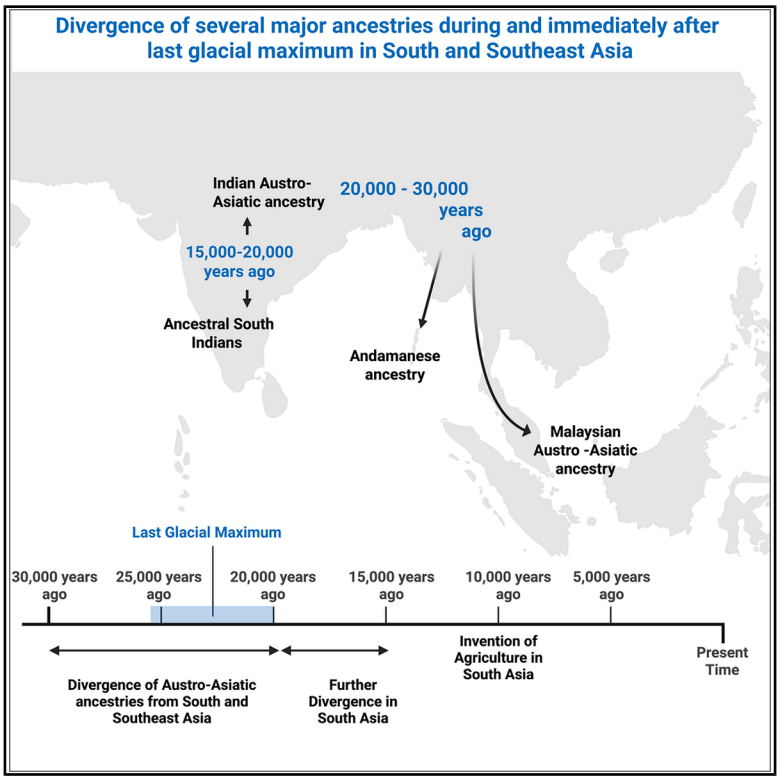
iScience

Article

Reconstruction of the genetic history of the Austro-Asiatic- and Dravidian-speaking tribal populations of South Asia

Graphical abstract



Highlights

- Tribal Dravidian and Austro-Asiatic ancestries in South Asia diverged just after LGM
- South and Southeast Asian Austro-Asiatic ancestries diverged earlier during the LGM
- Notable genetic-linguistic discordance was observed in Central Indian populations

Authors

Arghya Dey, Manisha Pal, Diptarup Nandi, Analabha Basu

Correspondence

diptarup.nandi@apu.edu.in (D.N.), ab1@nibmg.ac.in (A.B.)

In brief

Biological sciences; Human genetics

Dey et al., 2026, iScience *29*, 115241
April 17, 2026
https://doi.org/10.1016/j.isci.2026.115241

CellPress

iScience

CellPress
OPEN ACCESS

Article

Reconstruction of the genetic history of the Austro-Asiatic- and Dravidian-speaking tribal populations of South Asia

Arghya Dey,[1,2] Manisha Pal,[3] Diptarup Nandi,[4,*] and Analabha Basu[1,5,*]
[1]Biotechnology Research Innovation Council-National Institute of Biomedical Genomics, Kalyani, West Bengal 741251, India
[2]Department of Statistics, University of Calcutta, Kolkata, West Bengal, India
[3]Department of Statistics, Faculty of Science, St. Xavier's University, Kolkata, West Bengal, India
[4]School of Arts and Sciences, Azim Premji University, Bengaluru, Karnataka 562125, India
[5]Lead contact
*Correspondence: diptarup.nandi@apu.edu.in (D.N.), ab1@nibmg.ac.in (A.B.)
https://doi.org/10.1016/j.isci.2026.115241

SUMMARY

The ancestors of Austro-Asiatic- and Dravidian-speaking tribal populations are among the earliest settlers of South and Southeast Asia (S&SEA). While Austro-Asiatic speakers are distributed across S&SEA, Dravidian speakers are primarily confined to South India. Previous studies have identified the Indian Austro-Asiatic and Dravidian tribal populations to maximally represent two distinct genetic ancestral components, ancestral Austro-Asiatic (AAA) and ancestral South Indian (ASI), respectively. Leveraging the whole genome sequence (WGS) dataset from GenomeAsia 100K project, we investigated the genetic relationship of the tribal populations within the broader South Asian demographic landscape. Our analyses reveal that AAA and ASI components diverged ∼15,000–20,000 years before present (ybp), shortly after Last Glacial Maximum—a period marked by ecological shifts and regional isolation. This divergence postdates the separation (∼20,000–30,000 ybp) between the Indian (AAA) and Southeast Asian (AAM) Austro-Asiatic ancestries, indicating a deep and widespread pre-Neolithic distribution of this ancestral population across S&SEA. Additionally, recent (∼750–1500 ybp) gene flow between Central Indian Dravidian and Austro-Asiatic tribes produced notable genetic-linguistic discordances.

INTRODUCTION

Peopling of South and Southeast Asia (S&SEA) is complex. These were among the earliest regions colonized by the anatomically modern humans (AMHs) after they moved out of Africa nearly 50,000–100,000 years before present (ybp).[1–3] While the exact routes of the out-of-Africa (OoA) migration to S&SEA and the timings of their subsequent colonization remain to be determined, the hypothetical Southern exit route[4–6] has been supported more favorably by a growing body of evidence,[7,8] including some more recent challenges.[9,10] This route posits initial migrations through Ethiopia and the horn of Africa followed by a coastal trajectory that passed through S&SEA, eventually leading to Australasia.[4–6] In contrast, the Northern exit route passes through Egypt and Sinai toward Eurasia.[4,11,12] Irrespective of their precise route(s) of arrival, the earliest AMHs colonized S&SEA as early as 50,000–73,000 ybp.[13,14] Along with several lines of paleoanthropological and archaeological evidence, the expansion of the mtDNA M-haplogroup in this region attests to an early colonization event.[15–18]

The ancestors of the present-day tribal populations in South Asia (SA) that speak languages belonging to the Austro-Asiatic (AA) and Dravidian (DR) linguistic families are considered to be the initial AMH settlers of this region.[15,19–23] The AA-speaking populations are patchily distributed across SA and mainland Southeast Asia (SEA). In present-day SA, the AA speakers are found exclusively in isolated tribal populations of central, eastern, and north-eastern India. Majority of the AA speakers are, however, found in SEA, especially in Vietnam and Cambodia. Most of these populations are extensively admixed with other geographically proximate linguistic groups.[24–28] On the other hand, the DR-speaking populations are presently confined to SA. Unlike the AA languages that are spoken exclusively by tribals in SA, the DR languages are spoken by populations of all social hierarchies. They currently reside primarily in South India, with a notable exception of the Brahui people who reside in Pakistan.[29] However, they have been hypothesized to be more widespread in the past, beyond southern India, especially in northern SA, which might have included the inhabitants of Indus Valley Civilization, based on linguistic and genetic evidence.[15,29–32]

Studies using genome-wide data from South Asian populations have inferred four distinct genetic components that are present in differential compositions in the several ethnolinguistic groups inhabiting mainland SA.[33,34] The west Eurasian-related ancestral component, ancestral North Indian (ANI), was found

to show a geographic cline, with the highest representation among the Indo-European speakers of northern SA. The ancestral South Indian (ASI) and ancestral Austro-Asiatic (AAA) components were maximally represented by the DR- and AA-speaking tribal populations, respectively. Both these ethnolinguistic groups were found to completely lack the West Eurasian-related ANI component, whereas the caste populations of SA harbor a combination of these three ancestries along a north-south cline, with a higher proportion of ANI component in the north and ASI component in the south. The fourth component, ancestral Tibeto-Burman (ATB), is maximally present in the Tibeto-Burman (TB)-speaking populations inhabiting Northeast India and show an increasing west-to-east geographic cline. The AA and DR tribal populations share a unique demographic history that suggests a pre-Holocene presence in SA.[35] In fact, some of these DR-speaking tribal populations have been observed to share close genetic relationships with the Indian AA-speaking populations.[36,37] Using the Indian dataset, Basu et al.[33] also showed that in unsupervised clustering, the AAA and ASI components were the last to separate as the number of clusters (K) increased from $K = 2$ to 4. This distinction between ASI and AAA components raises the possibility of a deep, late-Pleistocene divergence within SA, predating the Neolithic transitions and suggesting long-term regional continuity, rather than recent introduction. However, the ancestral relationships between the populations predominantly harboring the AAA and ASI components remain unexplored despite their critical importance to the peopling of SA.

The ancestral relationship between the AA speakers from SA and SEA also remains to reach a consensus. Studies focusing on genome-wide datasets from SEA populations and ancient samples from Southern China have posited the origin of the earliest AA speakers in Southern China and subsequent southern spread to S&SEA around 4000 ybp.[25,38,39] On the other hand, inclusion of AA populations of SA suggests a widespread distribution of the ancestors of the AA-speaking people across mainland S&SEA and a subsequent divergence between the two around 10,500 ybp.[22] The southern migrations of East Asian farmers during the Neolithic might have led to the current isolated, patchy distribution of AA speakers in S&SEA. However, a holistic inference of the ancestral genetic relationships of the AA speakers of SA&SEA remains elusive.

In this study, we addressed these objectives by exploiting whole genome sequence (WGS) data from several SA and SEA populations that were generated by the GenomeAsia100K Consortium.[40] We specifically focused our analysis on the DR- and AA-speaking tribal populations of SA enriched in ASI and AAA components, respectively (see Tables S1 and S2 and Figure S1 for details). Our analysis also included the Malaysian AA-speaking populations of SEA to infer their relationship with the AA populations in SA. Firstly, we examined the extent of genetic relationship between the present-day AA and DR tribal populations with respect to other mainland SA populations, leveraging eleven DR tribal populations that varied in their geographic proximity to the four AA populations from SA. We examined whether environmental fluctuations—particularly those during the Last Glacial Maximum (LGM), as posited by Tagore et al.[22] along with subsequent cultural changes—played a pivotal role in shaping population splits and migrations across S&SEA. We further examined whether the genetic relationship between the AA and DR tribes is concordant with their geographical proximity or shared linguistic affiliations. Finally, we inferred the demographic history and divergence of populations representing ASI, Indian and Malaysian AA, to gain critical insights into their shared past.

RESULTS

Population structure of SA

We investigated the population structure of SA to understand how the AA and DR language-speaking tribal populations with the highest AAA and ASI proportions, respectively, are genetically related to each other, as well as with other populations of SA. Using a set of filtered SNPs (details in STAR Methods), we performed a principal component analysis (PCA) on 471 unrelated individuals from 55 mainland South Asian populations as implemented in the population genetic software toolset Eigensoft.[41] In the PCA, we excluded the Jarawa (JAR), Onge (ONG), and Nicobarese (NIC) populations from the Andaman and Nicobar Islands, as these have been shown to be genetically distant from the mainland populations, thus reducing the overall resolution of the PCA (Figure S2). The PC1-PC2 space revealed two broad clusters (Figure 1B), one including mainly the TB-speaking populations, and the other cluster with a spread along the PC2, comprising the populations belonging to Indo-European (IE), DR, and AA linguistic families. In this non-TB cluster, we observed a cline of decreasing ANI-related ancestry, which is genetically close to West Asians, Central Asians, and the Europeans, corroborating earlier findings.[33,34,40] Most of the AA and DR tribal populations clustered together, indicating their high genetic similarity within themselves when compared to other South Asian populations.

To further examine the population structure among the AA and DR tribes, we performed a second PCA with a subset of 141 unrelated individuals belonging to 4 AA tribes and 11 DR tribes (Figure 2). We observed a gradient along PC2 with the AA tribe Birhor (BIR) at one end and DR tribe Paniya (PNY) at the other one. Several individuals from the AA tribes clustered with the individuals belonging to the DR tribes from Central India. A few Kota (KTA) and Kaya Dora (KYD) individuals clustered separately due to relatively high ANI proportion, instead of AAA, in their genomes (Figure 1C).

We also performed ADMIXTURE[42] analysis to infer population structure among the 471 unrelated individuals from 55 populations of mainland SA, along with 26 unrelated individuals from 3 island populations, JAR, ONG, and NIC, which were not considered in PCA. We considered the same set of SNPs as in PCA for this analysis. At $K = 2$, we observed two distinct ancestral components: (1) ATB, maximized in the TB speakers, and (2) ANI (Figure S3). At $K = 3$, we observed an additional ancestral component maximized in the AA and DR tribal populations. The Andamanese ancestry separated from other ancestries for the first time at $K = 4$, which also corresponded to the minimum cross-validation error (CVE) of 0.25174 for $K = 4$. However, at $K = 4$, the ASI and AAA components did not separate (Figures 1C and S3). This was another line of evidence indicating genetic

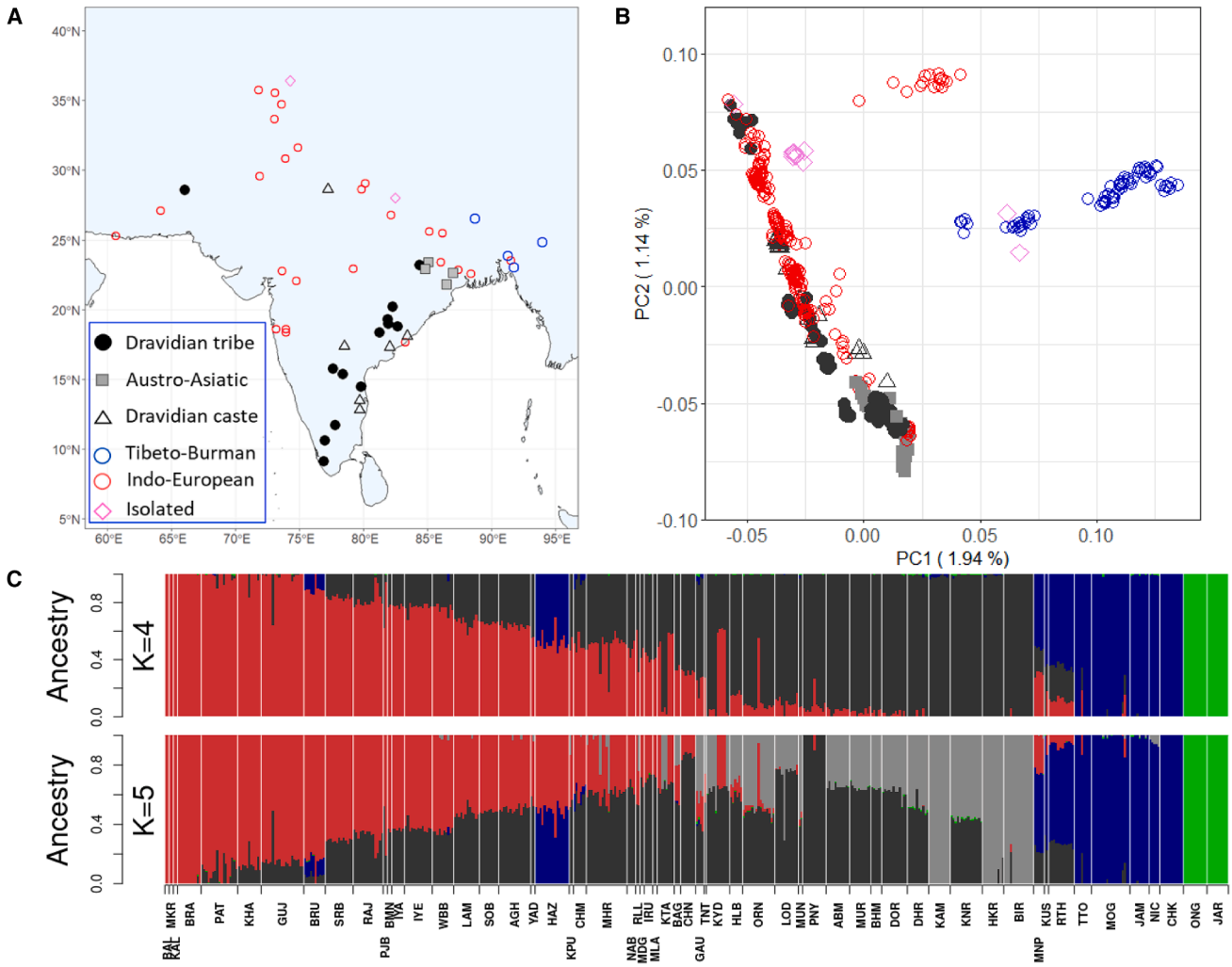


Figure 1. Population structure of SA

(A) Language distribution of mainland SA along with sampling locations, adapted from GenomeAsia100K Project.

(B) Principal component analysis (PCA) on 471 unrelated individuals from 55 populations from mainland SA. Each dot represents an individual. The color codes correspond to the linguistic groups as shown in (A).

(C) ADMIXTURE plots for 497 individuals from 58 South Asian populations (including the islanders: JAR, ONG, and NIC) at $K = 4$ and $K = 5$. Each individual is represented by a vertical line, which is further divided into colored segments. Each color represents an ancestral population, and the lengths of the colored segments represent the corresponding ancestral proportions (red, ANI; blue, ATB; black, ASI; green, Andamanese ancestry; gray, AAA). The three-letter labels at the bottom represent the corresponding populations, described in Table S1.

See also Figures S2 and S3. Both PCA and ADMIXTURE analysis suggest high genetic similarity among the AA-speaking populations and the DR-speaking tribal populations of SA.

SA, South Asia; ANI, ancestral North Indian; ATB, ancestral Tibeto-Burman; ASI, ancestral South Indian; AAA, ancestral Austro-Asiatic; AA, Austro-Asiatic; DR, Dravidian.

similarity between AA and DR tribes compared to other South Asian populations, similar to what we have asserted before in a subset of these populations.[33] Similar to the finding of Basu et al.,[33] the ASI and AAA components finally separated at $K = 5$ (CVE = 0.25300); the ASI component was maximized in DR-speaking PNY tribe, whereas the AAA component was maximized in the AA tribe BIR except one individual and Indo-European tribe Kamar (KAM). ADMIXTURE analysis, including only the AA and DR tribes, failed to reveal any discernible structure in these populations (minimum CVE corresponded to $K = 1$).

The AA and DR tribes had more recent shared ancestry between them compared with the ANI

The genetic similarity between the AA and DR tribes in PCA and ADMIXTURE analyses might be due to recent common ancestry. Hence, we wanted to understand whether the AA and DR tribes separated after their divergence from the ANI component. To answer this question, we used the *outgroup-f3* analysis,[43] with the assumption that there was no admixture between any of these tribal populations and the ANI, after they were separated from an outgroup (details in STAR Methods). We used the

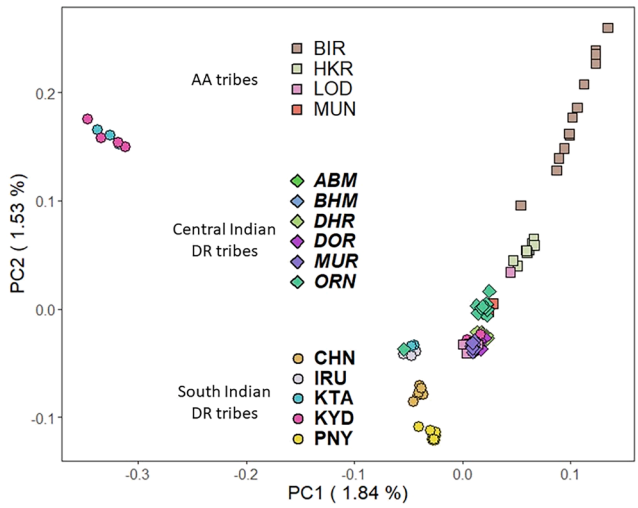


Figure 2. PCA of 141 unrelated individuals from AA- and DR-speaking tribal populations of SA

Each symbol represents an individual. The Central Indian DR tribes are labeled in bold and italic, whereas the South Indian DR tribes are labeled in bold but not italic. The AA tribes are labeled in the regular font. See Table S2 and Figure S1 for more details regarding the geographic and linguistic affiliations of the AA and DR tribal populations.

Indo-European language-speaking Pathan (PAT) population as a representative of high ANI in SA, and the African population Yoruba (YRI) as an outgroup. We observed that the *outgroup-f3* values of the form f3(YRI; PAT, an AA tribe) and f3(YRI; PAT, a DR tribe) were stochastically smaller than that of f3(YRI; an AA tribe, a DR tribe), (Kolmogorov-Smirnov test: D = 1 for both cases; p values = 1.028e-05 and 1.671e-11, respectively) (Figure 3A). This suggested that the DR tribes had more shared genetic drift with the AA tribes than their shared genetic drift with PAT. Using the Kolmogorov-Smirnov test, we found that the distributions of f3(YRI; PAT, an AA tribe) and f3(YRI; PAT, a DR tribe) were similar.

On repeating the *outgroup-f3* analysis with all populations with high ANI proportion (mean ANI proportion >50% in ADMIXTURE analysis for $K = 5$; we excluded HAZ due to high ATB representation) (Tables S3 and S4), we observed that both f3(YRI; an ANI-high population, an AA tribe) and f3(YRI; an ANI-high population, a DR tribe) were negatively correlated with ANI proportions in the corresponding ANI-high populations (correlation coefficients were −0.9298 and −0.9254, respectively) (Figure 3B). Thus, the *outgroup-f3* analyses suggest late separation of AA and DR tribes with respect to their divergence from ANI.

The ASI and AAA components separated nearly 15,000–20,000 ybp

We used joint frequency spectrums of SNPs in dadi (diffusion approximation for demographic inference)[44] to infer the demographic histories of populations with high ASI and AAA components with respect to an outgroup population and estimated the separation time between the ASI and AAA components. We considered BIR and PNY as representatives of AAA and ASI components, respectively. We considered the generation time as 25 years and the mutation rate as 1e-8/bp/generations.[45] Figure 4A shows the schematic of our 3-population demographic model.

In models with YRI as the outgroup, the estimated time of divergence between YRI and the ancestral populations of BIR and PNY (allowing no migration) was ∼64,000 ybp (SD ∼4,700 years), which corroborates well-established time period of the OoA event[1–3] when AMHs moved out of Africa and started spreading in other regions of the world. Furthermore, in accordance with various studies that suggested a severe bottleneck in the OoA founder population,[46,47] our model too estimated a severe decline in the effective population size (∼4% of that before the split) of the OoA population just after the split. Our model suggested a recovery in the effective population size of the OoA population before it split into BIR and PNY. The estimated time of the separation between BIR and PNY was nearly 17,500 ybp (SD ∼1,300 years). After the split, both BIR and PNY likely underwent bottlenecks, further corroborating an earlier finding.[35] On the other hand, the effective population size of the outgroup population YRI increased after the OoA split (Figure 4B), as expected based on previous results. We observed similar results on repeating the analysis with Hill Korwa (HKR) and Chenchu (CHN) as representative populations of AAA and ASI, respectively (Figure 4C; estimated OoA split time ∼61,000 ybp and estimated ASI-AAA split time ∼19,000 ybp) despite greater admixture in these populations compared with BIR and PNY.

With the populations BIR, PNY, and a European outgroup GBR (allowing no migration), the estimated split time between GBR and the ancestral population of BIR and PNY was ∼37,500 ybp. This split time corresponded with the Europe-Asia split, consistent with the split time estimated in other studies (around 40,000 ybp) using uniparental markers[48] and WGSs.[49–51] In this model, the estimated BIR-PNY split time was ∼15,000 ybp (Figure S10). With BIR, PNY, and Andamanese outgroup JAR (allowing no migration), the estimated split time between JAR and the ancestral populations of BIR and PNY was ∼24,500 ybp, corresponding to the split between mainland and island populations of SA.[33,52] The BIR-PNY split time was estimated to be ∼19,000 ybp (Figure S13). Thus, with various outgroups and populations representing AAA and ASI components, the AAA-ASI split times were found consistently between 15,000 and 20,000 ybp. When we considered symmetric and asymmetric migrations between populations representing high ASI and AAA proportions, the AAA-ASI split times appeared to be a bit more distant in the past, as expected. However, most of these split time estimates were still comparable (∼15,000–25,000 ybp) (Figures S4–S30; see Table S5 for the split time estimates considering various generation times and mutation rates).

We also considered similar demographic models with the outgroup YRI and the following target populations: a population having high AAA/ASI proportion and another having an ancestry corresponding to the present-day Austro-Asiatics of Malaysia (AAM), represented by two Malaysian tribal populations Kintak (KIN) and Senoi Semai (SNS). Although AA languages are majorly spoken in Vietnam and Cambodia, we considered only SNS and KIN as representative AA populations from SEA due to the following reasons: (1) Tätte et al.[37] showed that the Indian AA

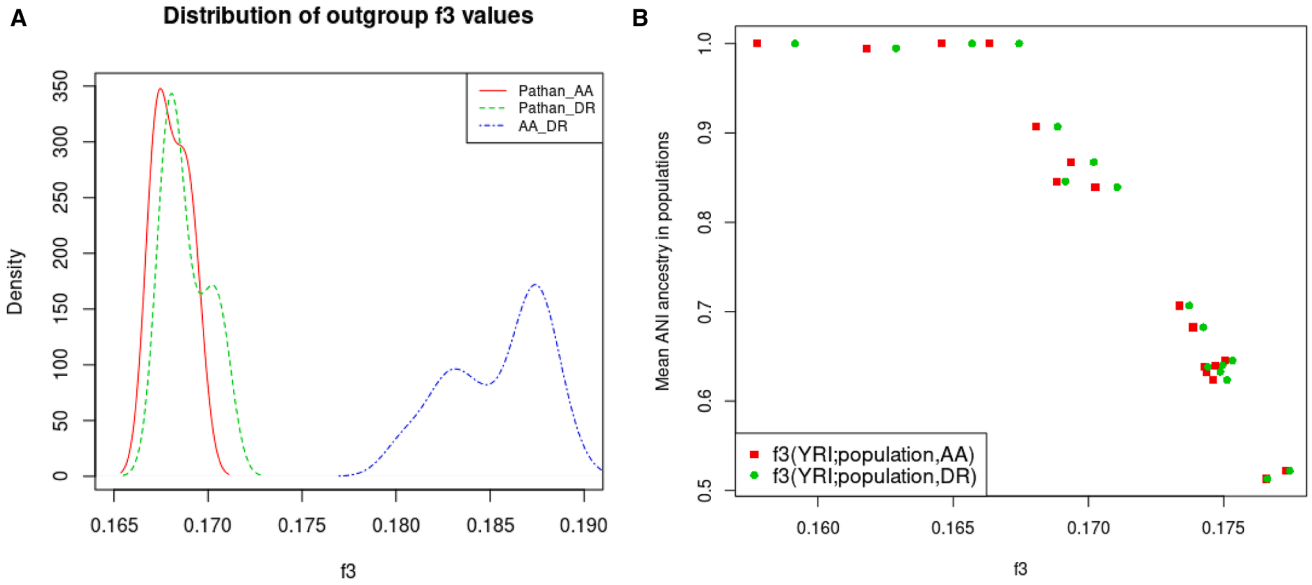


Figure 3. Shared genetic drift using *outgroup-f3* statistics

(A) Distribution of *outgroup-f3* values with the African outgroup Yoruba (YRI). Pathan (PAT) represents high ANI proportions. AA, Austro-Asiatic tribes; DR, Dravidian tribes.

(B) Association between *outgroup-f3* values and ANI proportions for populations with high ANI proportions (mean ANI proportion >50% in ADMIXTURE analysis for $K = 5$).

See also Tables S3 and S4.

populations have more IBD (identity by descent) sharing with Malaysian tribal populations than with Vietnamese and Cambodian AA, which they attributed to later East Asian admixture in Vietnam and Cambodia; (2) both SNS and KIN are less admixed compared with the populations from Vietnam and Cambodia sampled in the GenomeAsia100K dataset.[40] Allowing no migration, the estimated separation time between BIR (representing AAA component) and KIN (representing AAM component) corresponded to ~22,000 ybp (Figure 4D), whereas that between BIR and SNS, another Malaysian AA population, corresponded to ~21,000 ybp (Figure S22). The OoA split times estimated in both scenarios were comparable (~65,000 ybp and ~64,000 ybp, respectively). Similarly, the estimated separation time between PNY (representing ASI component) and KIN corresponded to ~28,500 ybp (Figure 4E), whereas that between PNY and SNS was ~31,000 ybp (Figure S25). The OoA split times were very similar (~64,000 ybp and ~62,500 ybp, respectively). We also considered scenarios where we allowed symmetric and asymmetric migrations between the populations representing ASI/AAA and AAM components. All the separation times between BIR and SNS/KIN were between 25,000 and 30,000 ybp, whereas those between PNY and SNS/KIN were between 33,000 and 37,000 ybp (Figures S4–S30 and Table S5). These results suggest that the separation between AAA and Malaysian AA ancestry predated the separation between ASI and AAA components (see Figures S4–S30 for visualizing the joint site frequency spectrums [SFS], residuals, and schematic for each demographic model used in dadi).

We validated the separation times estimated in dadi with those estimated using Relate,[53] which can estimate the within and across-group coalescence rates for pairs of groups to eventually estimate their separation times. We observed that the BIR and PNY clearly separated at ~15,000 ybp by visually inspecting the effective population size changes, corroborating our dadi result. Both BIR and PNY showed gradual separation from the Malaysian AA population KIN ~30,000 ybp, which culminated as a clear separation ~20,000 ybp. In the same time interval (~20,000–30,000 ybp), we observed that both BIR and PNY separated from another Malaysian AA population SNS and the Andamanese tribe JAR. Almost all the dadi estimates for separation times between ASI/AAA and AAM components were in this interval. Additionally, the Andamanese tribe JAR separated from the Malaysian AAs SNS and KIN at ~30,000 ybp. The OoA time was estimated in Relate by observing split times between YRI and various other populations. All these scenarios corresponded to a gradual split starting ~200,000 ybp and a clear split ~50,000 ybp, corroborating both dadi results and earlier findings (see Figure S31 for the plots corresponding to the effective population sizes estimated using Relate).

Central Indian DR tribes have more gene flow with AA tribes than other DR tribes

To understand the relationship between AA and DR tribes in terms of shared genetic drift, we generated a heatmap based on the *outgroup-f3* statistics[43] of the form f3(YRI; a DR/AA tribe, another DR/AA tribe). The dendrogram based on the f3 estimates showed two clusters, one comprising only South Indian DR tribes and the other comprising both AA tribes and Central Indian DR tribes. The cluster comprising different linguistic tribes of Central India consisted of several sub-clusters, which did not

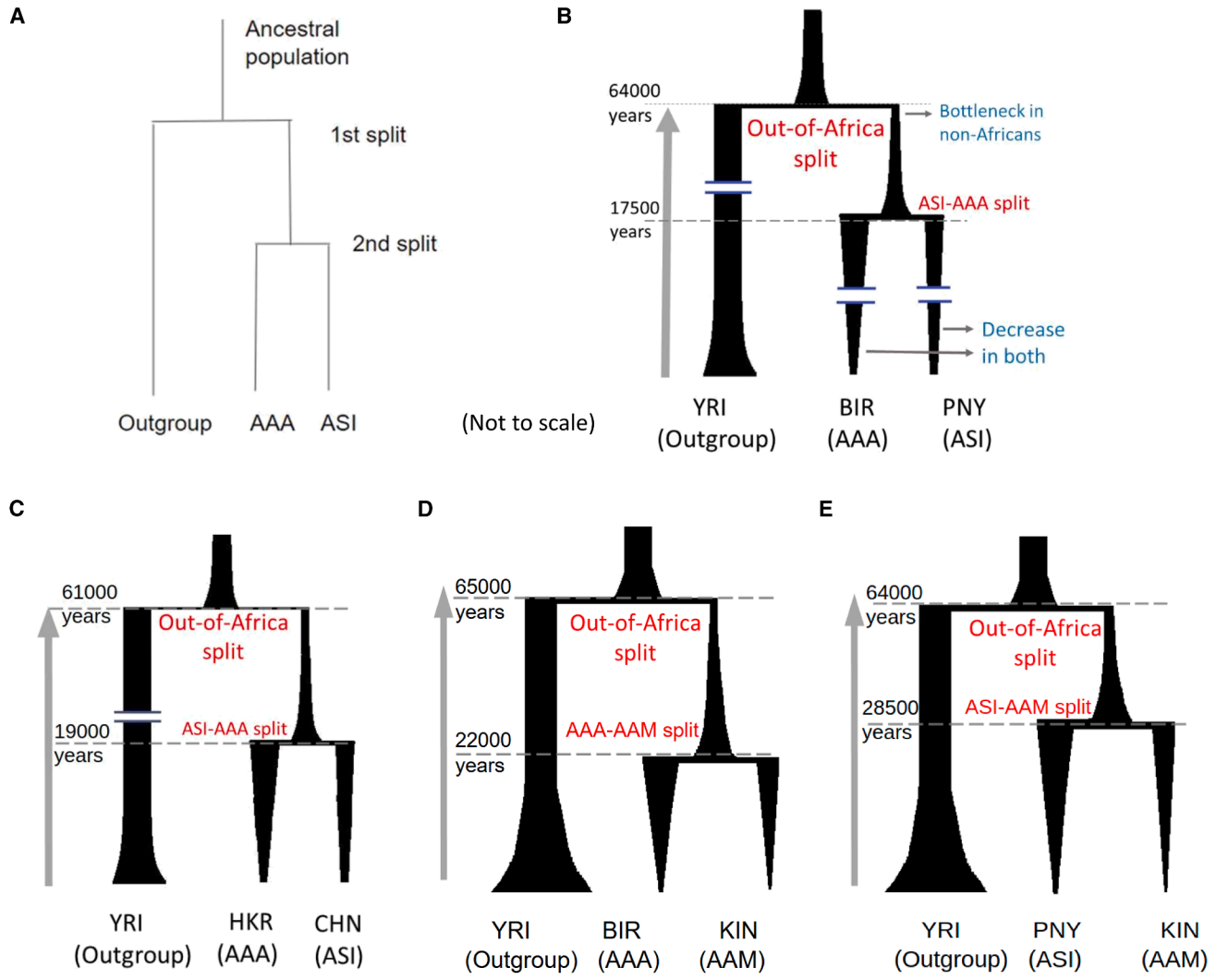


Figure 4. Demographic inference using dadi

(A) Schematic of a 3-population model in dadi (allowing no migrations).

(B) Inferred demographic history of the following populations: (1) Birhor (BIR), representing a high AAA proportion, (2) Paniya (PNY), representing a high ASI proportion, and (3) Yoruba (YRI), an African outgroup. The widths represent corresponding effective population sizes.

(C) Inferred demographic history of Hill Korwa (HKR) and Chenchu (CHN), representing AAA and ASI components, respectively, with outgroup YRI.

(D) Inferred demographic history of BIR and Malaysian AA population Kintak (KIN), with outgroup YRI.

(E) Inferred demographic history of PNY and KIN with outgroup YRI.

See Figures S4–S30 and Table S5 for more details.

clearly separate the DR and AA tribes (Figure 5A). We also observed that the AAA component (at $K = 5$) was highly correlated (correlation coefficient = 0.859) with the latitude for the DR tribes, showing that Central Indian DR tribes had more AAA component than the South Indian DR tribes in general (Figure S32).

To understand the phylogenetic relationship of the AA and DR tribes, we constructed a maximum likelihood tree by using Treemix,[54] considering 11 DR tribal populations and 4 AA populations, and a West African population YRI as the outgroup. The simplest model considering no migration showed that the AA tribal populations formed a clade along with DR tribe Oraon (ORN). The Central Indian DR tribes clustered closer to the AA clade (Figure 5B), corroborating our *outgroup-f3* analysis results. We used OptM[55] to estimate the optimum number of migration edges (m) in Treemix (Figure S33). At optimum m = 4, the tree topology was mostly unaltered, except Lodha (LOD) separating from the AA clade (Figure 5C). The strongest gene flow in terms of migration weights was estimated to be from ORN to LOD.

We further used D-statistics[56–58] of the form D(BIR, PNY; AA/DR tribe, YRI) to formally test for gene flow, taking the AA tribe BIR and the South Indian DR tribe PNY as proxies for AAA and ASI components, respectively. The D-statistics of the form D(BIR, PNY; Central Indian DR tribe, YRI) were all positive, and

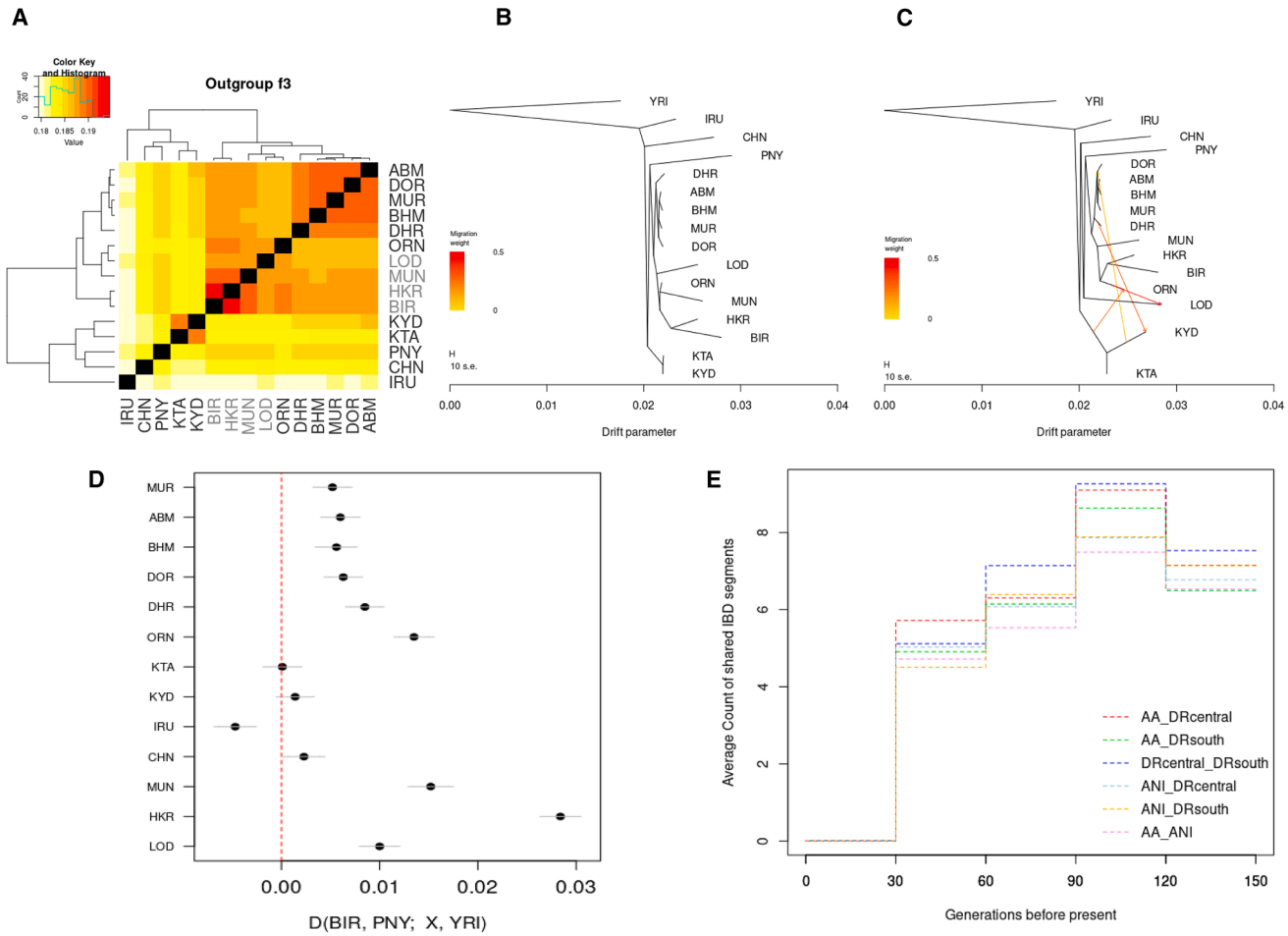


Figure 5. Gene flow between AA and DR tribal populations

(A) Heatmap of outgroup-f3 statistics for AA and DR tribal populations, with outgroup YRI. The population names in gray correspond to the AA tribes, whereas the population names in black represent the DR tribes. The corresponding dendrogram reveals broadly 2 clusters: one containing only the South Indian DR tribes, and another containing AA tribes and Central Indian DR tribes.

(B) Maximum likelihood tree generated by Treemix for AA and DR tribal populations, with outgroup YRI, allowing for no migrations (m = 0).

(C) Maximum likelihood tree generated by Treemix for AA and DR tribal populations, with outgroup YRI, for m = 4, which corresponds to the optimum value of m, as estimated using OptM analysis. See Figure S33 for more details.

(D) Distribution of D-statistics for various AA and DR tribes along with the error bars (± standard error).

(E) Average number of IBD segments shared between two populations in various intervals.

AA, Austro-Asiatic tribes; DRcentral, Central Indian DR tribes; DRsouth, South Indian DR tribes; ANI, populations with high ANI proportions (mean ANI proportion >90% in ADMIXTURE analysis for $K = 5$).

the corresponding Z scores were significant ($Z > 3$), suggesting that the Central Indian DR tribes had more gene flow with BIR than PNY (Figure 5D). For South Indian DR tribes, the Z scores corresponding to D(BIR, PNY; South Indian DR tribe, YRI) were not significant ($|Z| < 3$), as expected based on earlier results.

We used hap-IBD[59] to estimate IBD sharing between the AA and DR tribal populations, which can provide evidence for recent gene flow, using a phased dataset consisting of ~55 million SNPs. Longer shared IBD segments suggest more recent gene flow. We restricted the estimated lengths of IBD segments to ≥1 cM.[60] The length of 1 cM IBD segment corresponds to 150 generations[61] or ~3,750 years (generation time = 25 years). Hence, our IBD analysis corresponds to a time interval that spans ~3,750 ybp to the present. We grouped several study populations based on linguistic affiliation, geography, and/or ancestry proportions. We enumerated the average number of IBD segments shared between the individuals from different groups in various time intervals, each spanning 30 generations (750 years) (see STAR Methods). As expected, the AA and DR tribes had overall higher IBD sharing among themselves than with populations having high ANI proportions (mean ANI proportion >90% in ADMIXTURE analysis for $K = 5$) (Tables 1 and S3). In the Treemix analysis, the strongest signature of gene flow was also observed from the Central Indian DR tribe ORN to the AA tribe LOD. It might have occurred 60–90 generations ago (~1,500–2,250 ybp) considering higher IBD sharing between

Table 1. IBD sharing among various pairs of groups

Pair of groups	Average IBD shared (in cM)
AA, ANI	43.33789
ANI, DRsouth	45.52776
ANI, DRcentral	46.18654
AA, DRsouth	46.82481
AA, DRcentral	51.05741
DRcentral, DRsouth	51.50104

ANI, populations having high ANI representation (mean ANI ancestry proportion >90% in ADMIXTURE analysis for K = 5); AA, Austro-Asiatic tribes; DRsouth, South Indian Dravidian tribes; DRcentral, Central Indian Dravidian tribes.

LOD and ORN relative to any other AA population at that time interval. As the Central Indian DR tribe ORN formed a clade with AA tribes in the maximum likelihood tree generated by Treemix, we did not group it with other Central Indian DR tribes in all our IBD analyses. Between 60 and 150 generations before present, the IBD sharing was higher between Central Indian DR tribes and South Indian DR tribes than that between AA and Central Indian DR tribes. However, more recently between 30 and 60 generations before present, we estimated higher IBD sharing between the AA and DR tribes of Central India. This suggests that the recent gene flow has played a significant role in the Central Indian DR tribes being genetically closer to AA tribes, rather than their South Indian counterparts (Figure 5E).

DISCUSSION

Our results support a model of early and enduring human settlement in SA, grounded in a detailed analysis of autosomal DNA and corroborating with previously published uniparental genetic evidence. The divergence between the ASI and AAA lineages around 15,000–20,000 ybp—shortly after the LGM—indicates that these ancestries were shaped by late Pleistocene processes well before the spread of agriculture or Neolithic migrations. The deep-time ancestral similarity is mirrored in mtDNA patterns, particularly the widespread and high-frequency presence of haplogroup M among both AA- and DR-speaking tribal groups.[15] Haplogroup M, which represents one of the earliest maternal lineages outside Africa, is thought to have expanded rapidly across S&SEA following the OoA dispersal,[15–18] and its significantly high prevalence in these populations underscores their likely descent from some of the region's earliest AMHs. The concordance between deep autosomal divergence and early maternal lineage expansions points to long-term regional continuity, with these tribal populations preserving genetic signals of early settlements that are largely unaltered by the West Eurasian-related (ANI) or East Eurasian-related (ATB) admixtures seen in other South Asian groups. In addition, we observed that: (1) among the fifteen tribal populations used in our dataset, BIR and PNY, which maximally represented the distinct AAA and ASI components, respectively, among all SA populations, were genetically more similar relative to all other linguistic groups despite no admixture between the two populations, and (2) the population tree-based analyses corroborated the shared recent ancestry between the AA and DR tribes compared to the Indo-European-speaking populations and further confirmed a lack of gene flow between the BIR and PNY populations. However, the other geographically proximate tribal populations speaking AA and DR languages were found to be admixed with varying proportions of ASI and AAA components. Thus, the proto-ancestors of the AAA and ASI reflect the persistence of ancient, pre-Neolithic population structure in SA, rather than the outcome of recent external introductions.

The estimated time of divergence between the ASI and AAA components, around 15,000–20,000 ybp, is central to many of the inferences drawn in our study. The congruence of the estimated split times using two complementary approaches, SFS and genome-wide genealogies, provides confidence and adds validity. These estimates remained consistent under alternative modeling conditions, including variations in outgroup selection and reference South Asian populations, underscoring the robustness of our inference. Genetic, paleoanthropological, and archaeological lines of evidence have established the presence of AMH in SA for at least 40,000 years.[13,15] Our findings also resonate with recent genomic studies suggesting that the majority of South Asian genetic ancestry derives from a single OoA lineage dating to ~50,000 ybp,[62] further supporting a model of long-term regional continuity. In terms of continuity for a larger geophysical and temporal context, our estimates place the divergence of the AAM population, now found in Malaysian AA populations, between 20,000 and 30,000 ybp, preceding the ASI-AAA split, consistent with earlier results and suggesting that the Indian and Southeast Asian branches of AA had already diverged prior to the emergence of distinct SA ancestries. Together, these lines of evidence affirm that the genetic structure observed among present-day DR and AA tribal populations preserves a legacy of late Pleistocene differentiation, independent of the more recent waves of Neolithic or historical migration that reshaped the genetic landscape of much of S&SEA.

The results of this study suggest that the ancestral populations of ASI, AAA, and AAM were more widely distributed across mainland S&SEA. These ancestral populations could potentially represent the first AMH settlers of S&SEA. Previous studies have already established the presence of AMHs in these regions, which could have been colonized following a "southern" coastal route since OoA migration. Our finding predates the previous estimate of AAA-AAM split (10,500 years) proposed by Tagore et al.[22] The older estimates in this study could be attributed to the utilization of WGS data, which offer enhanced demographic inference capabilities compared with previously employed genotype-array data. Despite the difference in the divergence dates, both studies provide support to a widespread distribution of the ancestral tribal population across S&SEA. In fact, using ancient DNA from SEA,[38,63] Tagore et al.[22] showed that the pre-Neolithic Hòabìnhian hunter-gatherers had strong genetic affinity with the Indian AA populations. This view stands in contrast to an alternative model that posits a more recent origin of the AA ancestry in Southern China and its subsequent spread to mainland SEA and speculatively to SA.[25,38,39] It is difficult to reconcile with this model as it includes a very limited representation of South Asian AA-speaking populations, if at all, which can severely constrain a more holistic inference with respect to both the models. Using

Y chromosomal haplogroups, it has also been suggested that the AA speakers of India are derived from dispersal from SEA during the last 5,000 years,[36,64] with a possibility of using maritime routes.[65] However, the proposed AA dispersal in both overland[36] and maritime models[65] suggests a late separation of Indian AA branch Munda relative to other major AA branches of S&SEA,[66] in contradiction to the linguistic phylogeny of the AA languages.[67] The linguistic phylogeny supports an early linguistic separation of the Indian AA branch Munda from the Khasi-Aslian branches including Mon-Khmer and Khasi-Pakanic. On the other hand, some linguistic studies support a late separation of Munda from other major AA lineages, suggesting the origin of Munda languages as a result of creolization during the introduction of AA languages to the non-AA speakers in the Indian subcontinent.[66] We acknowledge that our analyses cannot definitively resolve the linguistic debates regarding the position of Munda in the phylogeny of the AA language family, and the linguistic separations can occur irrespective of the separation of genetic ancestries. Hence, our genetic inferences are compatible with both scenarios, albeit more aligned with the phylogeny of the AA languages.[67]

Our inferred split time between the South and Southeast Asian ancestries corresponds exactly to the LGM. During the same period, we also detected divergence of the island Andamanese ancestry from the mainland AAA and AAM components. While the LGM has often been associated with land connectivity via lowered sea levels that may have facilitated migration, paleoanthropological and geophysical studies also emphasize widespread habitat fragmentation, leading to human populations retreating into localized refugia in response to colder and more arid conditions.[68] Gavashelishvili et al.[69] suggested that humans mostly avoided dense forest cover during LGM, and they provided a map of LGM biomes, which showed that East and Northeast India had dense forest cover extending to SEA during LGM. We speculate that such ecological compartmentalization could have triggered initial genetic separations between the ancestral groups in SA and SEA. As climatic conditions changed post-LGM, characterized by increased temperatures and rising sea levels,[70,71] it created more widespread habitable conditions, facilitating human migrations to previously unoccupied regions.[72] Localized expansions and secondary migrations within SA could have driven the subsequent divergence of AAA and ASI components. These patterns collectively reflect how late Pleistocene climatic fluctuations shaped the demographic trajectories of some of SA's most deeply rooted populations.

It is important to emphasize that genetic proximity and linguistic affiliation do not always co-evolve and may reflect separate, temporally staggered processes of demographic and cultural transformation. The gene flow patterns, as inferred from several population genetic analyses deployed in this study, demonstrate a genetic-linguistic discordance among the Central Indian DR tribes. Despite using DR languages, these tribal populations were genetically closer to their neighboring AA-speaking populations than to the South Indian DR tribal populations. While our analyses do not allow for conclusive inferences on the linguistic identities of populations with ASI and AAA components at the time of their divergence or those of their ancestral population, the possibility of language replacement in at least a few of these tribal populations cannot be ruled out. However, it requires further attestation through linguistic studies to reach a definitive conclusion regarding language replacement. The DR languages are much more widely distributed across all socio-cultural hierarchies in present-day SA, comprising both castes and tribes, unlike the AA languages that are restricted exclusively to tribal populations. The spread of DR languages after the separation of ASI and AAA components through cultural diffusion from more dominant agrarian populations demands more elaborate considerations.

The presence of several agrarian DR kingdoms in Southern India since at least 3rd century BCE with military forces and the ensuing armed expansions suggest the possibility of dominance over weaker tribal populations.[73] For instance, the DR language-speaking ORN tribe, which clustered in the same clade along with the AA tribes in the maximum likelihood population tree in this study, has been hypothesized to undergo language replacement.[74,75] Bayesian phylogenetic analysis of the DR language family estimated the origin and spread of DR languages ~4500 ybp,[76] which is also supported by earlier archaeological and archaeo-botanical lines of evidence.[77–79] Moreover, our results based on IBD sharing indicate that the central Indian DR tribes were, in fact, genetically closer to the South Indian DR tribes between ~1,500 and 3,750 ybp and, only more recently, ~750–1500 ybp show higher genetic similarity with the AA tribes, possibly due to greater gene flow because of geographical proximity. Our findings, along with previous linguistic and archaeological lines of evidence, therefore, suggest that the tribal populations with a high ASI proportion may have started using DR languages long after their separation from the AAA populations. Among these, the Central Indian tribal populations speaking DR languages admixed with the geographically proximate AA tribal populations more recently, leading to the higher genetic similarity between these populations, over and above the similarity arising due to shared common ancestry.

Together, our results underscore the antiquity and continuity of human settlement in SA, shaped by pre-Neolithic population structure and more recent episodes of gene flow with a possibility of language shift. These patterns point to a complex interplay of deep ancestry and recent cultural processes in shaping the genomic and linguistic diversity of populations in the subcontinent. This study calls for a broader, regionally inclusive approach to understanding prehistoric human migrations and cultural transitions in SA and SEA by integrating genomic, linguistic, and archaeological research.

Limitations of the study

This study has some limitations, which should be considered while we interpret the results and plan for future studies. The major limitation remains the overall under-representation of SA and SEA populations in global genomic datasets. Within this framework of limited data availability, indigenous tribal populations like the AA and DRs, despite their critical importance for reconstructing population histories, are even less represented. Our study had to rely on high-coverage, whole-genome sequencing for the power of resolution that is indispensable for the inferences. This further constrains inclusion due to cost and data availability, rendering such analyses inherently selective and

difficult to scale across diverse tribal populations. Moreover, the scarcity of ancient DNA data of meaningful antiquity from SA precludes direct temporal validation of inferred demographic scenarios. Consequently, the interpretations presented here should be considered provisional and will require independent substantiation or refutation through complementary archaeological and linguistic evidence.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Analabha Basu (ab1@nibmg.ac.in).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data, accessible at https://doi.org/10.1038/s41586-019-1793-z.
- All original code has been deposited at Zenodo and is publicly available at https://zenodo.org/records/18533055, as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

A.D. was funded by DST-INSPIRE (DST/INSPIRE fellowship/[IF180204]). During the initial phases of this work, D.N. was supported by the National Post Doctoral Fellowship, Science and Engineering Research Board, Department of Science and Technology, Government of India (PDF/2018/004098). We acknowledge Dr. Saikat Chakraborty for his invaluable suggestions.

AUTHOR CONTRIBUTIONS

A.D., M.P., D.N., and A.B. designed research; A.D. analyzed data and performed research; A.D., D.N., and A.B. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Dataset and quality control
  - Population structure analysis
  - ADMIXTURE analysis
  - *Outgroup-f3* analysis
  - Dadi
  - Treemix analysis
  - D-statistics
  - Hap-IBD analysis
  - Relate
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2026.115241.

Received: July 29, 2025
Revised: October 29, 2025
Accepted: March 2, 2026
Published: March 5, 2026

REFERENCES

1. Cann, R.L., Stoneking, M., and Wilson, A.C. (1987). Mitochondrial DNA and human evolution. Nature *325*, 31–36. https://doi.org/10.1038/325031a0.
2. Harpending, H.C., Sherry, S.T., Rogers, A.R., and Stoneking, M. (1993). The Genetic Structure of Ancient Human Populations. Curr. Anthropol. *34*, 483–496. https://doi.org/10.1086/204195.
3. Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493–496. https://doi.org/10.1038/nature10231.
4. Lahr, M.M., and Foley, R. (1994). Multiple dispersals and modern human origins. Evol. Anthropol. *3*, 48–60. https://doi.org/10.1002/evan.1360030206.
5. Cavalli-Sforza, L.L., Menozzi, P., and Piazza, A. (1996). The History and Geography of Human Genes (Princeton University Press). https://doi.org/10.1515/9780691187266.
6. Quintana-Murci, L., Semino, O., Bandelt, H.-J., Passarino, G., McElreavey, K., and Santachiara-Benerecetti, A.S. (1999). Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat. Genet. *23*, 437–441. https://doi.org/10.1038/70550.
7. Forster, P. (2004). Ice Ages and the mitochondrial DNA chronology of human dispersals: a review. Philos. Trans. R. Soc. Lond. B Biol. Sci. *359*, 255–264. https://doi.org/10.1098/rstb.2003.1394.
8. Underhill, P.A., Passarino, G., Lin, A.A., Shen, P., Mirazón Lahr, M., Foley, R.A., Oefner, P.J., and Cavalli-Sforza, L.L. (2001). The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann. Hum. Genet. *65*, 43–62. https://doi.org/10.1046/j.1469-1809.2001.6510043.x.
9. Pagani, L., Schiffels, S., Gurdasani, D., Danecek, P., Scally, A., Chen, Y., Xue, Y., Haber, M., Ekong, R., Oljira, T., et al. (2015). Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians. Am. J. Hum. Genet. *96*, 986–991. https://doi.org/10.1016/j.ajhg.2015.04.019.
10. Wall, J.D. (2017). Inferring Human Demographic Histories of Non-African Populations from Patterns of Allele Sharing. Am. J. Hum. Genet. *100*, 766–772. https://doi.org/10.1016/j.ajhg.2017.04.002.
11. Forster, P., and Matsumura, S. (2005). Did Early Humans Go North or South? Science *308*, 965–966. https://doi.org/10.1126/science.1113261.
12. Reed, F.A., and Tishkoff, S.A. (2006). African human diversity, origins and migrations. Curr. Opin. Genet. Dev. *16*, 597–605. https://doi.org/10.1016/j.gde.2006.10.008.
13. Mellars, P., Gori, K.C., Carr, M., Soares, P.A., and Richards, M.B. (2013). Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. Proc. Natl. Acad. Sci. USA *110*, 10699–10704. https://doi.org/10.1073/pnas.1306043110.
14. Westaway, K.E., Louys, J., Awe, R.D., Morwood, M.J., Price, G.J., Zhao, J.-x., Aubert, M., Joannes-Boyau, R., Smith, T.M., Skinner, M.M., et al. (2017). An early modern human presence in Sumatra 73,000–63,000 years ago. Nature *548*, 322–325. https://doi.org/10.1038/nature23452.
15. Basu, A., Mukherjee, N., Roy, S., Sengupta, S., Banerjee, S., Chakraborty, M., Dey, B., Roy, M., Roy, B., Bhattacharyya, N.P., et al. (2003). Ethnic India: A Genomic View, With Special Reference to Peopling and Structure. Genome Res. *13*, 2277–2290. https://doi.org/10.1101/gr.1413403.
16. Chandrasekar, A., Kumar, S., Sreenath, J., Sarkar, B.N., Urade, B.P., Mallick, S., Bandopadhyay, S.S., Barua, P., Barik, S.S., Basu, D., et al. (2009). Updating Phylogeny of Mitochondrial DNA Macrohaplogroup M in India: Dispersal of Modern Human in South Asian Corridor. PLoS One *4*, e7447. https://doi.org/10.1371/journal.pone.0007447.

17. Passarino, G., Semino, O., Bernini, L.F., and Santachiara-Benerecetti, A.S. (1996). Pre-Caucasoid and Caucasoid genetic features of the Indian population, revealed by mtDNA polymorphisms. Am. J. Hum. Genet. *59*, 927–934.

18. Rajkumar, R., Banerjee, J., Gunturi, H.B., Trivedi, R., and Kashyap, V.K. (2005). Phylogeny and antiquity of M macrohaplogroup inferred from complete mt DNA sequence of Indian specific lineages. BMC Evol. Biol. *5*, 26. https://doi.org/10.1186/1471-2148-5-26.

19. Kumar, S., Padmanabham, P.B.S.V., Ravuri, R.R., Uttaravalli, K., Koneru, P., Mukherjee, P.A., Das, B., Kotal, M., Xaviour, D., Saheb, S.Y., and Rao, V.R. (2008). The earliest settlers' antiquity and evolutionary history of Indian populations: evidence from M2 mtDNA lineage. BMC Evol. Biol. *8*, 230. https://doi.org/10.1186/1471-2148-8-230.

20. Pattanayak, D.P. (1998). The language heritage of India. In The Indian human heritage, pp. 95–99.

21. Risley, H. (1915). The People of India, Calcutta, Thacker. Preprint at (Aufl), p. 2.

22. 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*, 61. https://doi.org/10.1186/s12915-021-00981-x.

23. Thapar, R. (1966). A History of India: Volume 1. Preprint at (Penguin Books).

24. Liu, D., Duong, N.T., Ton, N.D., Van Phong, N., Pakendorf, B., Van Hai, N., and Stoneking, M. (2020). Extensive Ethnolinguistic Diversity in Vietnam Reflects Multiple Sources of Genetic Diversity. Mol. Biol. Evol. *37*, 2503–2519. https://doi.org/10.1093/molbev/msaa099.

25. Kutanan, W., Liu, D., Kampuansai, J., Srikummool, M., Srithawong, S., Shoocongdej, R., Sangkhano, S., Ruangchai, S., Pittayaporn, P., Arias, L., and Stoneking, M. (2021). Reconstructing the Human Genetic History of Mainland Southeast Asia: Insights from Genome-Wide Data from Thailand and Laos. Mol. Biol. Evol. *38*, 3459–3477. https://doi.org/10.1093/molbev/msab124.

26. Larena, M., Sanchez-Quinto, F., Sjödin, P., McKenna, J., Ebeo, C., Reyes, R., Casel, O., Huang, J.-Y., Hagada, K.P., Guilay, D., et al. (2021). Multiple migrations to the Philippines during the last 50,000 years. Proc. Natl. Acad. Sci. USA *118*, e2026132118. https://doi.org/10.1073/pnas.2026132118.

27. He, Y., Zhang, X., Peng, M.-S., Li, Y.-C., Liu, K., Zhang, Y., Mao, L., Guo, Y., Ma, Y., Zhou, B., et al. (2025). Genome diversity and signatures of natural selection in mainland Southeast Asia. Nature *643*, 417–426. https://doi.org/10.1038/s41586-025-08998-w.

28. Wang, M., Huang, Y., Liu, K., Wang, Z., Zhang, M., Yuan, H., Duan, S., Wei, L., Yao, H., Sun, Q., et al. (2024). Multiple Human Population Movements and Cultural Dispersal Events Shaped the Landscape of Chinese Paternal Heritage. Mol. Biol. Evol. *41*, msae122. https://doi.org/10.1093/molbev/msae122.

29. Thapar, R. (2003). The Penguin History of Early India: From the Origins to AD 1300 (Penguin Books India).

30. Parpola, A. (2010). A Dravidian Solution to the Indus Script Problem (Central Institute of Classical Tamil).

31. Parpola, A. (2015). The Roots of Hinduism: The Early Aryans and the Indus Civilization (Oxford University Press).

32. Ansumali Mukhopadhyay, B. (2021). Ancestral Dravidian languages in Indus Civilization: ultraconserved Dravidian tooth-word reveals deep linguistic ancestry and supports genetics. Hum. Soc. Sci. Commun. *8*, 193. https://doi.org/10.1057/s41599-021-00868-w.

33. 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. USA *113*, 1594–1599. https://doi.org/10.1073/pnas.1513197113.

34. Reich, D., Thangaraj, K., Patterson, N., Price, A.L., and Singh, L. (2009). Reconstructing Indian population history. Nature *461*, 489–494. https://doi.org/10.1038/nature08365.

35. Chakraborty, S., and Basu, A. (2020). Whole genomes reveal severe bottleneck among Asian hunter-gatherers following the invention of agriculture. Preprint at bioRxiv. https://doi.org/10.1101/2020.06.25.170308.

36. Chaubey, G., Metspalu, M., Choi, Y., Mägi, R., Romero, I.G., Soares, P., van Oven, M., Behar, D.M., Rootsi, S., Hudjashov, G., et al. (2011). Population Genetic Structure in Indian Austroasiatic Speakers: The Role of Landscape Barriers and Sex-Specific Admixture. Mol. Biol. Evol. *28*, 1013–1024. https://doi.org/10.1093/molbev/msq288.

37. Tätte, K., Pagani, L., Pathak, A.K., Kõks, S., Ho Duy, B., Ho, X.D., Sultana, G.N.N., Sharif, M.I., Asaduzzaman, M., Behar, D.M., et al. (2019). The genetic legacy of continental scale admixture in Indian Austroasiatic speakers. Sci. Rep. *9*, 3818. https://doi.org/10.1038/s41598-019-40399-8.

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

39. Wang, T., Yang, M.A., Zhu, Z., Ma, M., Shi, H., Speidel, L., Min, R., Yuan, H., Jiang, Z., Hu, C., et al. (2025). Prehistoric genomes from Yunnan reveal ancestry related to Tibetans and Austroasiatic speakers. Science *388*, eadq9792. https://doi.org/10.1126/science.adq9792.

40. Wall, J.D., Stawiski, E.W., Ratan, A., Kim, H.L., Kim, C., Gupta, R., Suryamohan, K., Gusareva, E.S., Purbojati, R.W., Bhangale, T., et al. (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature *576*, 106–111. https://doi.org/10.1038/s41586-019-1793-z.

41. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909. https://doi.org/10.1038/ng1847.

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

43. Raghavan, M., Skoglund, P., Graf, K.E., Metspalu, M., Albrechtsen, A., Moltke, I., Rasmussen, S., Stafford, T.W., Jr., Orlando, L., Metspalu, E., et al. (2014). Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature *505*, 87–91. https://doi.org/10.1038/nature12736.

44. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genet. *5*, e1000695. https://doi.org/10.1371/journal.pgen.1000695.

45. 1000 Genomes Project Consortium; Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., McVean, G.A., Donnelly, P., et al. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073. https://doi.org/10.1038/nature09534.

46. Marth, G.T., Czabarka, E., Murvai, J., and Sherry, S.T. (2004). The Allele Frequency Spectrum in Genome-Wide Human Variation Data Reveals Signals of Differential Demographic History in Three Large World Populations. Genetics *166*, 351–372. https://doi.org/10.1534/genetics.166.1.351.

47. Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., and Lander, E.S. (2001). Linkage disequilibrium in the human genome. Nature *411*, 199–204. https://doi.org/10.1038/35075590.

48. Tateno, Y., Komiyama, T., Katoh, T., Munkhbat, B., Oka, A., Haida, Y., Kobayashi, H., Tamiya, G., and Inoko, H. (2014). Divergence of East Asians and Europeans Estimated Using Male- and Female-Specific Genetic Markers. Genome Biol. Evol. *6*, 466–473. https://doi.org/10.1093/gbe/evu027.

49. Malaspinas, A.-S., Westaway, M.C., Muller, C., Sousa, V.C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J.Y., Crawford, J.E., et al. (2016). A genomic history of Aboriginal Australia. Nature *538*, 207–214. https://doi.org/10.1038/nature18299.

50. Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. Nat. Genet. *46*, 919–925. https://doi.org/10.1038/ng.3015.

51. Yang, M.A., Gao, X., Theunert, C., Tong, H., Aximu-Petri, A., Nickel, B., Slatkin, M., Meyer, M., Pääbo, S., Kelso, J., and Fu, Q. (2017). 40,000-Year-Old Individual from Asia Provides Insight into Early Population Structure in Eurasia. Curr. Biol. *27*, 3202–3208.e9. https://doi.org/10.1016/j.cub.2017.09.030.

52. Mondal, M., Casals, F., Xu, T., Dall'Olio, G.M., Pybus, M., Netea, M.G., Comas, D., Laayouni, H., Li, Q., Majumder, P.P., and Bertranpetit, J. (2016). Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. Nat. Genet. *48*, 1066–1070. https://doi.org/10.1038/ng.3621.

53. Speidel, L., Forest, M., Shi, S., and Myers, S.R. (2019). A method for genome-wide genealogy estimation for thousands of samples. Nat. Genet. *51*, 1321–1329. https://doi.org/10.1038/s41588-019-0484-x.

54. Pickrell, J.K., and Pritchard, J.K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS Genet. *8*, e1002967. https://doi.org/10.1371/journal.pgen.1002967.

55. Fitak, R.R. (2021). *OptM* : estimating the optimal number of migration edges on population trees using *Treemix*. Biol. Methods Protoc. *6*, bpab017. https://doi.org/10.1093/biomethods/bpab017.

56. Durand, E.Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for Ancient Admixture between Closely Related Populations. Mol. Biol. Evol. *28*, 2239–2252. https://doi.org/10.1093/molbev/msr048.

57. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., et al. (2010). A Draft Sequence of the Neandertal Genome. Science *328*, 710–722. https://doi.org/10.1126/science.1188021.

58. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient Admixture in Human History. Genetics *192*, 1065–1093. https://doi.org/10.1534/genetics.112.145037.

59. Zhou, Y., Browning, S.R., and Browning, B.L. (2020). A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. Am. J. Hum. Genet. *106*, 426–437. https://doi.org/10.1016/j.ajhg.2020.02.010.

60. Browning, S.R., and Browning, B.L. (2020). Probabilistic Estimation of Identity by Descent Segment Endpoints and Detection of Recent Selection. Am. J. Hum. Genet. *107*, 895–910. https://doi.org/10.1016/j.ajhg.2020.09.010.

61. Baharian, S., Barakatt, M., Gignoux, C.R., Shringarpure, S., Errington, J., Blot, W.J., Bustamante, C.D., Kenny, E.E., Williams, S.M., Aldrich, M.C., and Gravel, S. (2016). The Great Migration and African-American Genomic Diversity. PLoS Genet. *12*, e1006059. https://doi.org/10.1371/journal.pgen.1006059.

62. Kerdoncuff, E., Skov, L., Patterson, N., Banerjee, J., Khobragade, P., Chakrabarti, S.S., Chakrawarty, A., Chatterjee, P., Dhar, M., Gupta, M., et al. (2025). 50,000 years of evolutionary history of India: Impact on health and disease variation. Cell *188*, 3389–3404.e6. https://doi.org/10.1016/j.cell.2025.04.027.

63. McColl, H., Racimo, F., Vinner, L., Demeter, F., Gakuhari, T., Moreno-Mayar, J.V., van Driem, G., Gram Wilken, U., Seguin-Orlando, A., de la Fuente Castro, C., et al. (2018). The prehistoric peopling of Southeast Asia. Science *361*, 88–92. https://doi.org/10.1126/science.aat3628.

64. Singh, P.P., Vishwakarma, S., Sultana, G.N.N., Pilvar, A., Karmin, M., Rootsi, S., Villems, R., Metspalu, M., Behar, D.M., Kivisild, T., et al. (2021). Dissecting the paternal founders of Mundari (Austroasiatic) speakers associated with the language dispersal in South Asia. Eur. J. Hum. Genet. *29*, 528–532. https://doi.org/10.1038/s41431-020-00745-1.

65. Rau, F., and Sidwell, P. (2019). The Munda maritime hypothesis. J. Southeast Asian Linguistics Society *12*, 35–57.

66. van Driem, G.L. (2021). Ethnolinguistic Prehistory: The Peopling of the World from the Perspective of Language, Genes and Material Culture (Brill).

67. Diffloth, G. (2012). The four registers of Pearic. In Conférence plénière at the 22nd Meeting of the Southeast Asian Linguistics Society, Agay.

68. Villalba-Mouco, V., van de Loosdrecht, M.S., Rohrlach, A.B., Fewlass, H., Talamo, S., Yu, H., Aron, F., Lalueza-Fox, C., Cabello, L., Cantalejo Duarte, P., et al. (2023). A 23,000-year-old southern Iberian individual links human groups that lived in Western Europe before and after the Last Glacial Maximum. Nat. Ecol. Evol. *7*, 597–609. https://doi.org/10.1038/s41559-023-01987-0.

69. Gavashelishvili, A., and Tarkhnishvili, D. (2016). Biomes and human distribution during the last ice age. Global Ecol. Biogeogr. *25*, 563–574. https://doi.org/10.1111/geb.12437.

70. Burroughs, W.J. (2005). Climate Change in Prehistory (Cambridge University Press). https://doi.org/10.1017/CBO9780511535826.

71. Clark, P.U., Dyke, A.S., Shakun, J.D., Carlson, A.E., Clark, J., Wohlfarth, B., Mitrovica, J.X., Hostetler, S.W., and McCabe, A.M. (2009). The Last Glacial Maximum. Science *325*, 710–714. https://doi.org/10.1126/science.1172873.

72. Lieberman, B., and Gordon, E. (2021). Climate Change in Human History: Prehistory to the Present (Bloomsbury Publishing).

73. Kulke, H., and Rothermund, D. (2016). A History of India (Routledge).

74. Kumar, N., and Mukherjee, D.P. (1975). Genetic distances among the Ho tribe and other groups of Central Indians. Am. J. Phys. Anthropol. *42*, 489–494. https://doi.org/10.1002/ajpa.1330420316.

75. Mondal, P.R., Saksena, D., Sachdeva, M.P., Murry, B., Meitei, K.S., Samtani, R., and Saraswathy, K.N. (2011). The Genomic Similarities with Linguistic Difference: A Study Among the Oraon and Munda Tribes of the Ranchi District, Jharkhand, India. Genet. Test. Mol. Biomark. *15*, 443–449. https://doi.org/10.1089/gtmb.2010.0187.

76. Kolipakam, V., Jordan, F.M., Dunn, M., Greenhill, S.J., Bouckaert, R., Gray, R.D., and Verkerk, A. (2018). A Bayesian phylogenetic study of the Dravidian language family. R. Soc. Open Sci. *5*, 171504. https://doi.org/10.1098/rsos.171504.

77. Fuller, D.Q. (2003). An Agricultural Perspective on Dravidian Historical Linguistics: Archaeological Crop Packages, Livestock and Dravidian Crop Vocabulary (Examining the Farming/language Dispersal Hypothesis), pp. 191–213.

78. Krishnamurti, B. (2003). The Dravidian Languages (The Cambridge University).

79. Southworth, F.C. (2006). Proto-Dravidian agriculture. In Proceedings of the Pre-symposium of Rihn and 7th ESCA Harvard-Kyoto Roundtable. Research Institute for Humanity and Nature, O. Toshiki, ed. (Kyoto), pp. 121–150.

80. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

81. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7. https://doi.org/10.1186/s13742-015-0047-8.

82. Delaneau, O., Coulonges, C., and Zagury, J.-F. (2008). Shape-IT: new rapid and accurate algorithm for haplotype inference. BMC Bioinf. *9*, 540. https://doi.org/10.1186/1471-2105-9-540.

83. Chang, C.C. (2020). Data Management and Summary Statistics with PLINK. Methods Mol. Biol. *2090*, 49–65. https://doi.org/10.1007/978-1-0716-0199-0_3.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
GenomeAsia 100K Consortium data	Wall et al.[40]	https://doi.org/10.1038/s41586-019-1793-z
Software and algorithms		
Codes for demographic inference using dadi	Zenodo	https://doi.org/10.5281/zenodo.18533055
PLINK version 1.9	Purcell et al.[80]; Chang et al.[81]	https://www.cog-genomics.org/plink/1.9/
PLINK version 2.0	Chang et al.[81]	https://www.cog-genomics.org/plink/2.0/
ADMIXTURE	Alexander et al.[42]	https://dalexander.github.io/admixture/download.html
ADMIXTOOLS	Patterson et al.[58]	https://github.com/DReichLab/AdmixTools
dadi	Gutenkunst et al.[44]	https://dadi.readthedocs.io/en/latest/
Treemix	Pickrell and Pritchard[54]	https://bitbucket.org/nygcresearch/treemix/downloads/
OptM	Fitak[55]	https://cran.r-project.org/web/packages/OptM/index.html
hap-IBD	Zhou et al.[59]	https://github.com/browning-lab/hap-ibd
SHAPEIT	Delaneau et al.[82]	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
Relate	Speidel et al.[53]	https://myersgroup.github.io/relate/
R software (version 4.4.1)	R Core Team	https://www.r-project.org/

METHOD DETAILS

Dataset and quality control

The GenomeAsia Pilot (GAsP) project dataset contained whole genome sequences (WGSs) of 1739 individuals belonging to 219 populations (70 from South Asia & 26 from Southeast Asia) and reported a total of nearly 66 million single nucleotide polymorphisms (SNP).[40] We performed quality control on this dataset using mostly PLINK version 1.9 and 2.0.[80,81,83] We exclusively considered bi-allelic autosomal SNPs for all our analyses. For most of our analyses, we removed SNPs which had missing genotype rates greater than 2% and minor allele frequencies (MAF) less than 0.01. To account for non-independence of SNPs, we further pruned the data for linkage disequilibrium (LD) with the following parameters: (i) window-size = 50, (ii) window-increment = 5, and (iii) r^2 threshold of 0.2. After the different levels of filtering, we retained 1,098,313 SNPs for further downstream analysis. We selected 497 unrelated individuals from 58 South Asian populations, out of which 471 individuals were from mainland South Asia and 26 were from the Andaman and Nicobar Islands (Table S1). In all the subsequent analyses, we used a subset of these individuals and alongside 30 Yoruba (YRI) individuals from Africa, 28 British (GBR) individuals from Europe, and 16 Kintak (KIN) and 9 Senoi Semai (SNS) individuals from Southeast Asia, in certain specific analyses. These individuals were also sourced from the GenomeAsia Pilot (GAsP) project dataset.

Population structure analysis

To explore the population structure among the 471 mainland South Asian individuals belonging to 55 populations, we performed the principal component analysis (PCA) using PLINK version 1.9.[80,81,83] We also performed PCA separately on 141 unrelated individuals from AA and DR tribal populations (Table S2), to detect finer structuring within this subset.

ADMIXTURE analysis

We further performed ADMIXTURE[42] analysis, a model-based clustering approach to infer the relative proportions of the different genetic clusters in the 497 individual genomes from 58 South Asian populations. We separately performed ADMIXTURE analysis considering 141 unrelated individuals from AA and DR tribal populations.

Outgroup-f3 analysis

We used *outgroup-f3* statistics[43] of the form f3(outgroup; pop1, pop2) to estimate the shared genetic drift between two unadmixed populations with respect to an outgroup population that diverged long before the separation of the two populations. In our analyses, we considered the African population YRI as the outgroup. We used the qp3Pop implementation of ADMIXTOOLS[58] to estimate *outgroup-f3* values. First, we estimated the shared genetic drift between the high-ANI population (e.g., PAT) and the AA/DR tribe. Then, we used the *outgroup-f3* statistics[43] of the form f3(YRI; an AA/DR tribe, another AA/DR tribe) to measure shared genetic drift between AA and DR tribal populations.

Dadi

dadi (Diffusion Approximation for Demographic Inference)[44] is an estimation method based on the joint site frequency spectrums (SFS) of SNPs, which is used to infer the demographic histories of populations. Given a parameterized demographic model, dadi simulates the joint SFS of the populations under the model and updates the model parameters using non-linear optimization to maximize the likelihood of the observed SFS given the model SFS. We performed dadi on the populations of interest considering various scenarios including no migration, symmetric migration and asymmetric migration. We used the full dataset of ~66 million SNPs for this analysis, after discarding SNPs with genotype missingness >2% and those which were monomorphic for the individual populations. In the demographic model that we tested on dadi, an ancestral population was allowed to evolve with changing effective population size (N_e) until it split into two populations with N_e estimated as fractions of the undivided ancestral population. One of these subdivided populations (outgroup) was allowed to evolve (N_e grows or declines exponentially) up to present. For the second population, the effective population size was allowed to change exponentially until it further split into two populations that represent our target present-day populations. These target populations corresponded to any of the three categories: (i) one population with high ASI ancestry and another with high AAA ancestry, (ii) one population with high AAA ancestry and another representing Malaysian AA-ancestry, (iii) one population with high ASI ancestry and another representing Malaysian AA-ancestry.

The estimated parameters corresponded to the split times, effective population sizes, and the split fractions. As the convergence of the model parameters depended on the initial values (input parameters for the simulation), we performed 25 iterations of simulations for each demographic model. These 25 iterations corresponded to 5 sets of randomly generated initial parameters. The optimal values of the parameters were chosen based on the maximum log likelihood values and the standard deviations were calculated based on the Godambe Information matrix incorporated in dadi, which performs well for composite likelihoods.

Treemix analysis

Treemix[54] is a powerful tool to construct population trees that enables inference of historical splits and migrations between populations. Along with building the maximum likelihood tree, Treemix also estimates gene flow between the populations by adding ''migration edges'' between the nodes. Each migration edge corresponds to an event of admixture. For the Treemix analysis, we used the African population Yoruba (YRI) as the outgroup population. We used 7,980,009 SNPs for our analysis, which were selected after discarding SNPs with missingness >2% and MAF <0.01. To account for LD, we grouped the nearby SNPs together in blocks with block size of 1000 SNPs. We also varied the number of migration edges (m) which corresponded to the events of gene flow among the populations. To estimate the optimum value of m, we further carried out OptM analysis.[55] To be able to use OptM analysis, we ran 5 iterations for each m and for each iteration, we considered a randomly selected block size of between 500 and 1000 SNPs (50 SNP increments). The Δm method incorporated in OptM, which is based on the second order rate of change of the likelihood, suggested that the optimum m = 4.

D-statistics

D-statistics[56–58] is a formal test of gene flow between populations. For any four populations (W, X, Y, Z), where Z is an outgroup population, D-statistics of the form D(W, X; Y, Z) can detect the gene flow between two populations. If $D > 0$ and the corresponding Z score >3, it suggests gene flow between W and Y. If $D < 0$ and the corresponding Z score < -3, it suggests gene flow between X and Y. We used the qpDstat implementation of ADMIXTOOLS[58] to compute the D-statistics. For all our D-statistics analyses, we considered YRI as the outgroup population Z, BIR as W, PNY as X and any other AA/DR tribe as Y.

Hap-IBD analysis

We utilized the hap-IBD[59] method, which can detect the IBD (identical by descent) segments in phased genotype data, to estimate separation time from a common ancestor. We phased the unfiltered autosomal genotype data comprising ~66 million SNPs using SHAPEIT,[82] which resulted in ancestral coded haplotypes comprising ~55 million SNPs. We performed hap-IBD on this phased data and restricted the estimated lengths of IBD segments to $\geq$1 cM, following Browning & Browning.[60] We set all other parameters used in hap-IBD as default. We estimated the number of generations (g) elapsed after the IBD segment was inherited from a common ancestor, based on its relationship with length of the IBD segment in Morgan (L) as follows: $g = 3/2L$.[61] We sorted the IBD segments in bins with lengths of $(5,\infty)$ cM, (2.5,5] cM, (1.667–2.5] cM, (1.25–1.667] cM and [1,1.25] cM, which corresponded to the estimated time intervals: 0–30, 30–60, 60–90, 90–120 and 120–150 generations respectively. In each time interval, we defined the average number of IBD segments shared between any two individuals from two different groups = Total number of segments shared between the groups/(no. of individuals in 1st group × no. of individuals in 2nd group).

Relate

Relate[53] can be used to estimate genome-wide genealogies in the forms of bifurcated trees, and it is scalable to thousands of samples. We used the same ancestral coded haplotypes comprising ~55 million SNPs that were used in the hap-IBD analysis in Relate to generate the genome-wide genealogy for 658 unrelated individuals, including individuals from 58 South Asian populations, 2 Southeast Asian populations (SNS and KIN), 3 Northeast Asian populations (HAN, JPN and KHL), 4 West Asian populations (IRN, JOR, MEJ and PAL), 1 African population YRI and 1 European population GBR. From the estimated full genealogy, we extracted the genealogy for a subsample including our populations of interest and re-estimated the within and across-group coalescence times for various populations using Relate, considering generation time = 25 years and mutation rate = 1e-8/bp/generations.[45] These estimates were used to estimate effective population sizes for the populations at different time points and their divergence times.

QUANTIFICATION AND STATISTICAL ANALYSIS

We performed the two-sample Kolmogorov-Smirnov test to test the equality of distributions of *outgroup-f3* values using ks.test() function in R software (version 4.4.1). To calculate Spearman's rank correlation between mean ANI ancestry proportions and respective *outgroup-f3* values, we used cor.test() function with the argument method = ''spearman'' in R. To calculate Pearson's product moment correlation between the DR tribes and their AAA ancestry, we used cor.test() function with the argument method = ''pearson'' in R.