

Supplementary Information for
**Exploring the genomic population structure and history of Austroasiatic speakers in
Mainland Southeast Asia**

Zi Yin¹, Yash Munnalal Gupta^{2,3}, Nonglak Prakhun², Jatupol Kampuansai⁴, Angkana Inta⁴, Metawee Srikumool⁵, Putsadee Rodcharoen⁶, Chatmongkon Suwannapoom⁷, Sengvilay Lorphengsy⁸, Wipada Woravatin⁹, Jae Joseph Russell B. Rodriguez¹⁰, Chawalit Khaokhiew⁶, Mark Stoneking^{11,12}, Wibhu Kutanan^{2,3*}, Dang Liu^{11,13,14*}, Ke Wang^{1*}

1. MOE Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, Center for Evolutionary Biology, School of Life Sciences, Fudan University, Shanghai, China
2. Department of Biology, Faculty of Science, Naresuan University, Phitsanulok, Thailand
3. Center of Excellence for Innovation and Technology for Detection and Advanced Materials (ITDAM), Naresuan University, Phitsanulok, Thailand
4. Department of Biology, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand
5. Department of Biochemistry, Faculty of Medical Science, Naresuan University, Phitsanulok, Thailand
6. Department of Archaeology, Silpakorn University, Bangkok, Thailand.
7. School of Agriculture and Natural Resources, University of Phayao, Muang Phayao, Thailand
8. The Biotechnology and Ecology Institute, Ministry of Science and Technology, Vientiane, Laos
9. Department of Biology, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand
10. Department of Archaeogenetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, D04103, Germany
11. Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, D04103, Germany
12. Biométrie et Biologie Évolutive, UMR 5558, CNRS & Université de Lyon, Lyon, France
13. Human Evolutionary Genetics Unit, Institut Pasteur, UMR 2000, CNRS, Paris, France
14. Current affiliation: Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland

***These authors jointly supervised and co-corresponded this work:** Wibhu Kutanan (wibhuk@nu.ac.th); Dang Liu (dang.liu@uzh.ch); Ke Wang (ke_wang@fudan.edu.cn; lead contact)

Supplementary Note 1: Overall IBD sharing among AA speakers and MSEA populations

In the effective population size estimates based on IBD, we also noticed that Katuic, Khmuic (except Khomu), Palaungic (except Blang), and Khmeric (except Cambodian) groups share a population size bottleneck estimated to be within the last 10 to 50 generations before present (Supplementary Fig. 5); this apparent population decline might be associated with historical events such as warfare and political instability, which resulted in migration and isolation of the minority groups^{1,2}. In contrast, around the same period, major ethnic groups in their respective countries, such as Viet-Muong speakers, Southern Laotian, Burmese, Cambodian and Blang, exhibit genetic signatures of population expansion. However, the absolute estimates of intensity and timing from hapNe may be biased by population demographic history, and the exact timing of expansions or bottlenecks should be interpreted with caution. We further calculated the average number and total length of shared IBD segments across all AA-speaking groups. The Mlabri and Maniq exhibited exceptionally high within-group IBD sharing, consistent with the strong genetic drift observed in the PCA and ADMIXTURE analyses. The Mang, HtinMal, Lua, and HtinPray groups also demonstrated high within-group sharing. In contrast, the Viet-Muong groups and Cambodian exhibited lower within-group sharing, which aligns with their signals of population expansion (Supplementary Fig. 6).

Furthermore, we analysed the recent genetic contact between AA speakers and other MSEA populations by summarizing the total length of all pairwise inferred IBD segments (Methods, Supplementary Fig. 7). Except for Kharia and Maniq, all MSEA AA groups share an average total IBD length of at least 10 cM, indicating recent common ancestry or contact³. The Viet-Muong groups, Mon, Palaung, and Cambodian show weaker connections with other AA groups, and distinct clusters are also observed within Eastern and Northern Mon-Khmer language branches, highlighting substantial genetic heterogeneity. Network analyses of strong IBD sharing with the average total length over 20 cM (Supplementary Fig. 7c, d) reveal that AA groups had more interactions with TK, AN, and ST groups than with HM groups. Notably, AN groups such as Malay, Ede, Giarai, and Borneo, strongly share with the Eastern Mon-Khmer cluster; ST Karen shares with the Northern Mon-Khmer cluster; and MSEA TK groups show moderate sharing across AA groups. These patterns are also supported by ChromoPainter (Supplementary Fig. 8), PCA, and ADMIXTURE results, indicate that differential contact with neighboring language families shaped the genetic structure of Eastern and Northern Mon-Khmer populations. Overall, these results further attest to the complex patterns of genetic heterogeneity among AA groups.

Supplementary Note 2: Modelling strategy for the ancient MSEA populations

Following the hypothesis that modern and ancient MSEA populations derive their ancestry mainly from a local source, we began with the pairwise qpWave for ancient MSEA populations and the five ancestry sources to assess their genetic cladality: namely, if the tested populations could be modeled by the same ancestry (Supplementary Fig. 12, Supplementary Data 3). In the cases when a pairwise qpWave test failed, it indicates that the tested populations cannot be explained by a single ancestry or the same admixture event. We observed that most ancient individuals showed cladality with Liangdao1, instead of with Laos_Hoabinhian.SG. However, N-NamTun, N-TamHang (Neolithic individuals from the same region in northern highland MSEA), and N-LoyangUjungCave (Neolithic individuals from western Indonesia) exhibited cladality with both Liangdao1 and Longlin. IA-BanRaiRockshelter (Iron Age individuals from northwestern Thailand) and Hi-SupuHujung (Historical era individuals from northern Borneo) showed cladality with both Liangdao1 and China_YR_MN, while H-LiangToge and H-Komodo (Historical era individuals from western Indonesia) showed cladality only with Longlin.

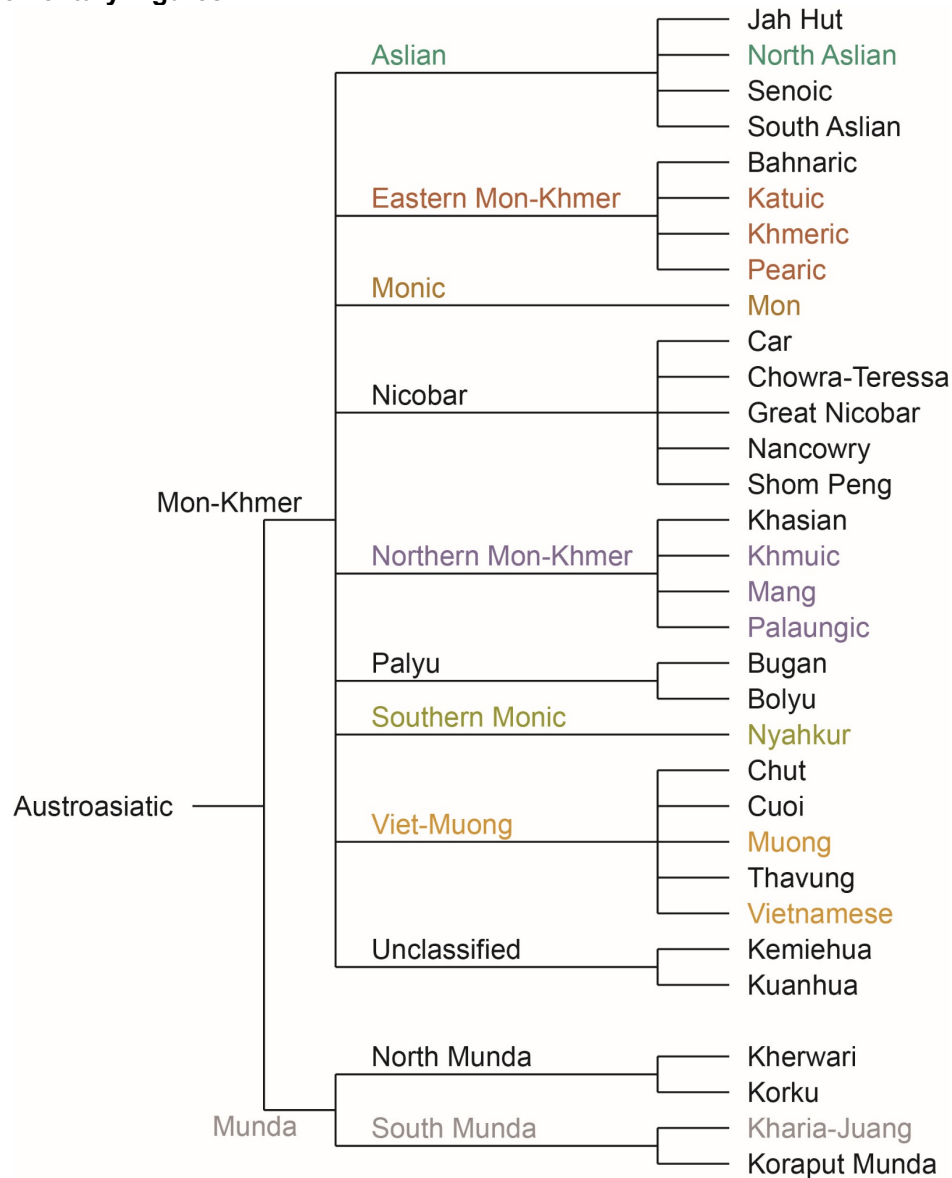
As Liangdao1 and Longlin are both early East Asian genetic lineages found in southern East Asia⁴, some distance from the MSEA continent, we considered multiple-sources models for ancient MSEA populations in the following modelling tests, using the indigenous MSEA ancestry as a fixed source (Laos_Hoabinhian.DG)⁵.

We then tried 2-sources models using Laos_Hoabinhian.SG as one source, and tested another possible ancestry (from South Asia or East Asia) as the other source, given that

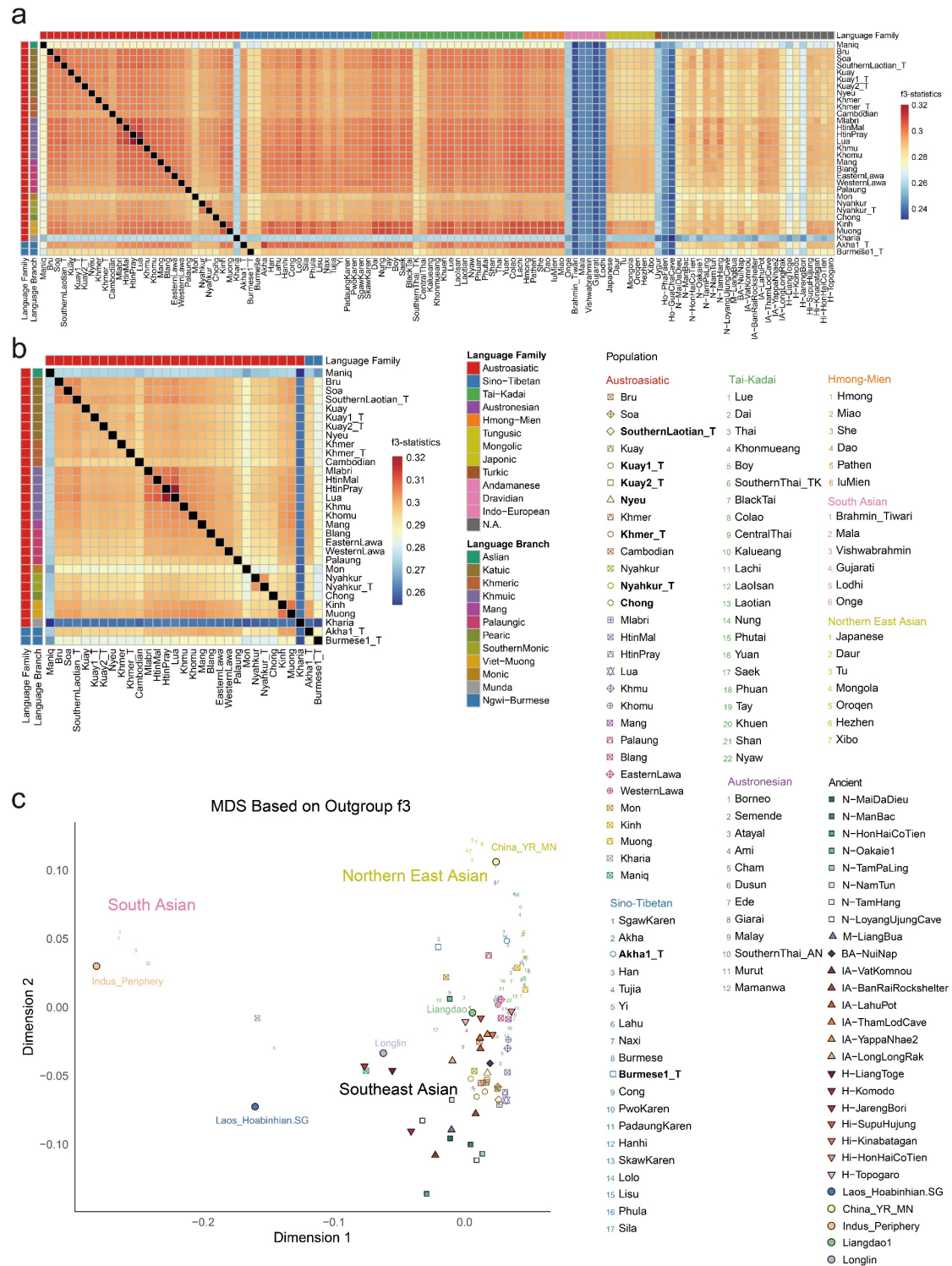
f_4 (ancient MSEA population, Maniq; South Asian/East Asian, Mbuti) statistics are clearly showing extra genetic affinity towards South Asian (Indus_Periphery) or East Asian (China_YR_MN) for some of the ancient MSEA groups (Supplementary Fig. 13).

Built on the successfully 2-sources models, we conducted qpAdm analyses with a rotating strategy by testing two potential sources while adding the other potential sources to the base outgroups, to examine whether the potential sources being added to the base outgroups would break the statistical fit of a 2-sources model (Supplementary Data 5).

Supplementary Figures

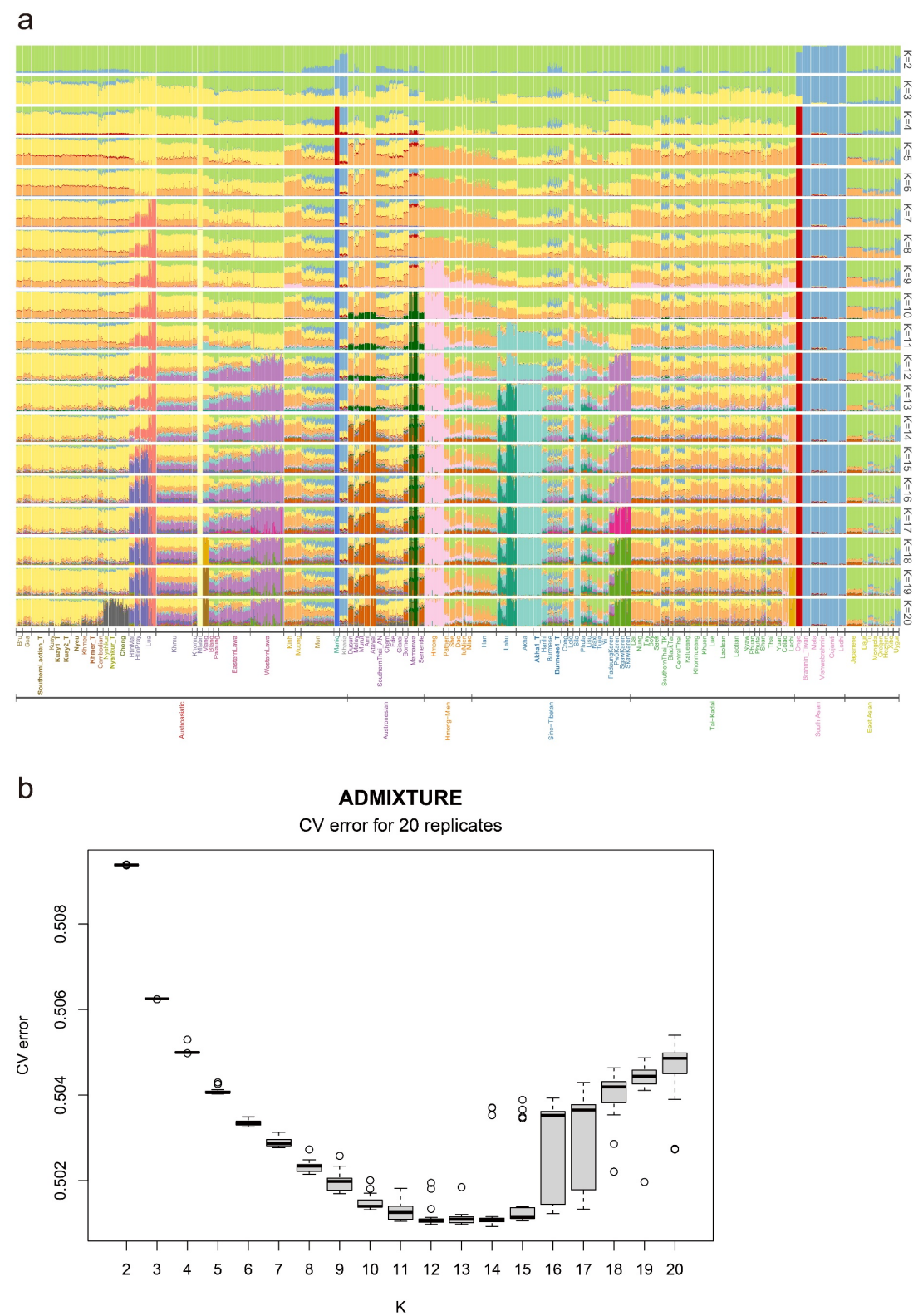


Supplementary Fig. 1 | Linguistic classification of the Austroasiatic language family in *Ethnologue*⁶. The AA language family comprises two primary branches - Mon-Khmer and Munda. We analysed speakers from the Aslian, Eastern Mon-Khmer, Northern Mon-Khmer, Monic, Southern Monic, Viet-Muong subbranches of the Mon-Khmer branch, as well as the South Munda subbranch of the Munda branch, which are highlighted in colored text in the phylogeny tree.



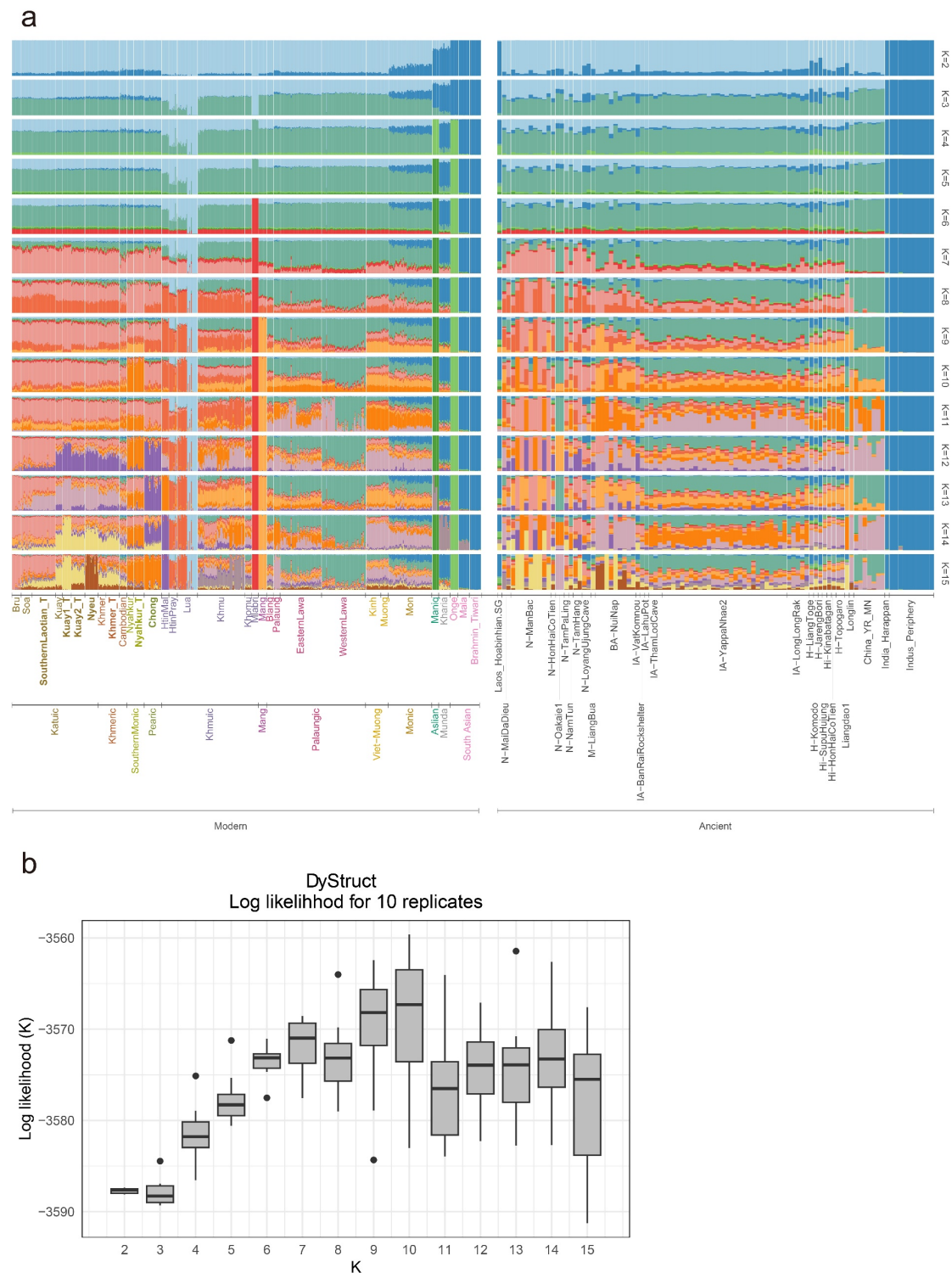
Supplementary Fig. 2 | Heatmap and MDS of outgroup f_3 statistics. In the heatmap, the colored bar in the column indicates the language family for comparative population, and the colored bar in the row indicates the language family and language branch for each group. The f_3 statistics were performed using 518,516 SNPs. **a** The estimated allele sharing between Austroasiatic populations/two Sino-Tibetan populations ($n = 31$) and all analyzed modern MSEA, South Asian, East Asian populations together with related ancient groups ($n = 115$). Austroasiatic populations are organized by language branches, while other populations are arranged by language family and location; **b** The estimated allele sharing among Austroasiatic populations and two Sino-Tibetan populations ($n = 31$); **c** MDS analysis based on the outgroup f_3 results, including both ancient and modern populations. To evaluate the fit of the MDS solution, we calculated the sum of squared differences between the observed dissimilarities

and the distances in the MDS solution. This sum serves as an indirect measure of the stress value (here is at 0.86).

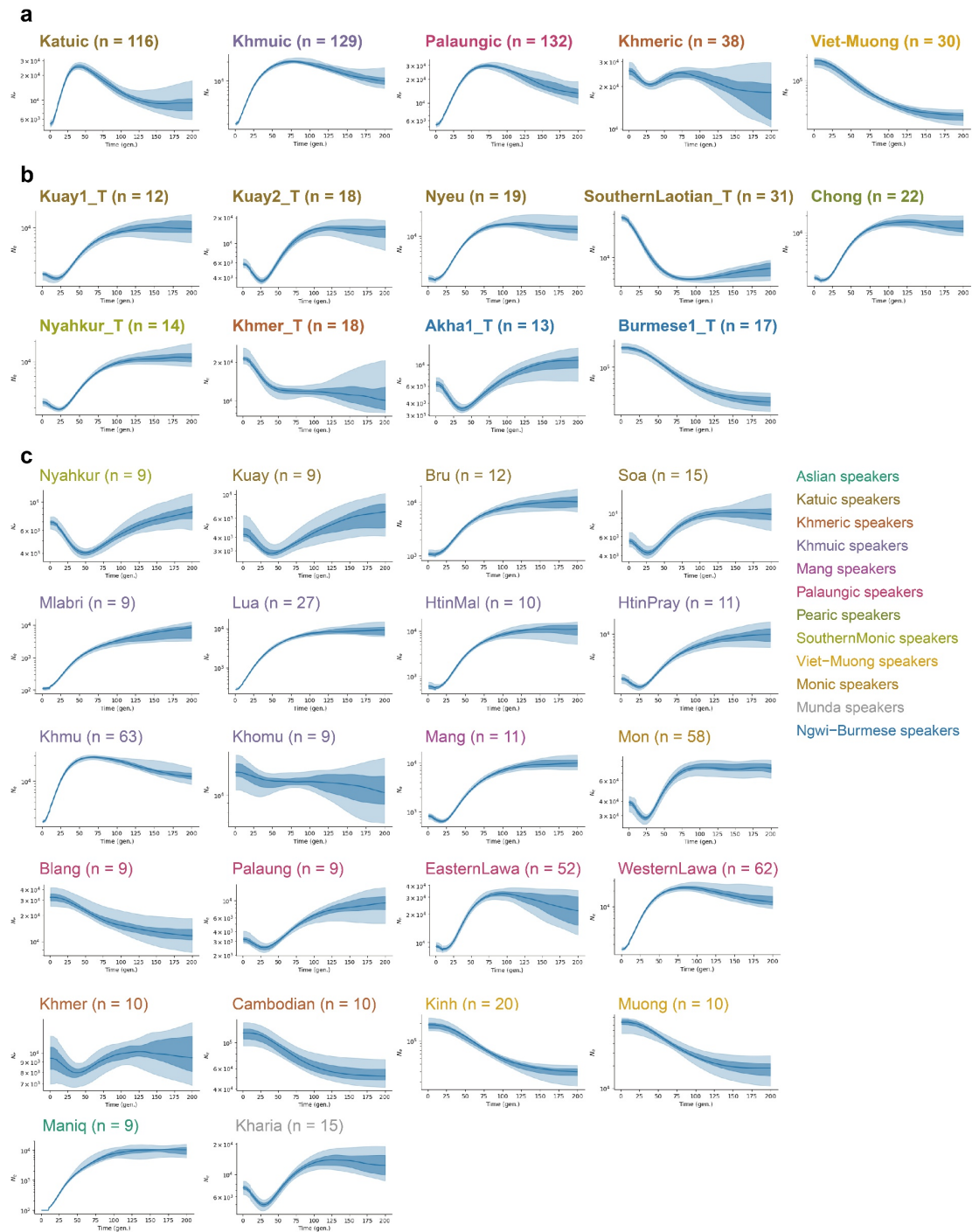


Supplementary Fig. 3 | ADMIXTURE results for each K. ADMIXTURE was conducted using 142,139 SNPs. **a** ADMIXTURE analysis ranging from K = 2 to K = 20 was conducted on modern populations from South Asia, Northeast Asia, and Southeast Asia; **b** The plot below displays the CV errors with the minimum CV error observed at K = 12. The boxes represent the interquartile ranging between 25th and 75th quartile, with the horizontal bar denoting the median

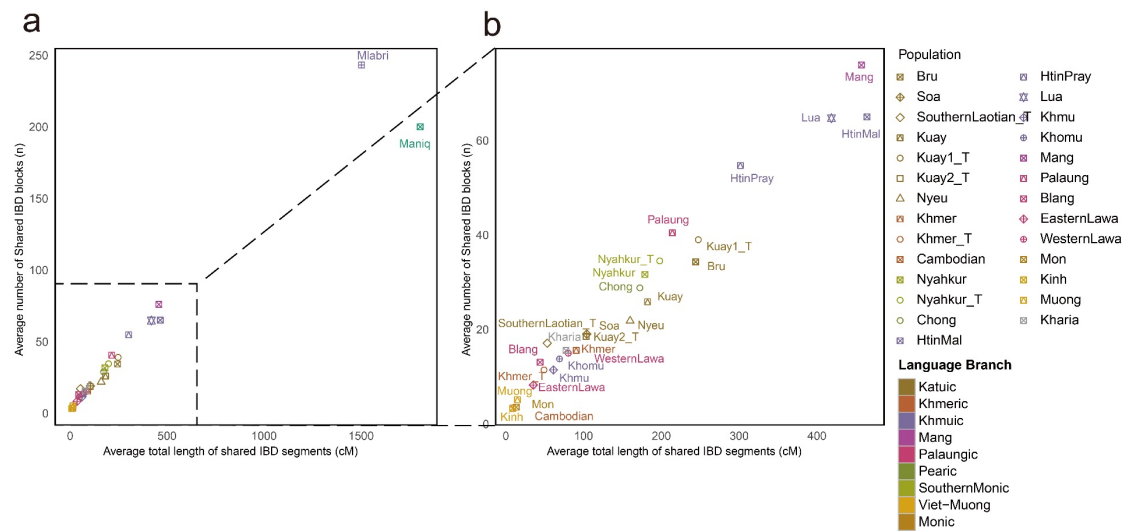
and the whiskers extending to the most extreme data points that fall within 1.5 times the interquartile range. Individual points falling outside this range are shown as outliers.



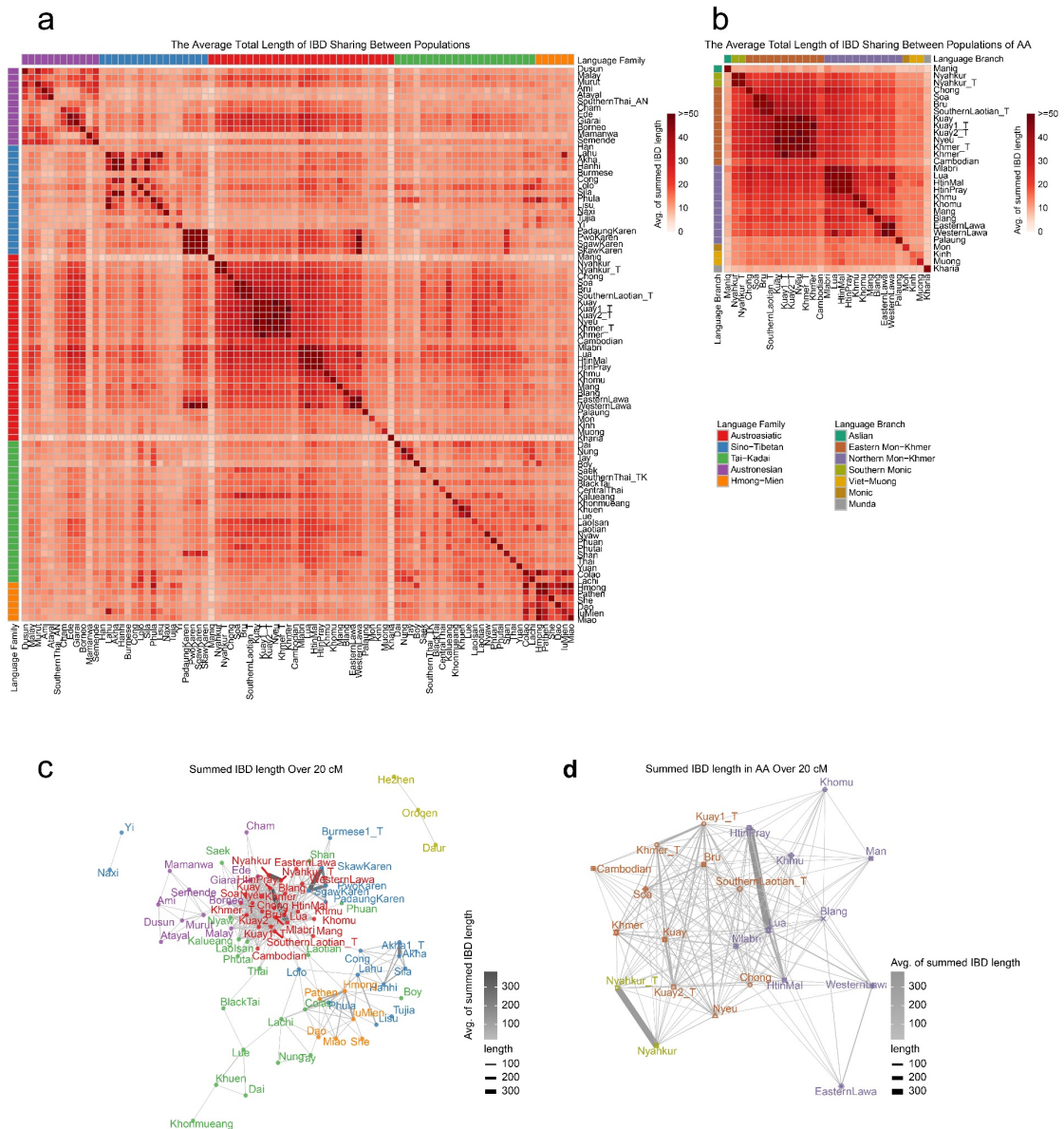
Supplementary Fig. 4 | DyStruct results for each K. DyStruct analysis was conducted using 142,139 SNPs. **a** DyStruct analysis conducted across modern populations from Austroasiatic (AA) and ancient individuals, ranging from K = 2 to K = 15; **b** The plot demonstrates the Log likelihood, revealing three prominent peaks in Log likelihood at K = 5, K = 9, and K = 12. The boxes represent the interquartile range (25-75%) of the 10 replicates, with the horizontal bar denoting the median and the whiskers extending to the most extreme data points that fall within 1.5 times the interquartile range. Individual points falling outside this range are shown as outliers.



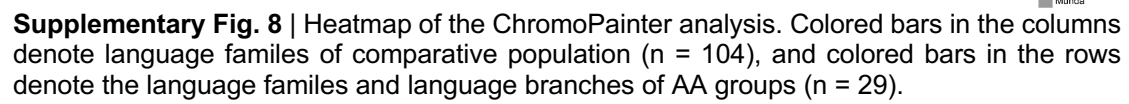
Supplementary Fig. 5 | Estimates of effective population size over the past 200 generations for individual and aggregate AA populations. Each population name is colored by the language branch. **a** Estimated N_e for the aggregated populations (n = 5); **b** Estimated N_e for each newly reported population (n = 9); **c** Estimated N_e for each AA population (n = 22). The light and dark-shaded areas correspond to 95% and 50% confidence intervals respectively, derived from bootstrap quantiles.

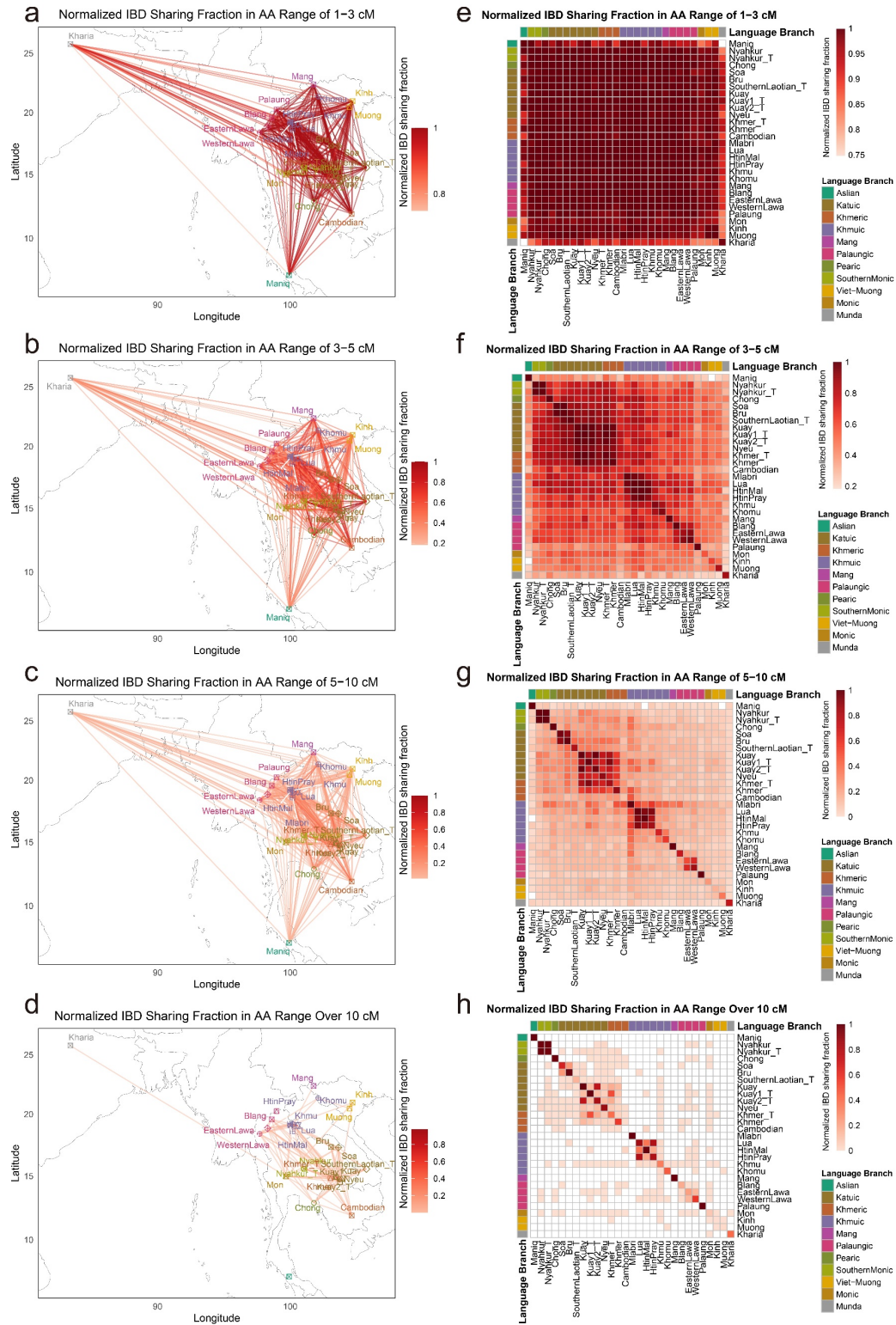


Supplementary Fig. 6 | The IBD sharing patterns within each AA population. The y axis and x axis indicate the average number of shared IBD blocks and the average of summed total length of IBD segments within each group, respectively. Each group is colored by the language branch. **a** Distribution of the average number of shared IBD blocks and the average of summed total length of IBD segments in all AA populations (n = 29); **b** A zoom-in visualization of AA populations (n = 27) except for Mlabri and Maniq.



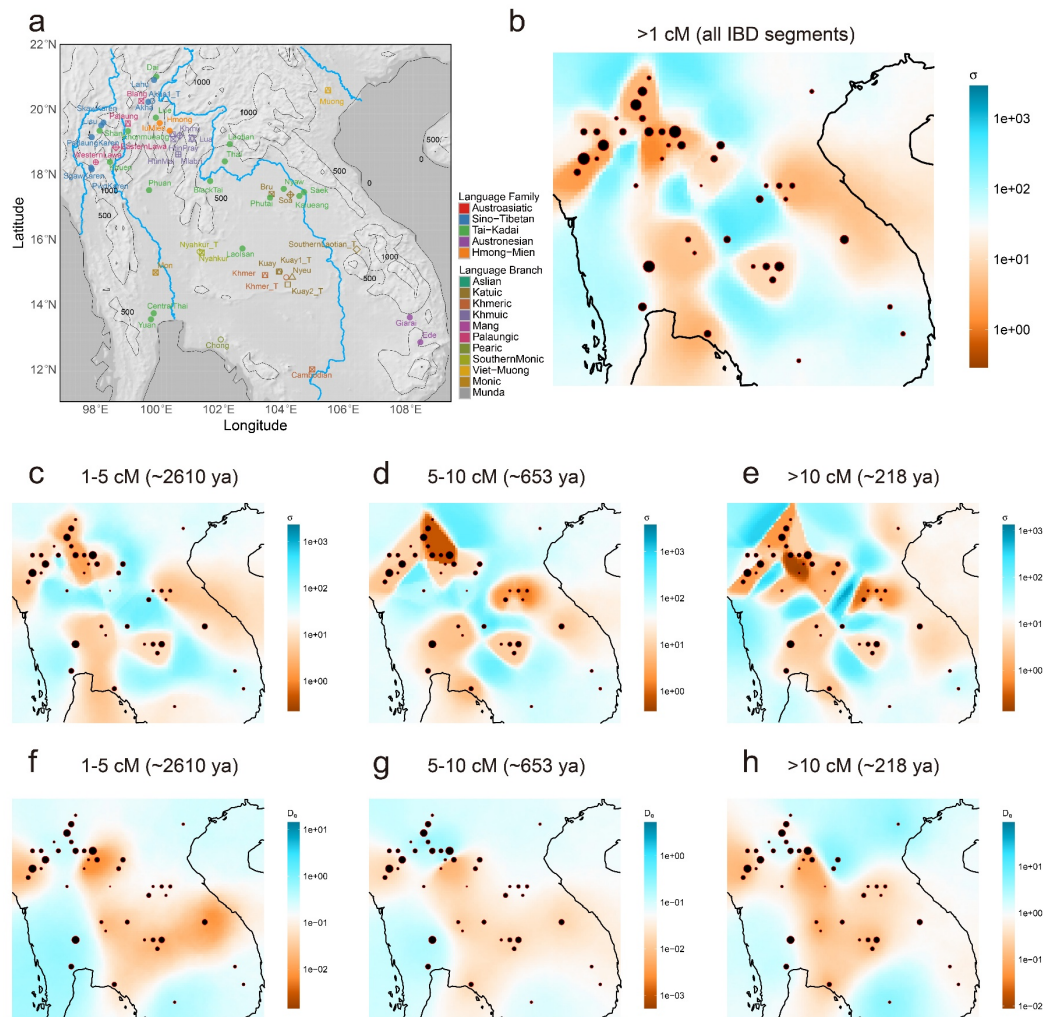
Supplementary Fig. 7 | Haplotype sharing across MSEA, East Asian, and South Asian populations. AA populations are colored by language group, while others are colored by language family. **a.** Heatmap of the average total length of summed all IBD blocks shared between individuals from Southeast Asian, South Asian, and East Asian populations ($n = 86$); **b.** Heatmap of the average total length of summed all IBD blocks shared between individuals from AA populations ($n = 29$); **c.** The network visualization of the average total length of summed all IBD blocks shared between Southeast Asian, South Asian, and East Asian populations ($n = 79$), focusing on the shared IBD blocks with an average summed length longer than over 20 cM; **d.** The network visualization of average total length of summed all IBD blocks shared between AA-speaking populations ($n = 23$), focusing on the shared IBD blocks with an average summed length longer over 20 cM.



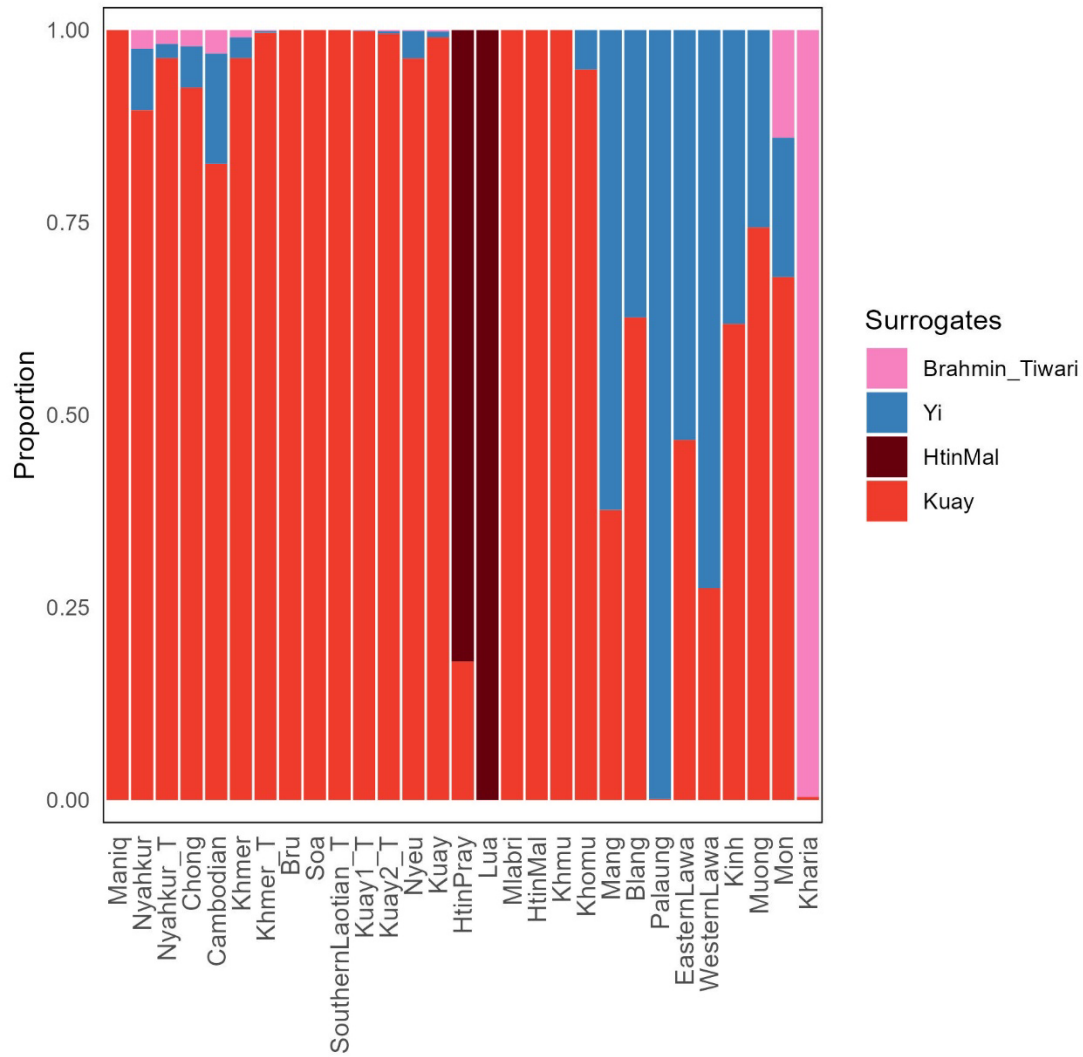


Supplementary Fig. 9 | The network of IBD sharing fraction across AA speakers on a

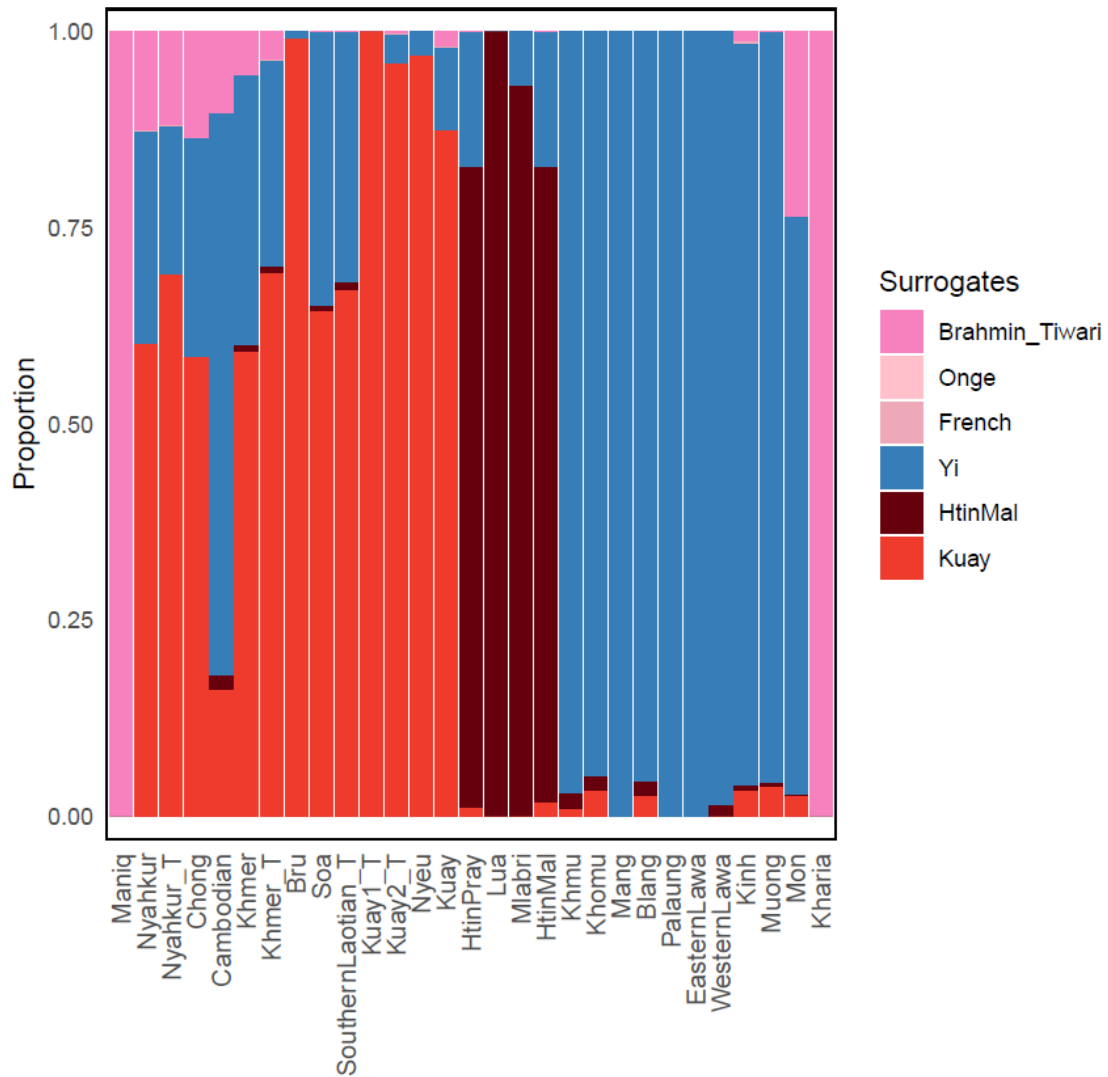
spatial map. We report results for identified IBD blocks in the range of 1 to 3 cM, 3 to 5 cM, 5 to 10 cM and over 10 cM, for AA-speaking populations ($n = 29$).



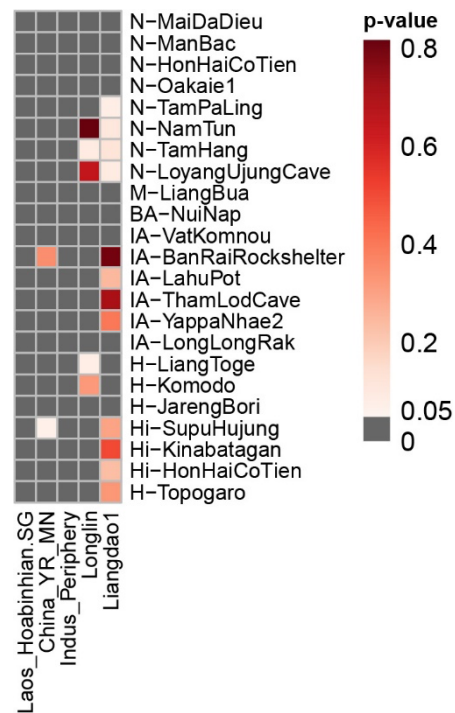
Supplementary Fig. 10 | Spatial pattern of effective migration rates through time. Black points indicate deme locations ($n = 42$), with dot size proportional to deme size. **a** Map indicating the contour of actual altitudes above sea level at 500, 1000, and 1500 meters, also indicating the locations of each MSEA population ($n = 52$). Map data were obtained from Natural Earth; **b** Contour plot of the scaled effective migration rates estimated with MAPS based on all IBD segments; **c-e** Contour plot of estimated dispersal rates from MAPS in the range of 1 to 5 cM, 5 to 10 cM and over 10 cM IBD length. Blue color represents regions of higher-than-average migration rates, while orange color shows regions of lower-than-average migration rates; **f-h** Contour plot of estimated population density from MAPS in the range of 1 to 5 cM, 5 to 10 cM and over 10 cM IBD length. Blue color represents regions of higher-than-average population sizes, while orange color shows regions of lower-than-average population sizes. The indicated time was estimated based on the length of IBD segments, assuming a generation time of 29 years.



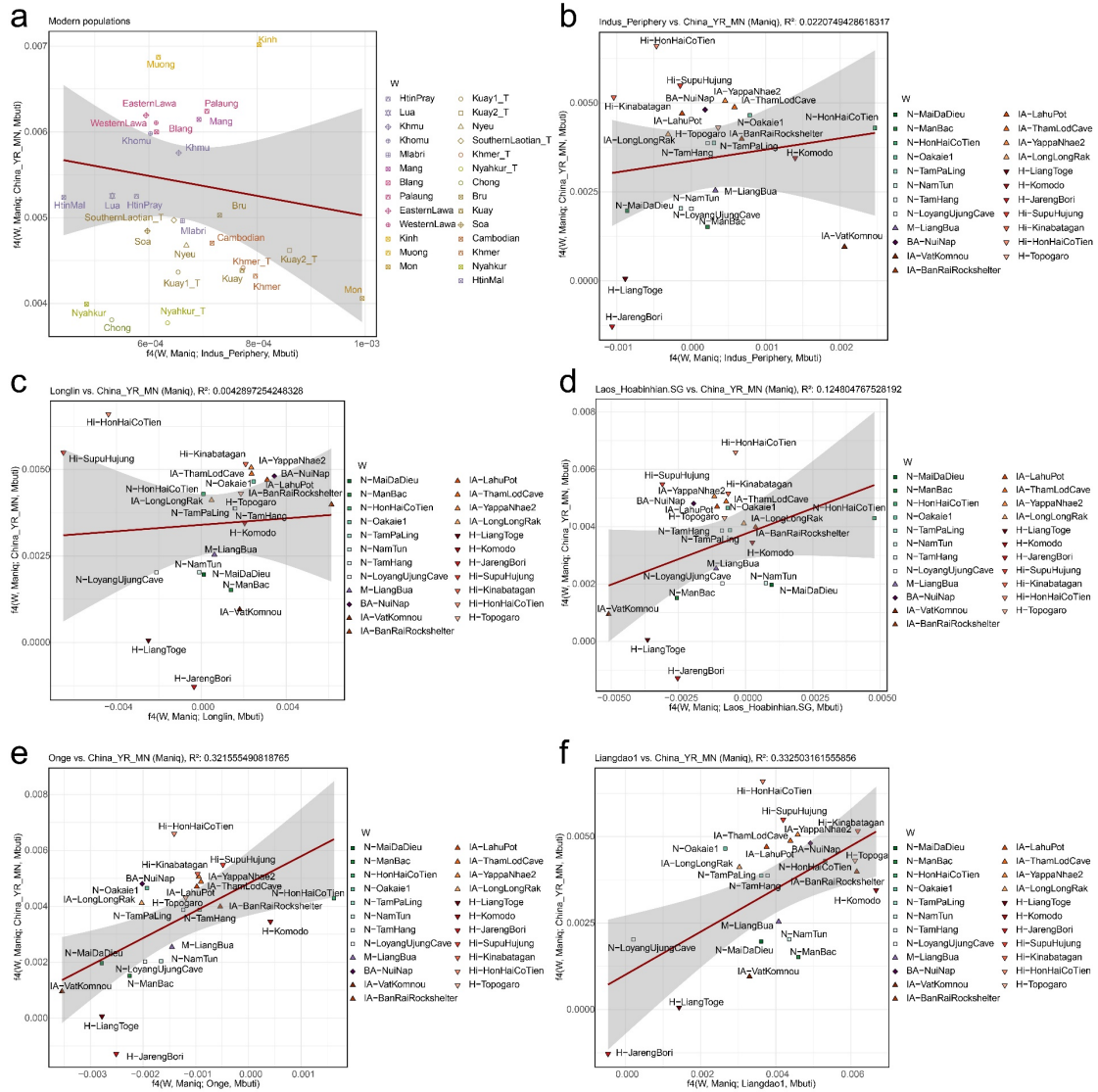
Supplementary Fig. 11 | Bar plot of SOURCEFIND results of AA-speaking groups. SOURCEFIND estimates ancestry proportions in AA groups (n = 29), using Brahmin_Tiwari as the South Asian ancestry, Yi as the East Asian ancestry, and HtinMal, Kuay, and Kuay1_T as the AA ancestry, which allows each AA-speaking group to be represented by one or two other AA surrogates.



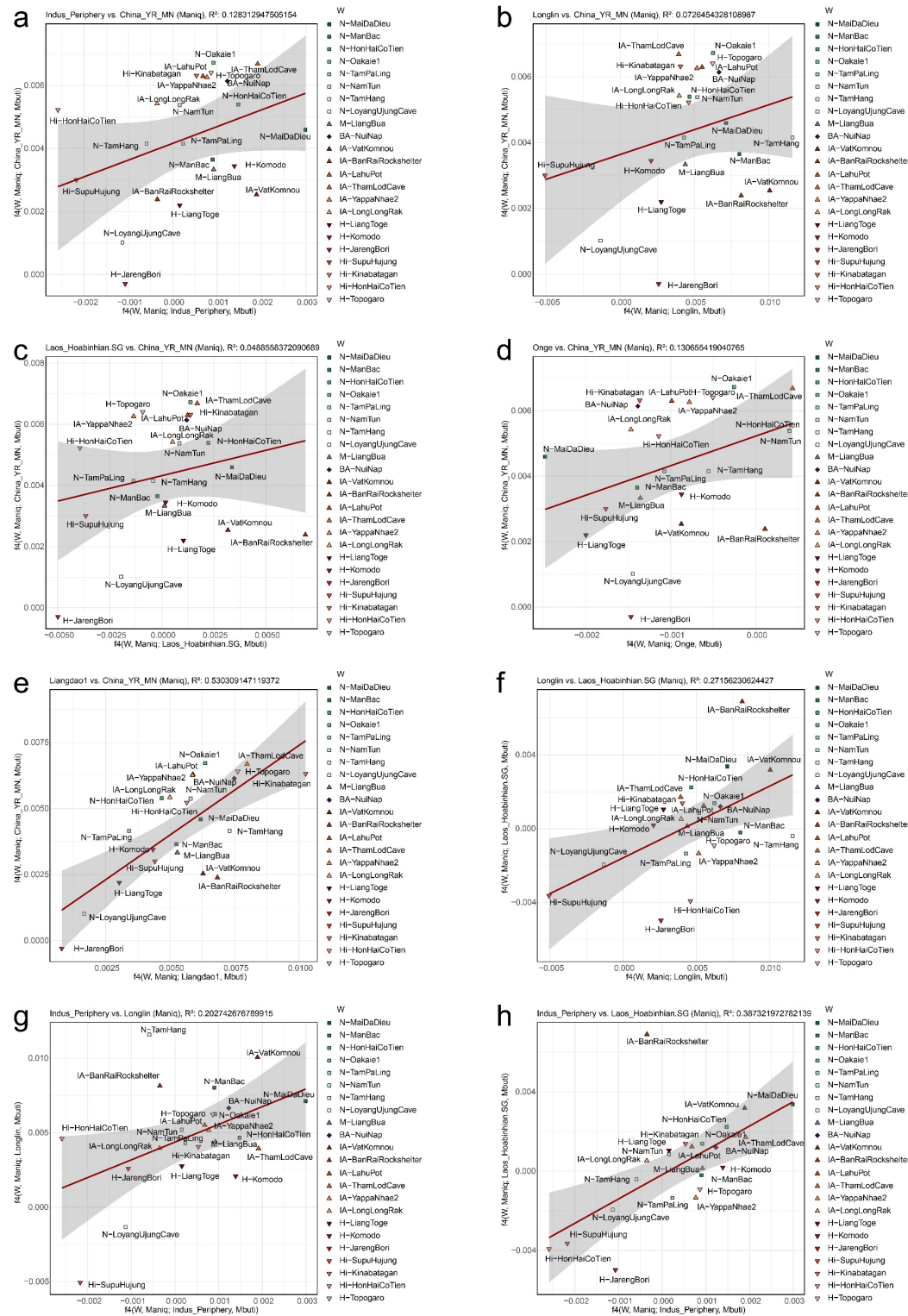
Supplementary Fig. 12 | Bar plot of SOURCEFIND results of AA-speaking groups using alternative South Asian surrogates. Bar plot showing SOURCEFIND estimates of ancestry proportions for AA-speaking individuals ($n = 29$). Brahmin_Tiwari was used as a surrogate for South Asian ancestry, Onge for ASI-related ancestry, and French for Steppe-related ancestry, representing alternative potential sources of South Asian-related components. Yi was included as the East Asian ancestry surrogate, whereas HtinMal, Kuay, and Kuay1_T served as AA ancestry surrogates.



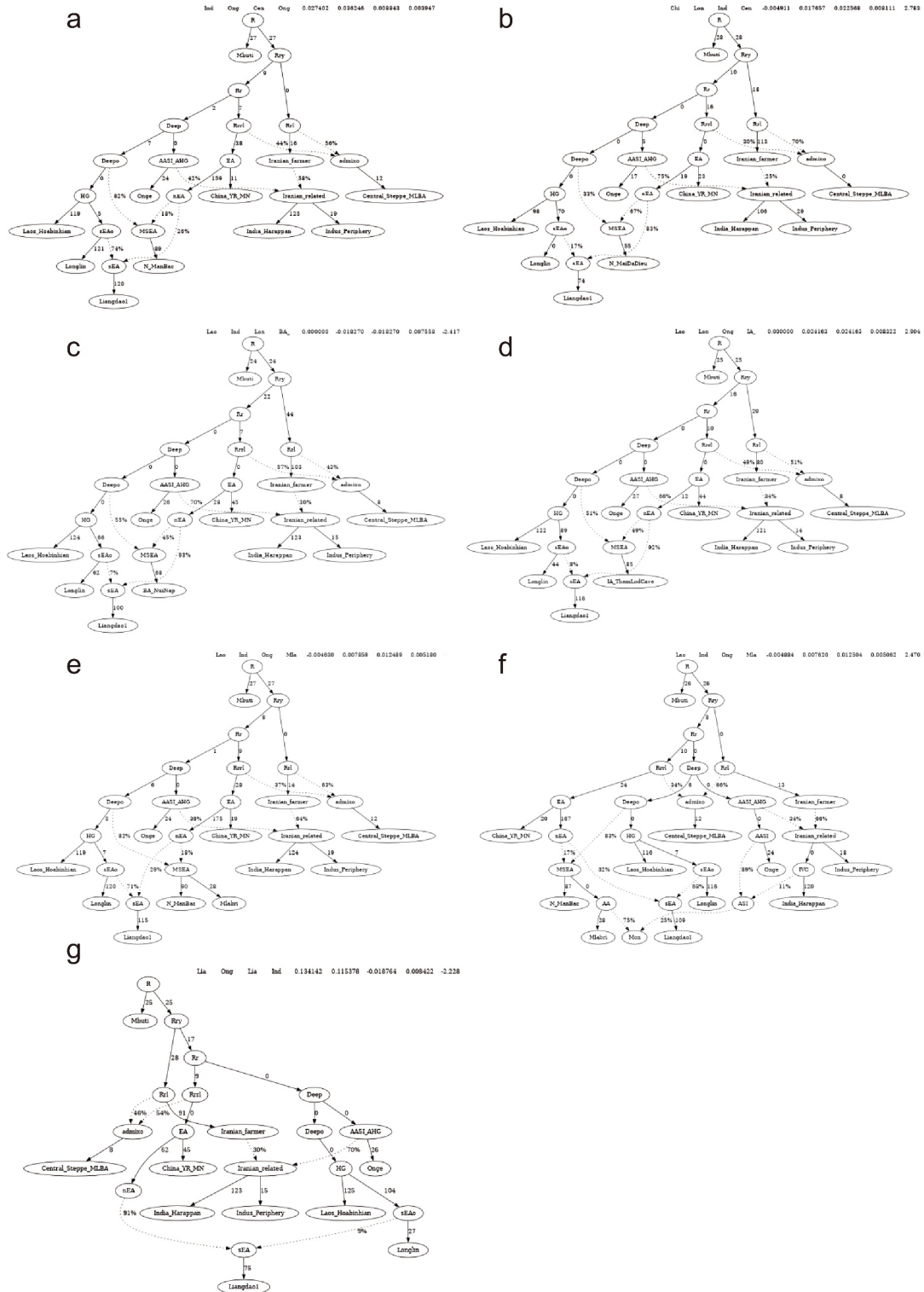
Supplementary Fig. 13 | Genetic cladality test between ancient individuals (n = 23) and the five ancestries by qpWave. The color scale represents p values, with gray blocks indicate p values less than 0.05 (Supplementary Data 4).



Supplementary Fig. 14 | Biplot of $f_4(W, \text{Maniq}; \text{South Asian/East Asian, Mbuti})$ statistics using all SNPs. For compared population W, we performed f_4 statistic using all SNPs (518,516 loci) to compare the shared genetic drift between **a** modern MSEA AA speakers ($n = 27$) and **b** ancient MSEA populations ($n = 23$) with ancient South Asian ancestry (Indus_Periphery) versus ancient ancient East Asian ancestry (China_YR_MN); **c-f** For ancient MSEA populations, we compared their shared genetic drift between ancient South Asian ancestry (Indus_Periphery, Onge) or East Asian ancestry (China_YR_MN, Longlin, Liangdao1) or Southeast Asian ancestry (Laos_Hoabinhian.SG). The red-colored solid line represents a linear fit of the plotted f_4 statistics tests, with the gray shading representing the 95% confidence interval.

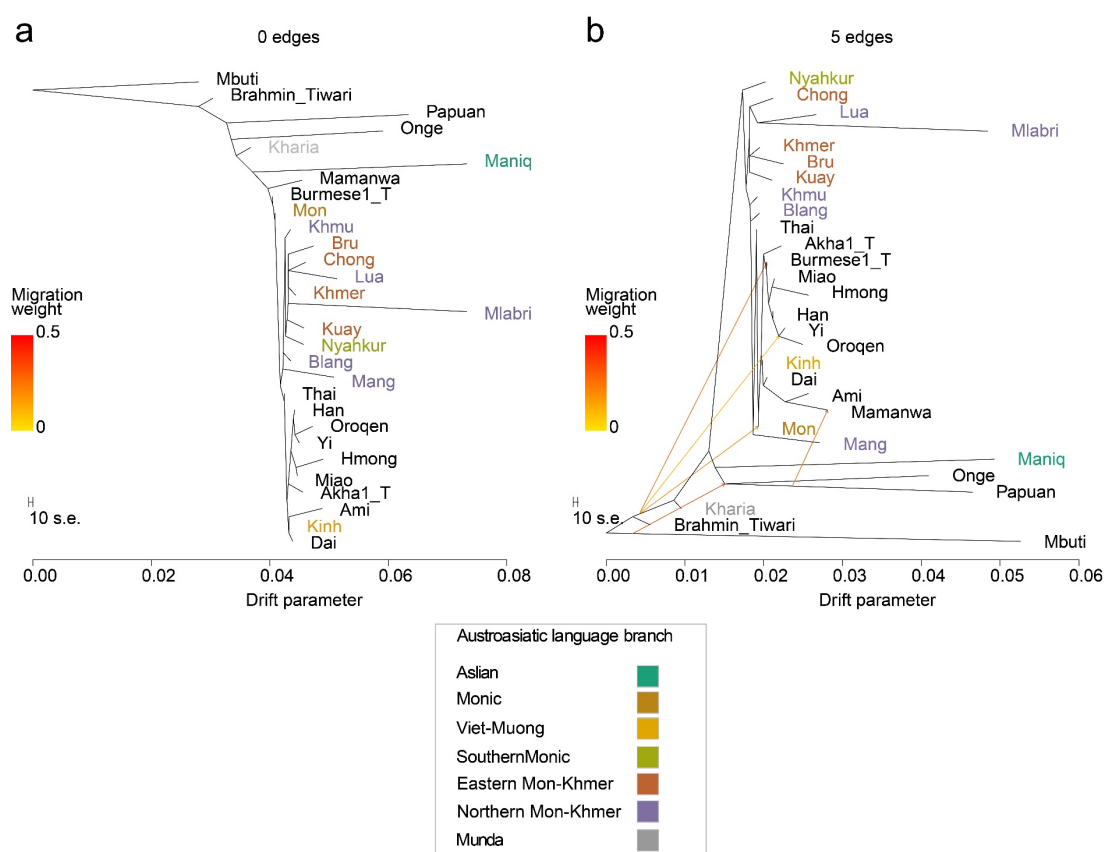


Supplementary Fig. 15 | Biplot of $f_4(W, \text{Maniq}; \text{South Asian/East Asian/Southeast Asian, Mbuti})$ statistics with transversos only SNPs. For ancient MSEA population W ($n = 23$), we performed f_4 statistic using transversos only (95,924 SNPs) to compare the shared genetic drift between ancient MSEA populations with South Asian ancestry (Indus_Periphery, Onge) or East Asian ancestry (China_YR_MN, Longlin, Liangdao1) or Southeast Asian ancestry (Laos_Hoabinhian.SG). The red-colored solid line represents a linear fit of the plotted f_4 statistics tests, with the gray shading representing the 95% confidence interval.

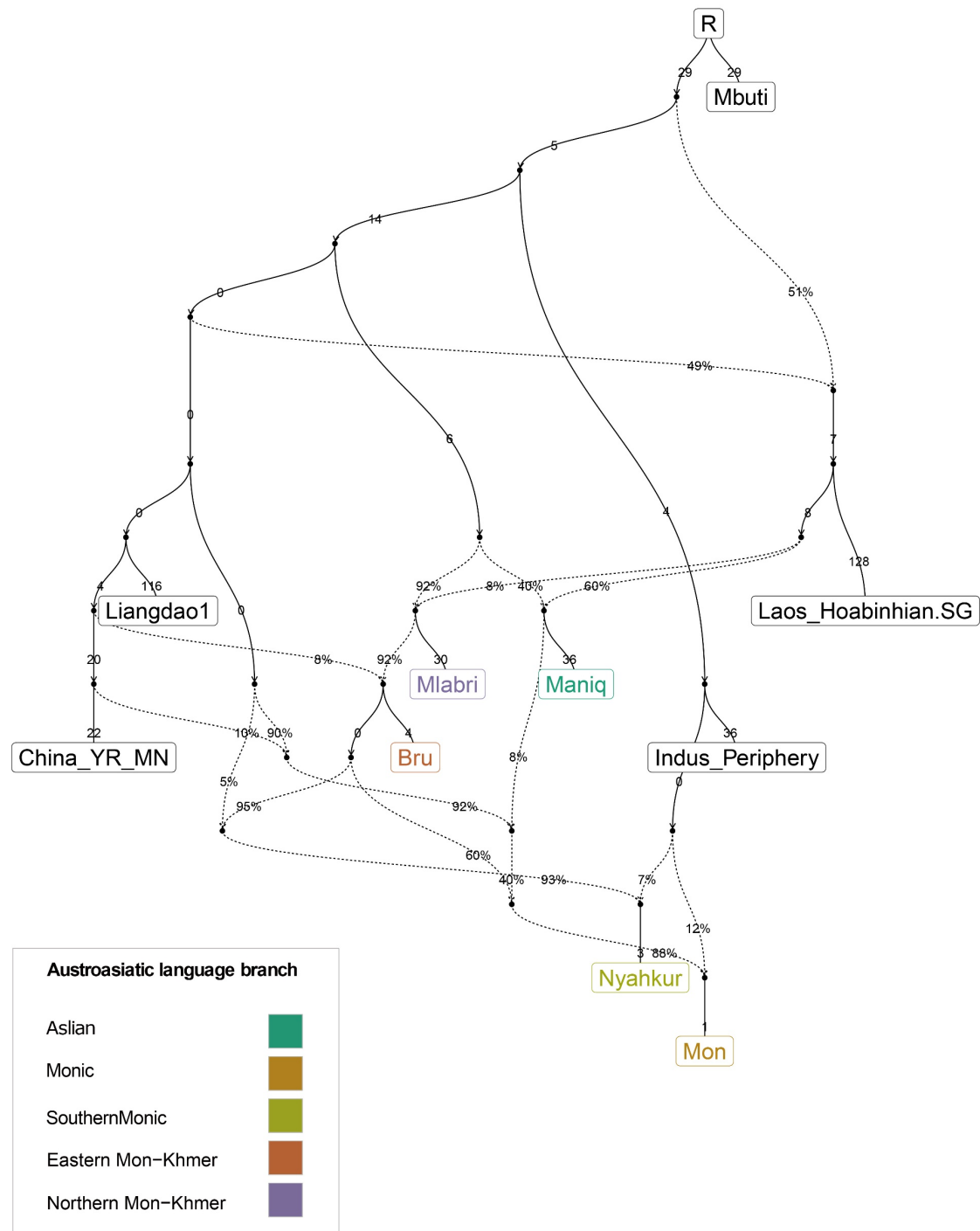


Supplementary Fig. 16 | Admixture graphs modeling with *qpGraph* for ancient MSEA, South Asian and East Asian groups. Dashed arrows represent admixture edges, whereas solid arrows are drift edges reported in the unit of $F_{ST} \times 1,000$. Laos_Hoabinhian represents ancient MSEA hunter-gatherers; China_YR_MN represents ancient northern East Asians; Liangdao1 represents ancient southern East Asians; Longlin represents ancient southern East Asian hunter-gatherers; Indus_Periphery_West refers to an Iranian farmer-related individual contributing to the Neolithic South Asian ancestry; India_Harappan corresponds to the ancestry of the Indus Valley Civilization (IVC); Central_Steppe_MLBA represents the Yamnaya-related ancestry component (associated with the Ancestral North Indians, ANI); and Onge stands in for Andamanese hunter-gatherer-related Ancestral South Indian (AASI) ancestry. We modeled

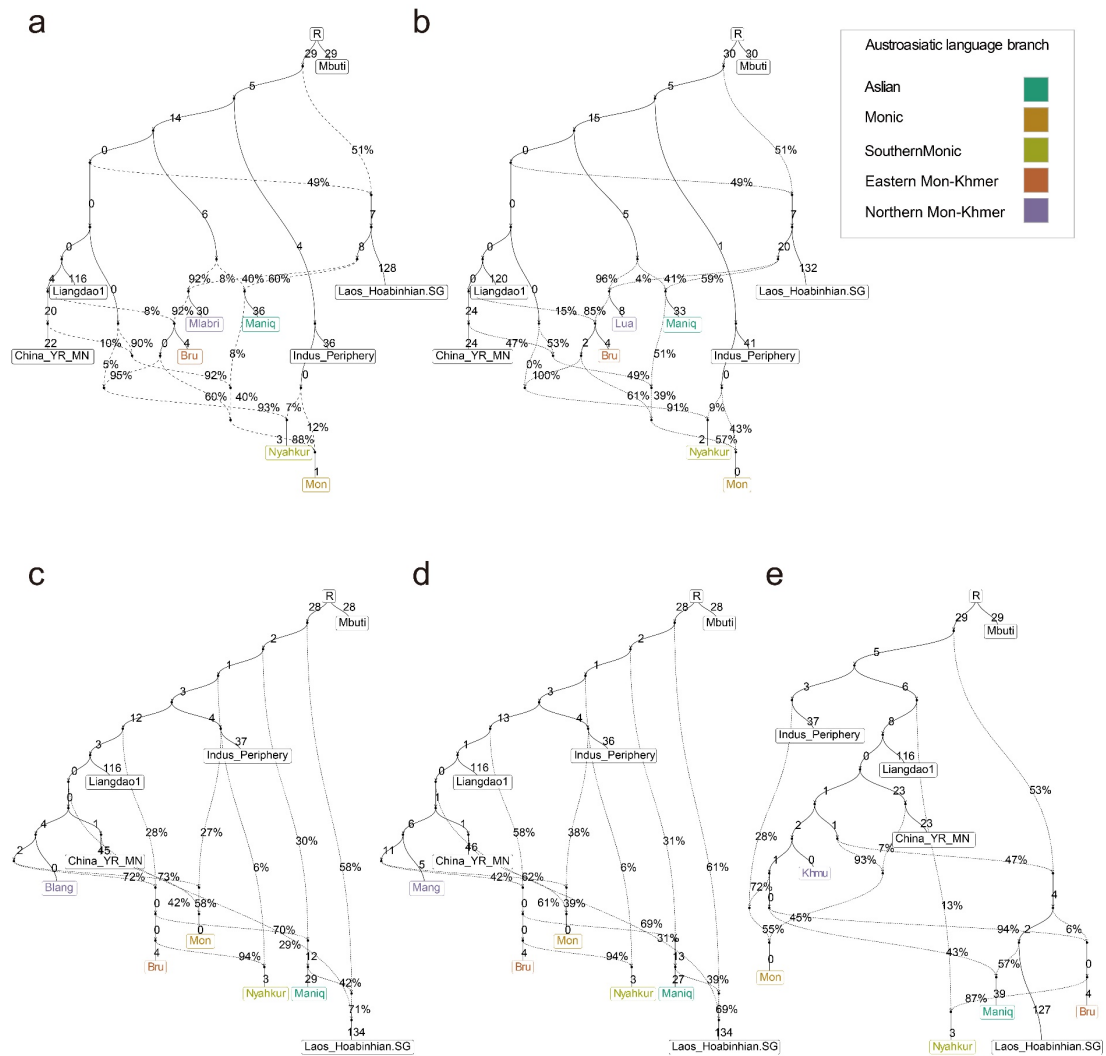
ancient MSEA populations as admixtures between northern East Asians and a deeply diverged lineage related to both MSEA and southern East Asian hunter-gatherers, and closely related to AASI. Model fit statistics for each population are as follows: **a** Z-score of 2.241 and residual of 0.008843 for N-ManBac; **b** Z-score of 2.783 and residual of 0.022568 for N-MaiDaDieu; **c** Z-score of -2.417 and residual of 0.01827 for BA-NuiNap; and **d** Z-score of 2.904 and residual of 0.024183 for IA-ThamLodCave. We further tested the robustness of this topology by adding modern AA groups: **e** Z-score of 2.411 and residual of 0.012489 when Mlabri was added as part of the same clade as N-ManBac; **f** Z-score of 2.470 and residual of 0.01250 when Mon was added as a mixture of Mlabri-related AA ancestry and Ancestral South Indian (ASI) ancestry; **g** Z-score of -2.228 and residual of -0.018764 for the backbone topology before adding any test MSEA groups.



Supplementary Fig. 17 | Inferred maximum-likelihood tree using TreeMix for Austroasiatic speakers and related South/East Asians. **a** The estimated tree typology without migration events; **b** The estimated tree typology with five migration events. This presented typology is identified as the optimal tree model based on the OptM estimation.



Supplementary Fig. 18 | Admixture graphs modeling with *qpGraph* for Austroasiatic speakers and related ancient populations. Dashed arrows represent admixture edges, whereas solid arrows are drift edges reported in the unit of $F_{ST} \times 1,000$. In the fitted admixture graph, the worst-fitting statistic test has a Z-score of 2.494 and a residual of 0.001524. Laos_Hoabinhian represents for ancient Mainland Southeast Asians; Indus_Periphery for ancient South Asians; China_YR_MN for ancient northern East Asians, and Liangdao1 for ancient southern East Asians. Modern Austroasiatic speakers are colored by their respective language branch.



Supplementary Fig. 19 | Admixture graphs modeling with *qpGraph* for Austroasiatic speakers and related ancient Asian populations. Dashed arrows represent admixture edges, whereas solid arrows are drift edges reported in the unit of $F_{ST} \times 1,000$. Laos_Hoabinhian represents for ancient Mainland Southeast Asians; Indus_Periphery for ancient South Asians; China_YR_MN for ancient northern East Asians, and Liangdao1 for ancient southern East Asians. Modern Austroasiatic speakers are colored by their respective language branch. Within the Northern Monic-Khmer branch, Mlabri and Lua fit the same topology, with the worst-fitting statistic test yielding **a** Z-score of 2.494 and residual of 0.001524 for Mlabri and **b** Z-score of 2.675 and residual of 0.000715 for Lua, respectively, while Blang and Mang fit another topology, with **c** Z-score of 2.186 and residual of 0.00029 for Blang and **d** Z-score of 2.527 and residual of 0.000388 for Mang. Khmu fits a third topology, with a worst-fitting statistic test of **e** Z-score of 2.006 and residual of 0.001059.

Supplementary References

1. Pittayaporn P. Layers of Chinese Loanwords in Protosouthwestern Tai as Evidence for the Dating of the Spread of Southwestern Tai. (2014).
2. Schliesinger, J. *Ethnic Groups of Thailand: Non-Tai-Speaking Peoples*. (White Lotus Press, Bangkok, 2000).
3. Palamara, P. F., Lencz, T., Darvasi, A. & Pe'er, I. Length Distributions of Identity by Descent

- Reveal Fine-Scale Demographic History. *Am. J. Hum. Genet.* **91**, 809–822 (2012).
4. Yang, M. A. *et al.* Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* **369**, 282–288 (2020).
 5. McColl, H. *et al.* The prehistoric peopling of Southeast Asia. *Science* **361**, 88–92 (2018).
 6. Eberhard, D., Simons, G. & Fennig, C. *Ethnologue: Languages of the world*. Twenty-third edition. Dallas, texas: SIL international. (2020).