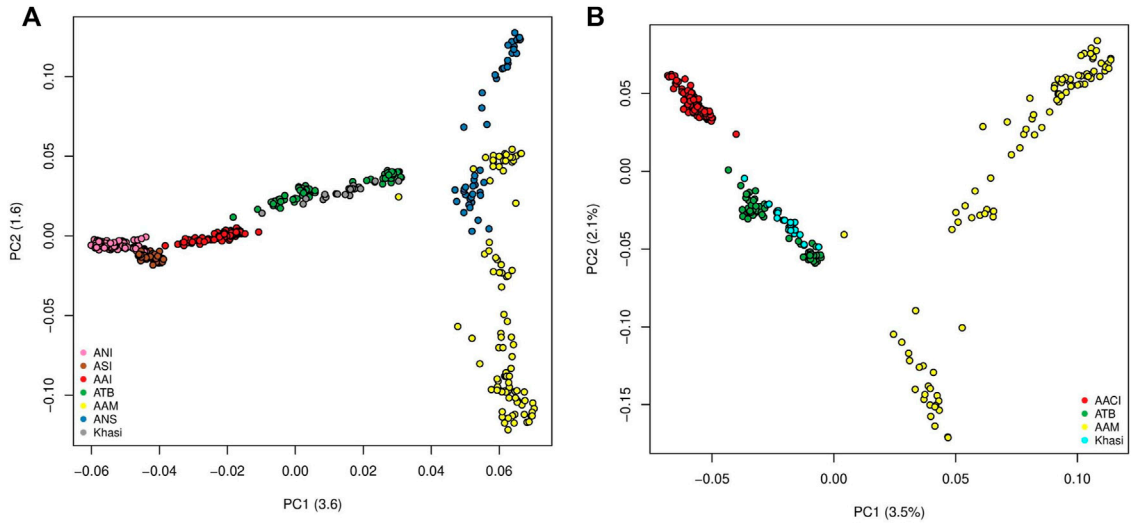

FIGURE 1
(A) PCA on Mainland Indian and Malaysian populations (ANI: Ancestral North Indian, ASI: Ancestral South Indian, AACI: Austroasiatics of Central India, ATB: Ancestral Tibeto Burmans, AAM: Austroasiatics of Malaysia, ANS: Austronesians); **(B)** PCA on the Austroasiatics and the Tibeto Burmans.

ancestry, Miazou for "Southern EA-like" ancestry and Birhor for "Austroasiatic-India or Mundari" ancestry. We chose Khasi and Jamatia as recipients (having ancestry from each of the source populations as a result of admixture) and estimated the time since the last admixture between the donors. We created co-ancestry curves for the surrogate donors (Miazou and Birhor) in both Khasi and Jamatia and another co-ancestry curve for donors Yakut and Birhor in Jamatia. The rate of decay in the curves was calculated which was equal to the number of generations since admixture took place.

Results

The first two Principal Components (PCs) of the PCA with all the mainland Indians (all Indian populations excluding those belonging to the "Island" group as in Supplementary Table S1A) and Malaysian populations explained 3.6% and 1.6% of the variation. In PC1-PC2 space the individuals belonging to the major population groups (as classified in the second column of Supplementary Table S1A), formed unique clusters. In the PC1 axis the Indian population, specifically the Ancestral North India-like (ANI-like) populations were on one extreme while the Malaysian populations were on the other. While most Indian populations were distinguishable along the first PC, the two Malaysian populations separated along the second PC. We found the Khasis to cluster with the Tibeto Burmans (Figure 1A) (It is to be noted that this is very similar to Figure 1D in Tagore et al., 2021 where we had all the populations except the Khasis). Using a smaller subset of the above we did a second PCA, where, we considered the Khasis along with the two Austroasiatic groups from our previous study i.e. Mon-Khmer speaking Austroasiatics from Malaysia (AAM), Mundari speaking Austroasiatics from Central India (AACI). We also included the Tibeto-Burman population (TB) who were geographically proximal to the Khasis (Figure 1B; Supplementary Figure S1). We considered three Principal Components (PCs) which together could explain 7.1% of the total variation. In the PC1-PC2 space, we found that the three Austroasiatic groups formed distinct clusters. The Khasi did not cluster with either of the other two Austroasiatic populations, instead, clustered with the TB subgroups (Figure 1B). Khasi and TB were distinguished as separate clusters in PC3. Here we could also identify two separate clusters within the TB: one comprising Jamatia and Tripuri (that clustered closer to the AAM) and the other comprising M-Brahmin and Tharu (Supplementary Figure S1), who are known to have been admixed with other populations of North India and the Upper Gangetic plains (Basu, Sarkar-Roy, and Majumder 2016). A similar pattern of clustering was found in the TreeMix analysis (Supplementary Figure S2). The Khasis clustered with the Tibeto Burmans in a branch separate from the other two Austroasiatic populations. The Munda and Mon-Khmer populations also formed distinct clusters.

On the same set of the population (as used in Figure 1B), we surveyed the allele frequencies and calculated pairwise F_{st} (Weir and Cockerham 1984) between them using PLINKv1.9. Here again, we found F_{st} between the Khasi and TB groups to be low

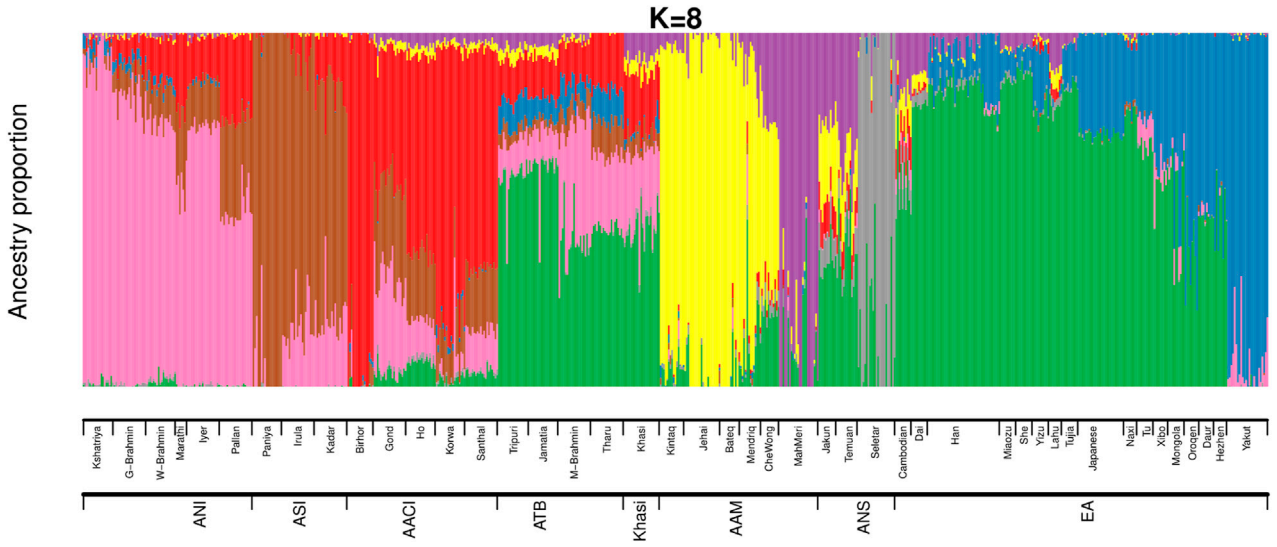

FIGURE 2
ADMIXTURE analysis on Mainland Indian populations, Malaysians and East Asians of HGDP (ANI: Ancestral North Indian, ASI: Ancestral South Indian, AACI: Austroasiatics of Central India, ATB: Ancestral Tibeto Burman, AAM: Austroasiatics of Malaysia, ANS: Austronesians, EA: East Asians of HGDP).

(mean = 0.019) (Supplementary Figure S3) which was even lower than those between Khasi and AAM (mean = 0.061) and between Khasi and AAI (mean = 0.033).

In our analysis, *f3* (Mbuti Pygmy; Khasi, X) we used the African Mbuti Pygmy as the outgroup. We measured the *f3* values of Khasi with AAM, AACI, and TB (details in Materials and Methods, Supplementary Table S3). The mean *f3* values are highest (mean = 0.279) between Khasi and TB, indicative of an exclusive and recent shared genetic history. The mean *f3* values were higher (mean = 0.277) for Khasi-AAM than between Khasi-AAI (mean = 0.265). The apparent discordance in the pattern of f3 and F_{st} values when Khasi are compared to AACI and AAM is largely due to the fact that F_{st} is affected by drift. It is to be noted here that we see the AAM populations be highly drifted in our Treemix analyses (Supplementary Figure S2).

We then estimated the genomic ancestries and admixture proportions at an individual level by considering all populations from India and Malaysia (the same populations we used in Figure 1A). We also included the East Asians in this analysis because we had observed in our previous study (Tagore et al., 2021) that an East Asian ancestry component was found among some Austroasiatic populations. We did an ADMIXTURE (Alexander, Novembre, and Lange 2009) analysis (Figure 2; Supplementary Figures S4A,B) where the Cross-Validation Error (CVE, details in Materials and Methods) was minimized at $k = 8$ (Supplementary Figure S4A). We also calculated the ancestry proportions for each of these populations (Supplementary Table S3).

The ancestry proportions of the Khasi, as estimated by ADMIXTURE, are distinct from the other two Austroasiatic groups (AAM and AACI) but are very similar to that of the Tibeto Burmans, especially to the Jamatia and Tripuri. The major ancestry identified among the Khasi was also the predominant ancestry identified among the Southern-EA populations (like Dai). In the ADMIXTURE plot (Figure 2), it is depicted by the green color (nearly 44%, "Southern-EA-major" in Supplementary Table S3). This green-colored component is the ancestry modal to the Southern East Asians (henceforth referred to as "Southern EA-like" ancestry) such as Dai. The Khasi genome also has a substantial proportion of AACI-like ancestry (16%, "AACI-major"; red in color) and AAM-like ancestry (4%; yellow color modal to Jehai and 7%; purple color modal to MahMeri). 20% of the Khasi genome is of ANI-like ancestry ("ANI-major", pink in color). Neither Khasi nor the TB groups had in them any distinctly identified "Khasi-like" or "TB-like" component respectively. Alternatively, the AACI and AAM had genomic components mostly exclusive to them: 64% "AACI-major" component (red in color) and 62% "AAM-major" components (yellow in color) respectively. The East Asian component present in Khasis (green in color) was also present in AACI and AAM, though in lesser proportions of 3.5% and 7.6% respectively. The TBs, on the other hand, had an even higher proportion of this component (51%). In addition to this East Asian component, TBs also have a substantial proportion (7.4%) of a second East Asian component (blue in color). This East Asian component (henceforth referred to as "Northern-EA-major") is modal in the East Asian populations residing in today's Northern China (e.g. the Yakut). It is to be noted here that these "Northern-EA" and "Southern-EA"

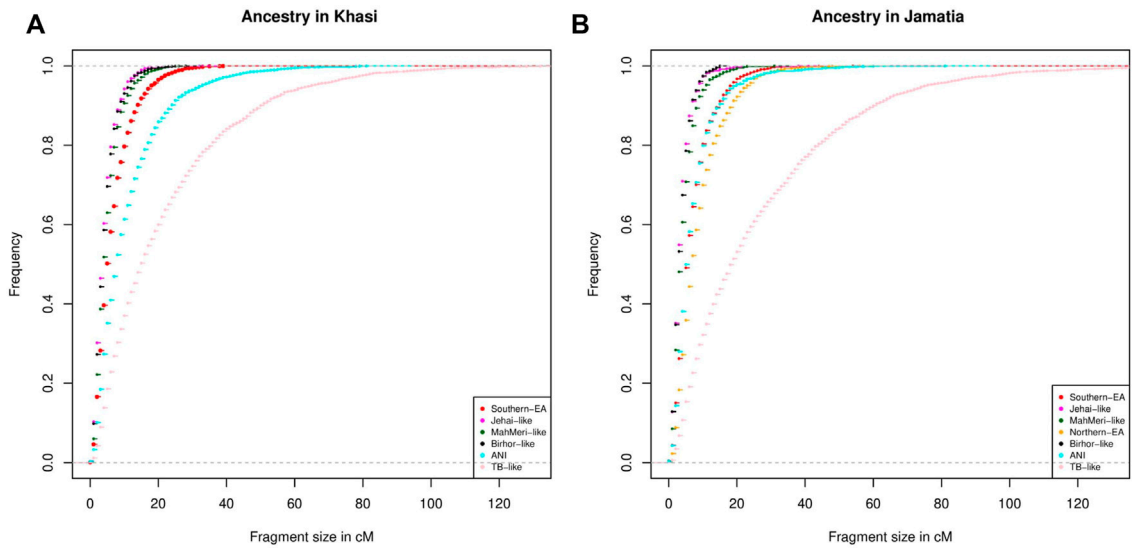

FIGURE 3
(A) Distribution of tract lengths corresponding to different ancestries in Khasi; **(B)** Distribution of tract lengths corresponding to different ancestries in Jamatia.

components were also identified in our previous study. Although the TB (particularly Jamatia and Tripuri) and Khasi individuals cluster close in the PCA, this ancestry is negligible in the Khasis. Compared to the Southern EA-like ancestry, the Northern EA-like ancestry is negligible in the other two Austroasiatics groups (AACI and AAM) as well.

To investigate the chronology of admixture events into Khasi, we performed local ancestry estimation, using RFMix (Maples et al., 2013). We identified regions within genomes of Khasi individuals representing different ancestries as inferred in the ADMIXTURE analysis. We estimated the length of admixed tracts representing the following ancestries: "Jehai-like", Birhor-like", "MahMeri-like", ANI-like" and "Southern -EA-like". We looked into the cumulative frequency distribution of these tract lengths. Larger tracts (segments) would correspond to a recent introduction of the corresponding ancestry, and hence can be used as an indicator of the sequence of admixture events. We found that the tract lengths corresponding to Birhor-like and Jehai-like ancestry are the smallest in the Khasis. This is followed by MahMeri-like, ANI-like and Southern EA-like ancestry tracts. This indicates that Southern EA-like ancestry is most recently introduced in the Khasis (Figure 3A).

We then repeated the same analysis with the Jamatia (a subgroup of the TB) and with the same five ancestries i.e. Birhor-like, Jehai-like, MahMeri-like, ANI-like, and Southern EA-like ancestries. Furthermore another ancestry: the Northern EA-like was also included in the analyses because this was an additional ancestry present substantially in the TBs (as evident from ADMIXTURE analysis) but was absent among the Khasis. Similar to what was observed in the Khasis, tract lengths corresponding to the Birhor-like and Jehai-like ancestry are the smallest followed by MahMeri-like, ANI-like, Southern EA-like ancestries. This mimics the chronology of the admixture events that we see in the Khasis. However, the largest length of admixture tracts corresponded to the Northern EA-like ancestry (Figure 3B). This indicates that though the overall sequence of admixture i.e. introduction of ancestries within the Khasis and TBs are similar, the introduction of the Northern EA-like ancestry is the most recent event and unique to the TBs. Thus we conclude that the admixture of the East Asian populations and ancestors of present-day Khasi and Tibeto Burmans is a relatively recent event; of the two distinct East Asian genetic ancestries, the Northern-EA ancestry was introduced in the Tibeto Burmans subsequent to the Southern EA-like ancestry. While both Khasis and TBs have experienced multiple admixture events, the Northern East Asian admixture largely with the TBs is the one which is unique and recent.

We further dated these local admixture events using a method implemented in MOSAIC (Salter-Townshend and Myers 2019) that infers admixture time by fitting an exponential decay coancestry curve (details in Supplementary Material). We chose homogeneous representative populations such as Yakut for Northern EA-like ancestry, Miazou for Southern EA-like ancestry and Birhor for Central Indian Austroasiatic ancestry, as a source population for admixture in populations such as Khasi and Tibeto-Burman. We found that the last evidence of admixture between Southern EA-like ancestry-bearing populations and Austroasiatics and TB took

place 13.9 and 10.5 generations ago when Khasi and Jamatia (a representative subgroup for the Tibeto Burman population) were chosen as recipients (Supplementary Figures S5A,B). We also found that the incorporation of Northern EA-like ancestry in the Jamatia happened as recently as 8.3 generations ago (Supplementary Figure S5C). These findings were in accordance with the chronology of events we inferred from the RFMix analysis. This substantiates our conclusion that there were at least two distinct admixture events in NEI populations with East Asians where populations bearing Southern EA-like ancestry admixed first with both the ancestors of TB and Khasi and later populations bearing Northern EA-like ancestry admix mostly with the ancestors of TB.

Discussion

The Khasi are a relatively large population, subdivided into groups owing to geographical barriers, and show considerable heterogeneity as evident from an anthropometric study (Das 1970). We find from our analyses (PCA, ADMIXTURE, TreeMix), that the Khasis are genetically very similar to the Tibeto-Burmans. The PCA cannot identify the Khasis as a distinct cluster, separate from other TB populations when compared with AACI and AAM. The Khasi Austroasiatics are distinct from the other Austroasiatics (Mundari or AACI and Mon-Khmer or AAM) in our study which conforms to the linguistic classification by Diffloth (2005b). The ancestral components inferred using ADMIXTURE in the Khasis and TBs are also very similar. In our previous study, we had observed that the AACI, TB, and AAM shared a deep common ancestry and proposed that all of their ancestors likely spoke some proto-AA language. The observed genomic profile of the Khasis suggests that the Austroasiatic-speaking Khasis fit well into the proposed model. We had also postulated that the ancestors of the present-day TB and the AAM populations experienced admixture with southward migrating EA agriculturists. Here we find that the Khasis, residing in the same region as the TBs, experienced the same sequence of admixture events as the Jamatia (TB). This indicates they likely share a common history. The southward migration of East Asians led to the incorporation of East-Asian ancestry in the Tibeto Burmans and the Khasis. This migration was extensive and as we have also previously observed, the admixture signals of this migration can also be found among other AA speakers (predominantly the AAM).

Despite an overall genomic similarity of the TBs and Khasis, there is a distinct difference between the TB populations and the Khasis: unlike the TBs, the Khasis lack Northern East Asian ancestry. Results from our RFMix analysis suggest that there were at least two distinct waves of East Asian migration. The first wave brought the Southern East Asian ancestry that got incorporated in both the Khasis and the TBs and the second wave brought the Northern East Asian component. This migration started possibly from Northern EA and led to the introduction of the Northern EA-like genomic ancestry into the TB population but not in the Khasis. It is to be noted that the Northern EA component is absent from other Austroasiatic speakers as well (AACI and AAM), although some of these populations have substantial EA ancestry, i.e Southern EA ancestry. Wang et al. (2018) suggested that the TBs were an admixed group resulting from two distinct ancient populations: a hunter-gatherer population, (which we believe were the proto-Austroasiatics) and a millet farmer population from middle Yellow River basin. A genetic link between the millet farming proto-Sino Tibetans of the Yellow River basin and Tibeto Burmans has also been proposed by Guo et al. (2022). Though Wang et al propose a two wave migration leading to the formation of TBs, they propose that out of the two, only one wave of migration formed both the TBs of India and populations of MSEA. However, our study suggests that though there were atleast two waves of migration, the second wave solely affected the TBs while the first affected both the TBs and the AAs of MSEA.

We, therefore, propose, in agreement with our previous study, that the ancestors of extant Austroasiatic speakers were widespread across Central India and Southeast Asia encompassing the present-day location of the TBs and Khasis. This is in agreement with other studies (Cordaux et al., 2004). It is hence plausible that the ancestors of present-day Tibeto-Burman speakers spoke some form of an Austroasiatic or Proto-Austroasiatic language. With time, the Austroasiatic populations evolved into three major branches as we see them today namely Mundari, Mon-Khmer, and Khasi.

Higham has suggested that before Neolithic expansion, this region was inhabited by hunter gatherers (Higham 2017). He also suggested that the expansion of farming communities happened from two regions that reached mainland Southeast Asia: one of millet cultivators from the Yellow River basin and another of the rice cultivators from the Yangtze River basin. Such migration events are also supported by morphological studies. Cranial (Matsmura 2011) and Dental (Matsmura 2010) morphological studies found two groups of individuals at the Man Bac excavation site in Southeast Asia: one close to the Neolithic inhabitants of Weidun in the Yangtze Valley and the other to the local hunter gatherers. Archaeological studies in Southeast Asia also supports presence of hunter-gatherers in Southeast Asia as well as Southern China (Higham 2013) and that archaeological sites provide indication that immigrants from Southern China encountered these hunter gatherers on their way. Infact, Neolithic migration has also diluted the genetic differentiation within China (Yang et al., 2020). An extensive documentation of rice spread also supports the spread of rice from China to Southeast Asia (Fuller et al., 2010).

We argue that the language of the extant TBs is a result of this linguistic shift, possibly evidence of elite dominance, which is a

consequence of the migration and gene flow from Northern East Asia. When we look at the ancestral segments of Northern-EA ancestry, we find that they are among the longest ancestral segments in TB, preceded by segments of Southern EA-like ancestry. We postulate that the two migration events from East Asia were such that initially, populations bearing Southern EA-like ancestry arrived in NEI, and later came the populations of Northern EA-like ancestry. The Southern EA-like ancestral segments are also present in Khasis and AAM, the two Austroasiatic groups with substantial East Asian ancestry. In these populations, Southern EA-like ancestral segments are among the longest. The AACI however have negligible East Asian components in their genome. It is to be noted here that the language of the AACI, i.e. Mundari is much more distant from the other two branches of the AA family, namely Khasi-Khmuic and Mon-Khmer. The Khasi-Khmuic and Mon-Khmer are more similar to the Sino-Tibetan language. This is expected as our genetic data also confirms closer proximity and longer admixture of Khasi-Khmuic and Mon-Khmer speaking populations (Khasi and AAM) with Southern-EA populations. The admixture with populations of Northern EA-like ancestry is unique among the TB and their languages belong to Sino-Tibetan, a different language family altogether. In TBs this ancestry has been incorporated after the second migration wave. This leads us to conclude that the ancestral populations of TB have experienced a language shift, from a more proto-Khasi-Khmuic language to a language closer to that of the East Asians (the Tibeto Burman languages) and this has occurred due to the most recent admixture with populations with Northern EA-like ancestry.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://share.nibmg.ac.in/d/0b373e011a1d4f689cb2/, BMCB2021; https://www.nibmg.ac.in/fp/khasi_data.html, Khasi_data_2022.

Ethics statement

The studies involving human participants were reviewed and approved by Institutional Ethics Committee for Human Samples/Participants (IECHSP/2014/07), North-Eastern Hill University, Shillong, India. The patients/participants provided their written informed consent to participate in this study.

Author contributions

AB and DT designed the study with the active participation of AC and PM. DT analyzed the data and prepared the final figures and tables. DT and AB wrote the manuscript with inputs from AC and PM. All authors read and approved the final manuscript.

Funding

This work was supported by the Department of Science and Technology (DST), Government of India. The infrastructure was provided by NIBMG and supported by the Department of Biotechnology (DBT), Government of India.

Acknowledgments

The authors are thankful to Diptarup Nandi at the National Institute of Biomedical Genomics for his valuable comments and suggestions. We also thank Genomics lab of NIBMG for the sequencing of the Khasi genomes and Arnab Ghosh for initial processing of the sequence data.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1023870/full#supplementary-material

References

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi:10.1101/gr.094052.109

Basu, A., Mukherjee, N., Roy, S., Sengupta, S., Banerjee, S., Chakraborty, M., et al. (2003). Ethnic India: A genomic view, with special reference to peopling and structure. *Genome Res.* 13 (10), 2277–2290. doi:10.1101/gr.1413403

Basu, A., Sarkar-Roy, N., and Majumder, P. P. (2016). 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), 1594–1599. doi:10.1073/pnas.1513197113

Borkar, M., Ahmad, F., Khan, F., and Agrawal, S. (2011). Paleolithic spread of Y-chromosomal lineage of tribes in eastern and northeastern India. *Ann. Hum. Biol.* 38 (6), 736–746. doi:10.3109/03014460.2011.617389

Cavalli-Sforza, L. L., Piazza, L. L. C. S. P. M. A., Cavalli-Sforza, L., Menozzi, P., Piazza, A., and Princeton University Press (1994). *The history and geography of human genes*. Princeton, NJ, USA Princeton University Press.

Chazée, L. (1999). *The peoples of Laos: Rural and ethnic diversities: With an ethno-linguistic map*. Limited: White Lotus Company.Chennai, India (Thailand).

Cordaux, R., Saha, N., Bentley, G. R., Aunger, R., Sirajuddin, S. M., and Stoneking, M. (2003). Mitochondrial DNA analysis reveals diverse histories of tribal populations from India. *Eur. J. Hum. Genet.* 11 (3), 253–264. doi:10.1038/sj.ejhg.5200949

Cordaux, R., Weiss, G., Saha, N., and Stoneking, M. (2004). the Northeast Indian passageway: A barrier or corridor for human migrations? *Mol. Biol. Evol.* 21 (8), 1525–1533. doi:10.1093/molbev/msh151

Das, B. M. (1970). Somatic variation among the Khasi populations of Assam, India. *Zmorph_anthropol.* 3, 259–266. doi:10.1127/zma/62/1970/259

Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nat. Methods* 9 (2), 179–181. doi:10.1038/nmeth.1785

Diffloth, G. (2005a). "The contribution of linguistic palaeontology and Austroasiatic," in *The peopling of east Asia: Putting together archaeology, linguistics and GeneticsRoger blench and alicia sanchez-mazas laurent sagart* (Newyork, NY, USA: Routledge Curzon), 77–80.

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

Endicott, P., Metspalu, M., and Kivisild, T. (2007). "Genetic evidence on modern human dispersals in South Asia: Y chromosome and mitochondrial DNA perspectives: The world through the eyes of two haploid genomes," in *The evolution and history of human populations in south Asia* (Springer), NewYork, NY, USA, 229–244.

Fuller, D. Q., Sato, Y.-I., Castillo, C., Qin, L., Weisskopf, A. R., Kingwell-Banham, E. J., et al. (2010). Consilience of genetics and archaeobotany in the entangled history of rice. *Archaeol. Anthropol. Sci.* 2 (2), 115–131. doi:10.1007/s12520-010-0035-y

Gadgil, M., Shambu Prasad, U. V., Manoharan, S., and Patil, S. (1993). *Peopling of India*. Chennai, India: IHC.

Gayden, T., Mirabal, S., AliciaCadenas, M., Lacau, H., M Simms, T., Morlote, D., et al. (2009). Genetic insights into the origins of Tibeto-Burman populations in the Himalayas. *J. Hum. Genet.* 54 (4), 216–223. doi:10.1038/jhg.2009.14

Guo, J., Wang, W., Zhao, K., Li, G., He, G., Zhao, J., et al. (2022). Genomic insights into Neolithic farming-related migrations in the junction of east and southeast Asia. *Am. J. Biol. Anthropol.* 177 (2), 328–342. doi:10.1002/ajpa.24434

GurdonThornhaghand Philip Richard (1914). *The Khasis*.NewYork, NY, USA Macmillan.

Higham, C. (2013). Hunter-gatherers in Southeast Asia: From prehistory to the present. *Hum. Biol.* 85 (1/3), 21–43. doi:10.3378/027.085.0302

Higham, C. F. (2017). First farmers in Mainland southeast Asia. *J. Indo-Pacific Archaeol.* 41, 13–21. doi:10.7152/jipa.v41i0.15014

Hill, C., Soares, P., Mormina, M., Macaulay, V., Meehan, W., Blackburn, J., et al. (2006). Phylogeography and ethnogenesis of aboriginal southeast Asians. *Mol. Biol. Evol.* 23 (12), 2480–2491. doi:10.1093/molbev/msl124

Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., Parik, J., et al. (2003). The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am. J. Hum. Genet.* 72 (2), 313–332. doi:10.1086/346068

Krithika, S, Trivedi, R., Kashyap, V. K., Vasulu, T. S., and Kashyap, V. K. (2005). Genetic diversity at 15 microsatellite loci among the Adi Pasi population of Adi tribal cluster in Arunachal Pradesh, India. *Leg. Med.* 7 (5), 306–310. doi:10.1016/j.legalmed.2005.04.002

Krithika, S, Trivedi, R., Kashyap, V. K., Bharati, P., and Vasulu, T. S. (2006). Antiquity, geographic contiguity and genetic affinity among tibeto-burman populations of India: A microsatellite study. *Ann. Hum. Biol.* 33 (1), 26–42. doi:10.1080/03014460500424043

Krithika, S., Maji, S., and Vasulu, T. S. (2008). A microsatellite guided insight into the genetic status of Adi, an isolated hunting-gathering tribe of Northeast India. *PloS one* 3 (7), e2549. doi:10.1371/journal.pone.0002549

Langstieh, B. T., B Mohan Reddy, K. T., Kumar, V., and Singh, L. (2004). Genetic diversity and relationships among the tribes of Meghalaya compared to other Indian and Continental populations. *Hum. Biol.* 76, 569–590. doi:10.1353/hub.2004.0057

Lipson, M., Cheronet, O., Mallick, S., Rohland, N., Oxenham, M., Pietrusewsky, M., et al. (2018). Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science* 361, 92–95. doi:10.1126/science.aat3188

Liu, C.-C., Witonsky, D., Gosling, A., Lee, J. H., Ringbauer, H., Hagan, R., et al. (2022). Ancient genomes from the Himalayas illuminate the genetic history of Tibetans and their Tibeto-Burman speaking neighbors. *Nat. Commun.* 13 (1), 1203–1214. doi:10.1038/s41467-022-28827-2

Longmailai, M. (2015). Language and culture in northeast india and beyond, 126.*Lexical and morphological resemblances of Khasi and dimasa*

Macaulay, V., Hill, C., Achilli, A., Rengo, C., Douglas, C., Meehan, W., et al. (2005). Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* 308 (5724), 1034–1036. doi:10.1126/science.1109792

Maity, B., Nunga, S. C., and Kashyap, V. K. (2003). Genetic polymorphism revealed by 13 tetrameric and 2 pentameric STR loci in four Mongoloid tribal population. *Forensic Sci. Int.* 132 (3), 216–222. doi:10.1016/s0379-0738(02)00436-x

Majumder, P. P. (2008). Genomic inferences on peopling of south Asia. *Curr. Opin. Genet. Dev.* 18 (3), 280–284. doi:10.1016/j.gde.2008.07.003

Maples, B. K., Gravel, S., Kenny, E. E., and Bustamante, C. D. (2013). RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* 93 (2), 278–288. doi:10.1016/j.ajhg.2013.06.020

Mastana, S. S., Murry, B., Sachdeva, M. P., Das, K., Young, D., Das, M. K., et al. (2007). Genetic variation of 13 STR loci in the four endogamous tribal populations of Eastern India. *Forensic Sci. Int.* 169 (2-3), 266–273. doi:10.1016/j.forsciint.2006.03.019

Matsumura, H. (2010). Quantitative and qualitative dental-morphology at man bac. Man bac. *Excav. a Neolithic Site North. Vietnam* 33, 43–63.

Matsumura, H. (2011). *Quantitative cranio-morphology at man bac. Man bac: The excavation of a late neolithic site in northern vietnam*, 21–32.

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2 (12), e190. doi:10.1371/journal.pgen.0020190

Peiros, I. (2011). Some thoughts on the problem of the Austro-Asiatic homeland. *J. Lang. Relatsh.* 6 (1), 101–114. doi:10.31826/9781463234119-009

Pickrell, J., and Pritchard, J. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *Nat. Prec.* 1. 1. doi:10.1038/npre.2012.6956.1

Pinnow, H.-J., Kuiper, F. R. S., Greenberg, J. A. S., and Emeneau, M. (1942). The position of the Munda languages within the Austroasiatic language family. *Language* 18, 206.

Racimo, F., Sriram, S., Nielsen, R., and Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nat. Rev. Genet.* 16 (6), 359–371. doi:10.1038/nrg3936

Reddy, B. M., Langstieh, B. T., Kumar, V., Nagaraja, T., Reddy, A. N. S., Meka, A., Reddy, A. G., et al. (2007). Austro-Asiatic tribes of Northeast India provide hitherto missing genetic link between South and Southeast Asia. *PLoS One* 2 (11), e1141. doi:10.1371/journal.pone.0001141

Salter-Townshend, M., and Myers, S. (2019). Fine-scale inference of ancestry segments without prior knowledge of admixing groups. *Genetics* 212 (3), 869–889. doi:10.1534/genetics.119.302139

Shaun, P., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81 (3), 559–575. doi:10.1086/519795

Su, B., Xiao, C., Deka, R., Seielstad, M. T., Kangwanpong, D., Xiao, J., et al. (2000). Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. *Hum. Genet.* 107 (6), 582–590. doi:10.1007/s004390000406

Tagore, D., Aghakhanian, F., Naidu, R., Phipps, M. E., and Basu, A. (2021). Insights into the demographic history of Asia from common ancestry and admixture in the genomic landscape of present-day Austroasiatic speakers. *BMC Biol.* 19 (1), 61–19. doi:10.1186/s12915-021-00981-x

Thangaraj, K., Chaubey, G., Kivisild, T., Reddy, A. G., Singh, V. K., Rasalkar, A. A., et al. (2005). Reconstructing the origin of andaman islanders. *Science* 308 (5724), 996. doi:10.1126/science.1109987

Van Driem, G. (2005). *The peopling of east Asia: Putting together archaeology, linguistics and genetics*, Rouledge, England, UK, 81.Implications for population geneticists, archaeologists and prehistorians

Wang, L.-X., Lu, Y., Zhang, C., Wei, L.-H., Shi, Y., Huang, Y.-Z., et al. (2018). Reconstruction of Y-chromosome phylogeny reveals two neolithic expansions of Tibeto-Burman populations. *Mol. Genet. Genomics* 293 (5), 1293–1300. doi:10.1007/s00438-018-1461-2

Weir, B. S., and Cockerham, C. C. (1984). evolution, 1358–1370.Estimating F-statistics for the analysis of population structure

Yang, M. A., Fan, X., Sun, B., Chen, C., Lang, J., Ko, Y.-C., Tsang, C.-h., et al. (2020). Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* 369 (6501), 282–288. doi:10.1126/science.aba0909

Yu, X., and Hui, L. (2021). Origin of ethnic groups, linguistic families, and civilizations in China viewed from the Y chromosome. *Mol. Genet. Genomics.* 296 (4), 783–797. doi:10.1007/s00438-021-01794-x