UCLA
UCLA Previously Published Works

Title

Reconstructing the ancestral gene pool to uncover the origins and genetic links of Hmong-Mien speakers.

Permalink

https://escholarship.org/uc/item/4557m7h5

Journal

Journal of Biology, 22(1)

Authors

Gao, Yang

Zhang, Xiaoxi

Chen, Hao

et al.

Publication Date

2024-03-13

DOI

10.1186/s12915-024-01838-9

Peer reviewed

Reconstructing the ancestral gene pool to uncover the origins and genetic links of Hmong–Mien speakers

Yang Gao[1,2,3†], Xiaoxi Zhang[2,3†], Hao Chen[3], Yan Lu[1], Sen Ma[3], Yajun Yang[4], Menghan Zhang[5] and Shuhua Xu[1,2,3,4*]

Abstract

Background Hmong–Mien (HM) speakers are linguistically related and live primarily in China, but little is known about their ancestral origins or the evolutionary mechanism shaping their genomic diversity. In particular, the lack of whole-genome sequencing data on the Yao population has prevented a full investigation of the origins and evolutionary history of HM speakers. As such, their origins are debatable.

Results Here, we made a deep sequencing effort of 80 Yao genomes, and our analysis together with 28 East Asian populations and 968 ancient Asian genomes suggested that there is a strong genetic basis for the formation of the HM language family. We estimated that the most recent common ancestor dates to 5800 years ago, while the genetic divergence between the HM and Tai–Kadai speakers was estimated to be 8200 years ago. We proposed that HM speakers originated from the Yangtze River Basin and spread with agricultural civilization. We identified highly differentiated variants between HM and Han Chinese, in particular, a deafness-related missense variant (rs72474224) in the *GJB2* gene is in a higher frequency in HM speakers than in others.

Conclusions Our results indicated complex gene flow and medically relevant variants involved in the HM speakers' evolution history.

Keywords Reconstructing genomes, Hmong-Mien, Next-generation sequencing, Genomic diversity, Local adaptation

†Yang Gao and Xiaoxi Zhang contributed equally to this work.

*Correspondence:
Shuhua Xu
xushua@fudan.edu.cn
[1] State Key Laboratory of Genetic Engineering, Human Phenome Institute, Zhangjiang Fudan International Innovation Center, Center for Evolutionary Biology, Department of Liver Surgery and Transplantation, Liver Cancer Institute, Zhongshan Hospital, Fudan University, Shanghai 200032, China
[2] School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China
[3] Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[4] Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, Fudan University, Shanghai, China
[5] Institute of Modern Languages and Linguistics, and Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China

Gao *et al. BMC Biology* (2024) 22:59

Page 2 of 22

Main text

Background

The Hmong–Mien (HM), also known as Miao–Yao, speakers are a group of linguistically related people living primarily in the mountains of Southern China and Southeast Asia. The majority of the HM speakers living in China include three main populations belonging to the Hmongic branch (Miao and She) and the Mienic branch (Yao) [1]. Previous studies on the genetics of HM populations mainly focus on the Y chromosome and mitochondrial DNA (mtDNA). The genetic difference between the Hmong and Mien populations has been found. The Hmong populations had more contact with the northern East Asians [2]. A specific and highly frequent Y haplotype group (O-M7) that existed in the modern HM population has been found in the ancient DNA samples from the Daxi site in the middle reach of the Yangtze River 5300–6400 years ago, which suggests that the Daxi people might share ancestry with modern HM populations [3]. However, the O-M7 haplotype also appeared in a group of Mon–Khmer speakers in the investigation of a larger sample [4]. The history of the HM population may not be simple. In the autosomal study, the HM population was found to carry a specific genetic component [5, 6] and receive gene flow from the ancestors of southern East Asians [7] and Sino–Tibetan-related ancestry [6]. However, including some other recent studies [8–11], genetic research in the HM population is mainly based on sparse single-nucleotide polymorphisms (SNPs), with the Miao population as the main population and the lack of Yao population data. Limited by data, the demographic history model and adaptive evolution of the HM populations are still not comprehensive and clear. At present, only one ancient individual (500 years old) related to the HM population has been discovered in Guangxi [12], which limits the study of HM history. Nonetheless, HM language populations still have been under-investigated in recent genomic studies using whole-genome sequencing [13, 14]; in particular, these studies only involve whole-genome sequencing data from the Miao and She populations but lack Yao population, which limits a full investigation of the ancestral origins and evolutionary history of HM speakers.

We collected 80 Yao blood samples from Guangxi Province, where more than half of the Yao people reside, and sequenced all of them to high coverage (30×) (see Additional file 1: Table S1) (see "Methods"). After quality control, 12.8 million high-quality variants were called autosomes, with more than 80% of them being biallelic single-nucleotide variants (SNVs) (see Additional file 2: Text S1). Through dbSNP (version 154) annotation, a total of 504,927 novel variants were discovered, accounting for 4.92% of the total number of biallelic SNVs.

Further functional annotation indicated strong biological effects of 2889 novel variants. We also identified 4423 fragments of Yao with a unique or specific archaic infiltration with a total length of 71.56 MB using Archaicseeker2.0 [15] (Additional file 2: Text S2; Additional file 1: Tables S2-6). We integrate the Yao dataset with the public dataset of other populations (Additional file 2: Text S3, Fig. S1), such as Miao and She. Taking advantage of these datasets, we attempted to reveal the genetic structure of HM populations and provide genetic evidence for the origin of HM populations.

Results

The genetic structure of the present-day HM population

All 80 Yao samples were collected from Guilin and Laibin. These two subgroups are closely clustered in the phylogenetic tree (Additional file 2: Fig. S2). The focus of this study is on the history of the HM population on a larger scale. Therefore, we regard two subgroups as a group in the analysis. To determine the genetic coordinates of present-day HM speakers, we performed principal component (PC) analysis (PCA) (see "Methods," "PCA"). In the Eurasian context (Additional file 2: Fig. S3), three HM groups formed a cluster. HM speakers were located at the end of the East Asian cluster and far away from the European and Siberian clusters, indicating that HM speakers are typical East Asian descendants. In the context of East Asia (Fig. 1a), we applied linear fitting to project the genetic coordinates to the geographical coordinates of locations where population samples were collected. PC1 fitted the latitude well (Fig. 1b), while PC2 fitted the longitude well (Fig. 1c). Compared with the Han Chinese population, the HM speakers were located more in the south of East Asia and clustered together with Sino–Tibetan speakers and Tai–Kadai speakers. HM speakers were mainly distributed along the north–south axis. The Miao and She populations were located in the northern part of the HM language family, while the Yao population was located in a more southern part (Fig. 1a).

We next dissected the ancestry composition of HM speakers (see "Methods," "ADMIXTURE"). When the number of ancestral populations (K) was assumed to be five, the results exhibited a low cross-validation error compared to others (Additional file 2: Fig. S4c). East Asian populations were mainly composed of the northern component, represented by Nganasan, and the southern component, represented by Taiwan Aboriginal. The Yao population harbored a higher southern genetic component than the Miao and She populations and the ancestral compositions of She and Miao were almost identical (Fig. 1d). These patterns are consistent with that shown in PCA (Fig. 1a). The analysis of shared genetic drift showed that Miao and She shared the most
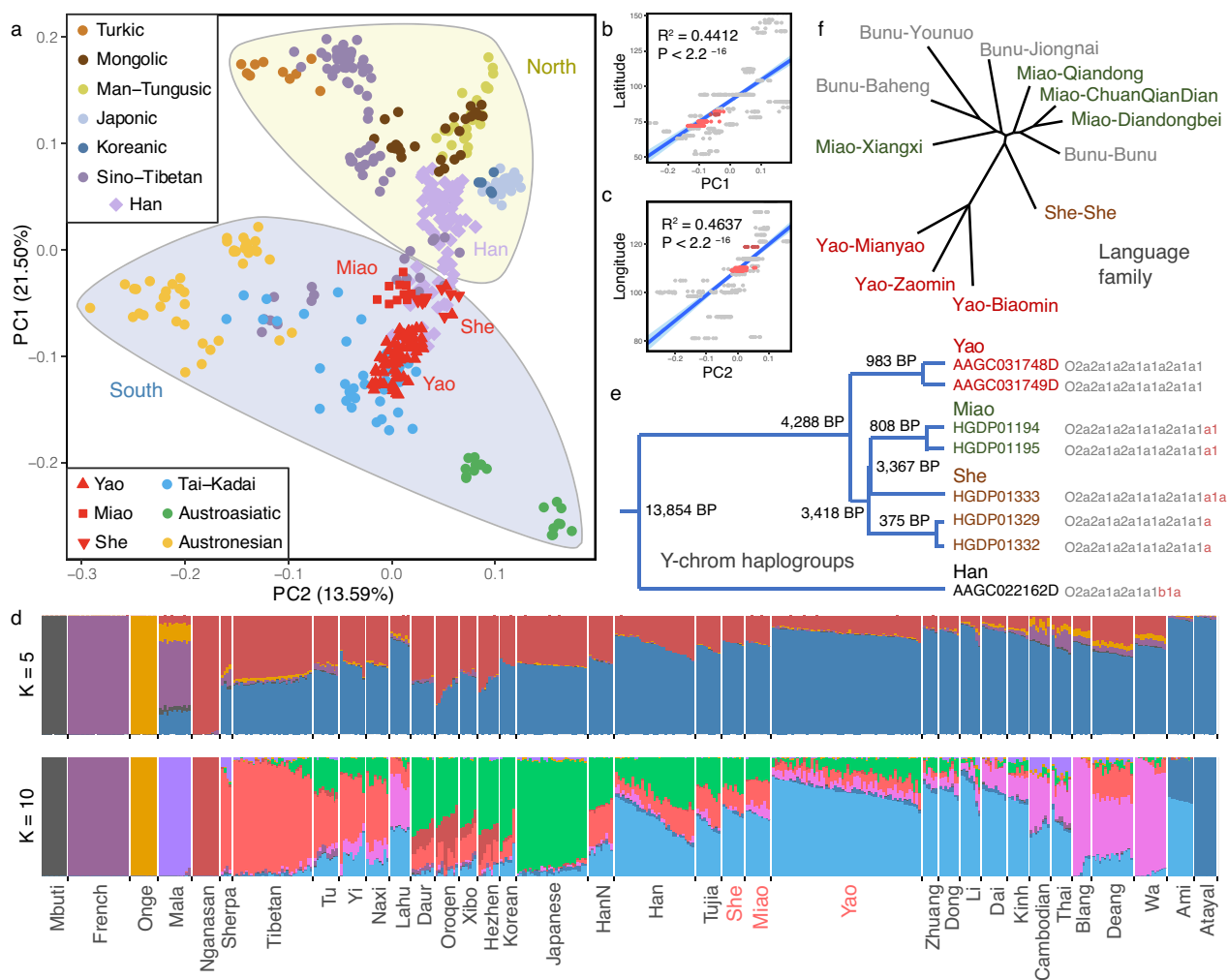
Gao *et al. BMC Biology*   (2024) 22:59

Page 3 of 22

Fig. 1 Genetic diversity among Hmong-Mien-speaking (HM) subpopulations. **a** The hierarchical genetic coordinates of HM speakers by PCA in the context of East Asia. The north–south population is divided by the geographical boundary, along the Qinling Mountains—Huai River line. **b** Linear fitting between PC1 and latitude. The red dots mean HM groups. The regression lines fitted using all the groups in the PCA. The blue shadow is a 95% confidence interval. **c** Linear fitting between PC2 and longitude. **d** The ancestral component inference of the HM speakers by ADMIXTURE software. **e** Paternal genetic model and divergence time estimation of HM population by Y chromosome O-M7 haplogroup. **f** The genealogy tree of the HM language family was constructed by the distance matrix generated by the comparison of basic words

genetic drift (Additional file 2: Fig. S5) (see "Methods," "Outgroup F3"). Therefore, the genetic evidence obtained so far supports that the split of Yao from other HM groups occurred earlier than that between She and Miao. We estimated that the divergence time between Yao and She was 5790 years by MSMC [16] and MSMC-IM [17], and we also estimated that the divergence time of Yao and Miao to be 5899 years and that of Miao and She was 7197 years (Additional file 2: Fig. S6a) (see "Methods," "Divergence time"). These divergence patterns inferred with MSMC were not concordant with the genetic structure of the HM population. We speculate that the divergence time might have

been affected due to population isolation or recent gene flow from peripheral populations.

Previous studies have shown that O2a2a1a (O-M7) is found in high frequency in some HM populations [18]. Daxi relics (5300–6400 years ago) in the middle reaches of the Yangtze River that carry this haplogroup are also considered to be related to the ancestors of modern HM populations [3]. Therefore, O-M7 can be considered as a specific Y haplogroup of HM speakers. In our analysis (see "Methods"), O-M7 was indeed enriched in the HM sub-branch groups (She: 3/7; Miao: 2/7; Yao: 2/44) but absent or rare in non-HM populations (Tibetan: 0/18; Dai: 0/6; Han: 1/21) (Additional

Gao *et al. BMC Biology*   (2024) 22:59

Page 4 of 22

file 1: Table S7-8). Eight O-M7 samples could be further classified into different subgroups (Fig. 1e). Unlike the Han Chinese population, the seven HM populations belonged to the O-N5 lineage. Consistent with the results obtained from the autosomal data, the evidence from the Y haplogroup analysis also supports the closer genetic relationship between Miao and She. Considering the specificity of O-N5 in the HM speakers, we estimated the time of the most recent common ancestor (MRCA) of this haplogroup in the HM speakers to be around 4288 years ago and that between Miao and She to be around 3418 years ago (Fig. 1e, Additional file 2: Fig. S7).

To further examine the relationship between genetic affinity and linguistic affinity, we calculated the distance between several branches of the HM language family using the similarity of basic words and by constructing the language tree based on the distance matrix (see "Methods"). The results show that the Miao and Yao language branches clustered respectively (Fig. 1f, Additional file 1: Table S9). Suffixes represent branches within a language family. Among them, the languages spoken by the Miao and Bunu populations belong to the Hmongic language branch of the HM language family, and the languages spoken by the Yao population belong to the Mienic language branch (Additional file 2: Text S4). The distance between the She and Miao languages was closer than that between the She and Yao languages, indicating that the Yao language split first, followed by the She and Miao languages. These patterns are highly consistent with the genetic data.

On a finer scale, we also examined the genetic diversity among the HM branches. Overall, the HM population showed lower genetic diversity than the Han Chinese and Tibetan populations belonging to the Sino–Tibetan language family (Additional file 2: Fig. S8). Among them, the genetic diversity of the Yao population was equivalent to that of Dai belonging to the Tai–Kadai language family, and higher than that of the She and Miao populations. The genetic diversity of the She population was the lowest among the three HM populations. From the historical changes in the effective population size, both She and Miao populations experienced a bottleneck event after the divergence of the HM populations (Additional file 2: Fig. S9), which may partly explain the differences in genetic diversity observed. We also found the sex-biased admixture in the HM population, where the admixture of HM populations was prone to combinations of southern males and northern females (Additional file 2: Fig. S10). Moreover, we did not observe HM-specific mtDNA haplogroups (Additional file 1: Table S10-11).

Admixture-driven differentiation of HM subpopulations

Our further analysis indicated that gene flow from surrounding populations played an important role in shaping the genetic structure within the HM population. First, Yao showed a closer relationship with Tai–Kadai speakers (Fig. 2a, Additional file 1: Table S12), especially Zhuang, compared with Miao and She (Fig. 2b,c, Additional file 2: Fig. S5). PCA in the context of Southern East Asia (Additional file 2: Fig. S11) showed that Dong and Zhuang from the Tai–Kadai language family were the two populations with the closest relationship to Yao. These patterns were also confirmed by GLOBETROT-TER [19] analysis, that is, the present-day Yao population was affected by gene flow from the ancestors of Zhuang, Han, and Miao by about 1100 years (Additional file 2: Text S5). The extra affinity between Yao and Zhuang may have resulted from the geographical overlapping of Yao and Zhuang, which can be seen from the sample distribution map (Fig. 2b&c). Second, Miao was more influenced by Tujia than She (Fig. 2d) and more influenced by northern populations than Yao (Fig. 2b), while She was relatively isolated and had fewer genetic connections with other populations (Fig. 2a,c,d). The isolation of the She population can also be confirmed with the analysis of the run of homozygosity (ROH) (Additional file 2: Fig. S12) and the rare allele sharing (Additional file 2: Text S6). This pattern can be largely explained by the geographical distribution of these populations (Fig. 2c,d). These results reflect recent genetic admixtures that occurred between the HM and surrounding populations. In particular, the three HM populations all showed a close relationship with the Han Chinese population. The genetic differentiation measured by the F_{ST} of each HM population with the Han population was even smaller than that of any pair of the HM populations (Fig. 2a), and they share quite a lot of rare variants with the Han Chinese population (Additional file 2: Text S6). The results of MSMC-IM also supported the recent gene flow between the three HM subpopulations and the Han population (Additional file 2: Fig. S6b).

Genetic origins of the HM population

To gain further insight into the genetic history of the present-day HM populations, we tried to explore the origin of the HM population. The main ancestral source of the HM speakers was the southern populations (>70%) (Fig. 1d). In the higher dimension of admixture analysis ($K=10$), the genetic components of East Asia were further subdivided. The southern component was divided into the southeast islands' component (deep blue), represented by the Austronesian speakers; the southwest inland component (purplish-red), represented by the Austroasiatic speakers; and the South China components
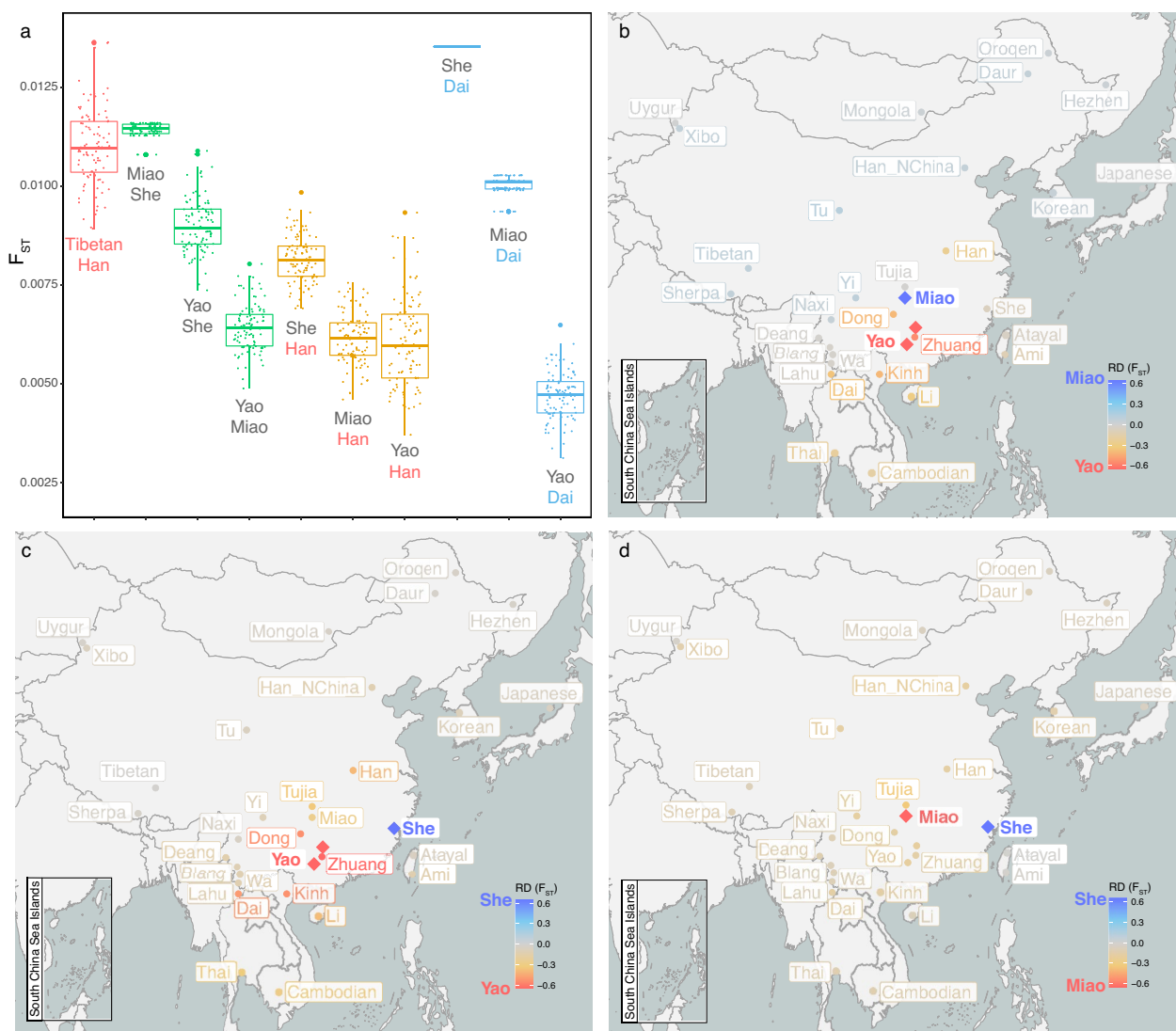
Gao *et al. BMC Biology*   (2024) 22:59

Page 5 of 22

Fig. 2 Gene flow from surrounding populations promoted the formation of a subpopulation structure within the HM population. **a** F_{ST} of the pairwise population was the result of randomly selecting 9 samples and repeating it 100 times. Each point represents a repetition. **b–d** The diamonds denote two target populations, and the other dots denote the populations tested. The Yao samples come from two sites, Guilin and Laibin. The color represents the relative difference in F_{ST} between the population to be tested and the two target populations (see "Methods"). The results show that surrounding populations contributed gene flow to the HM populations

(blue), shared by the HM and the Tai–Kadai speakers. The proportion of the South China genetic composition was more than 50% in HM and Tai–Kadai speakers, and they were probably derived from the same ancestral population (Fig. 1d). We estimated the divergence time between the HM population and the Tai–Kadai population (see "Methods," "Divergence time"). To minimize the influence of recent gene flow, we used the relatively isolated She population to represent HM speakers and estimated the divergence time between the She and Dai as 8200 years (Additional file 2: Fig. S6c). We also estimated that the divergence time between HM and Han

was 10,800 years based on the historical change of effective population size (N_e). Therefore, HM and Tai–Kadai have a closer genetic relationship compared with the Han Chinese population.

Furthermore, the South China component shared by HM and TK ancestors could also be observed in the Han Chinese population (Fig. 1d). The South China component showed a high proportion in the southern and northern Han Chinese populations (46.52%, 22.39%) but not in the Tibetan people of the Sino–Tibetan language family. These results suggest that the infiltration of the South China component into the Han

Gao *et al. BMC Biology* (2024) 22:59

Page 6 of 22

Chinese population occurred after the Han-Tibetan divergence and before the expansion of the Han Chinese population. Through further analysis, we found that the Han Chinese population shared more genetic drift with HM populations than with Tai–Kadai populations (Additional file 1: Table S13). Therefore, the South China components in the Han Chinese population may have originated from the common ancestor of HM speakers.

We further built a fine evolution model of HM speakers with the qpGraph in ADMIXTOOLS2 [20] (see "Methods"). Based on the score of the model generated by the automatic search, we speculate that at least three genetic admixture events have occurred in the HM, TK, and ST populations (Additional file 2: Text S7). We designed the preliminary skeleton of the model based on our conclusion. According to the score, branch length, admixture ratio, and other information of alternative models (Additional file 2: Text S7, Fig. S13), the preliminary model was further fine-tuned, and finally, the software scored our model as 7.53×10^{-6}, giving strong support to our model (Fig. 3, Additional file 2: Fig. S13b). According to the model, HM and Tai–Kadai populations shared the MRCA. About half of the genetic components of the early Han Chinese population came from the MRCA of the HM speakers, which may explain the substantial genetic differences between the Han and the Tibetan

populations. The Yao was first separated from the HM speakers and was influenced by the Tai–Kadai speakers. Affected by the expansion of the Han Chinese population, the present-day HM populations independently received the gene flow from the Han Chinese populations of slightly different genetic backgrounds. Compared with Miao and Yao, She experienced a longer period of population isolation.

Reconstruction of HM ancestral genomes

The ancestors of HM may have played an important role in the history of East Asia. However, the lack of sufficiently ancient HM-representative samples has limited a deeper analysis. In previous studies [5, 6], a special genetic component shared by HM populations was observed, but the corresponding local genomic regions were not identified. Here, we designed a scheme to extract ancestral fragments from the genomes of present-day HM populations for the reconstruction of the ancestral HM genomes (Fig. 4) (see "Methods"). We first identified the ancestral fragments in the genome of present-day HM populations using local ancestral inference [21] (see "Methods"). To dissect the recent linkage disequilibrium, all the obtained ancestral fragments were broken into segment sizes of 6 kb, and the ancestral genome was reconstructed by random sampling
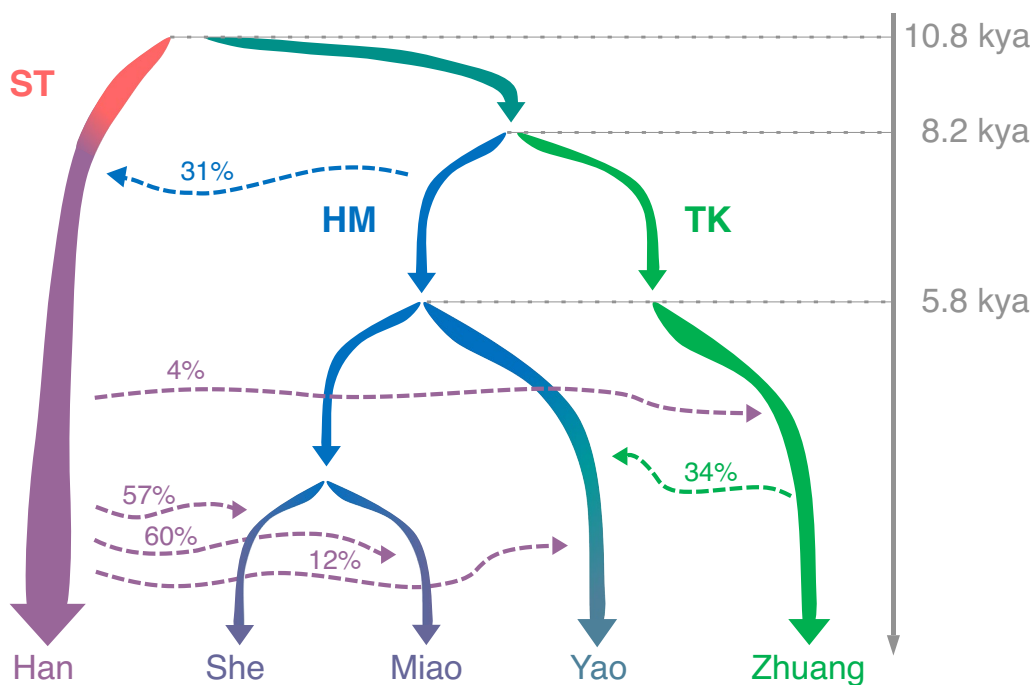

Fig. 3 A fine origin history model of the HM speakers by qpGraph. The basic skeleton was obtained through the analysis of the above group structure and history and then further adjusted according to the score. The final score was 7.53×10^{-6}. The dotted line represents the admixture events, and the percentage was the infiltration proportion of the current admixture event. The divergence time in the figure is the result of MSMC-IM analysis
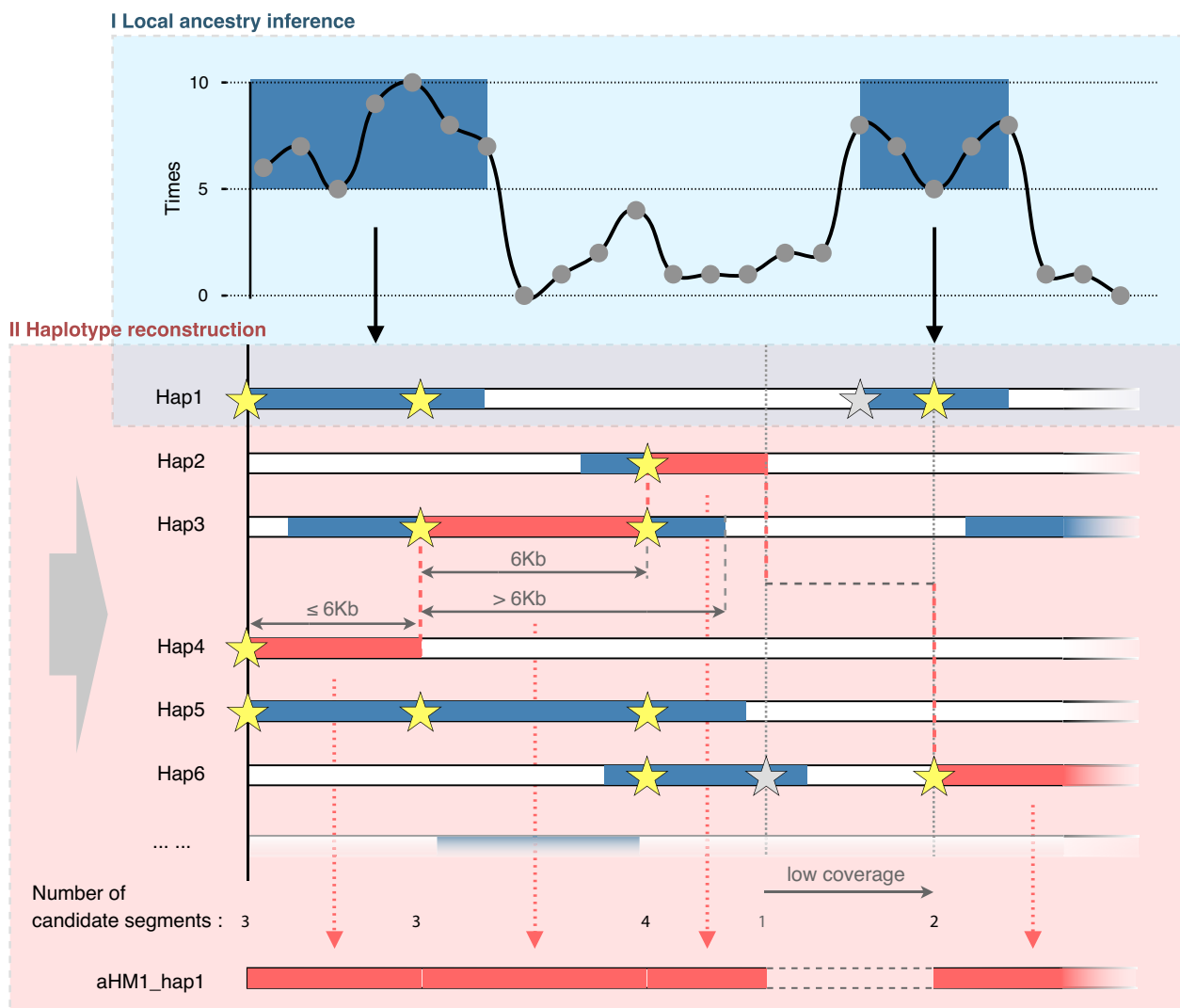
Fig. 4 A workflow for reconstructing the ancestral genome based on present-day populations. I. Local ancestry inference of modern human genome. The segments inferred as HM in 5 of the 10 results were treated as candidate segments. Grey dots denote variants. II. Assembly of HM ancestral genomes using candidate segments. Each time, a segment was randomly selected from the optional starting sites and extended back no more than 6 kb. The low-coverage regions of the ancestral segments were skipped. The pentagrams denote the starting sites for candidates.

with replacement. In addition, to avoid the diversity bias in local regions resulting from uneven coverage of the genomic pool, we treated the genomic regions with low coverage of ancestral segments as missing data. Eventually, we constructed 30 high-fidelity HM ancestral (aHM) genomes using this method.

Next, we evaluated the quality of reconstructed aHM genomes. It turned out that the coverage of these reconstructed genomes was about 74.73% of the whole genome, a great gain compared with the traditional single-nucleotide polymorphism array and a typical ancient DNA sequencing effort. A significant positive correlation was observed between identity-by-descent (IBD) analysis

and local ancestry inference [22] (Fisher's exact test, $P < 2.2^{-16}$) (Additional file 2: Text S8), indicating that the result of local ancestral inference is reliable (see "Methods"). The inbreeding analysis also showed that no pair of samples had a relationship of the third degree or closer in all the reconstructed ancestral genomes (see "Methods"). The reconstructed aHM genomes showed good quality. Analysis of shared genetic drift between the reconstructed ancestral genome and 29 modern East Asian populations [23] showed that the present-day Hmong-Mien populations and the reconstructed ancestral population shared the most genetic drift (Additional file 2: Fig. S14) (see "Methods," "Outgroup F3"). Moreover, the

results of ancestral inference show that the proportion of the genetic components of reconstructed aHM genomes in the ancient DNA samples (Gaohuahua, around 500 ya) [12] related to the HM population in Guangxi was higher than that in the present-day HM population (Additional file 2: Fig. S15). The ancient Gaohuahua genomes have been reported to share the most genetic drift with HM-speaking groups [12]. Thus, reconstructed aHM genomes largely represent the ancestors of HM speakers or proto-HM populations.

Ancient founders in the Yangtze River Basin

We further analyzed the genetic architecture of the proto-HM populations together with present-day populations. With the high coverage of aHM genomes constructed, we performed PCA of the aHM genomes with modern Eurasian populations (see "Methods"). The aHM showed distinction from the present-day HM populations on PC2, corresponding to a north–south geographical differentiation (Additional file 2: Fig. S16). The geographic coordinates of the aHM were more southeast than those of the other present-day East Asian populations. This pattern can be explained by the persistent influence of northern East Asian ancestry in southern East Asia [24].

Next, we performed an ADMIXTURE analysis of combined data of the aHM and present-day populations assuming a different number of ancestral populations ($K=2$–15) (Additional file 2: Fig. S17) (see "Methods"). Interestingly, the aHM showed a single pure genetic component throughout all of the ADMIXTURE analyses. Assuming three ancestral populations ($K=3$), our reconstructed aHM genomes represented a genetic component of Southern East Asia. When $K=6$, this genetic component of aHM was in a low frequency among all of the East Asian populations, suggesting that the aHM population may be one of the founder ancestral groups of East Asians. In addition, the aHM component showed a higher frequency in southern populations than in northern populations, and even slightly higher in the present-day HM populations, the Tai–Kadai populations, the Taiwanese aboriginal people, and the Han Chinese population. These results suggest that the aHM population likely originated in the southeast region of East Asia.

In contrast to the ancient DNA samples of known geographical location but unknown ethnic information, our reconstructed aHM genomes provided a more explicit message in tracing human migration history. We integrated 29 East Asian populations with the reconstructed HM ancestral population as a reference dataset as well as 968 ancient DNA samples located in East, Southeast, and Central Asia [24–38]. By analyzing the shared genetic drift of each ancient sample and each reference population [23], we identified eight high-quality ancient samples

that shared the most genetic drift with the aHM genomes (Fig. 5, Additional file 1: Tables S14-15) (see "Methods"). These samples mainly lived about 3000–4000 years ago, that is, the period after the divergence of HM populations. The geographical locations of these ancient samples indicated the habitation of the aHM population and the dispersal routes since the initial divergence. In brief, the aHM population was mainly distributed in the south of East Asia but later arrived in Thailand in the South and the lower reaches of the Yellow River Basin in the north. The aHM population was likely to live between the two places, that is, in the middle and lower reaches of the Yangtze River Basin. A study based on the Y chromosome in Daxi Culture also identified traces of HM ancestors' activities in the Yangtze River Basin [3]. According to the dating of eight ancient DNA samples, the earliest sample appeared in the southeast coastal area of China around 4300–4400 years ago. The ancient DNA samples found in the China–Indochina Peninsula and the Yellow River Basin also provided evidence that the aHM population spread in East Asia to both directions, north and south.

Footprints of natural selection in the HM population

We separately extracted 10 samples from the Yao, Miao, and She populations to construct the HM population group, and calculated F_{ST} with the Han population. A total of 13,180 variants (Additional file 1: Table S16, Additional file 2: Fig. S18) were identified based on between-population analysis using site-specific F_{ST} (top 0.1%), of which 35.30% were expression quantitative trait loci (eQTL), which is much higher than the average density of eQTL in the whole genome (15.47%) ($P<2.2\times10^{-16}$; Additional file 2: Text S8). There were 1854 genes covered by the 13,180 variants significantly enriched in multiple regulatory pathways and the composition of protein structure such as protein binding (FDR $P=2.5\times10^{-58}$), plasma membrane (FDR $P=7.2\times10^{-36}$), and obesity-related traits (FDR $P=5.1\times10^{-22}$) (Additional file 1: Table S17). In addition, we also found that some variants associated with disease risk were different among populations. For example, rs7756992, a variant in the *CDKAL1* gene, associated with type-2 diabetes [39], had a lower frequency in the Yao population (Yao-G:0.492; Han-G:0.625), which means that the risk of disease might be reduced. Notably, a deafness-related missense mutation rs72474224 (p.Val37 Ile) located in the *GJB2* gene showed a higher allele frequency in the southern East Asian population including HM [39] (Yao-T:0.153; Miao-T:0.250; She-T:0.278) but a lower frequency in other populations in the world (Fig. 6a) [40]. The variant rs72474224 is highly conservative (GERP: 5.21); the derived allele is located only on a specific haplotype from Southern East
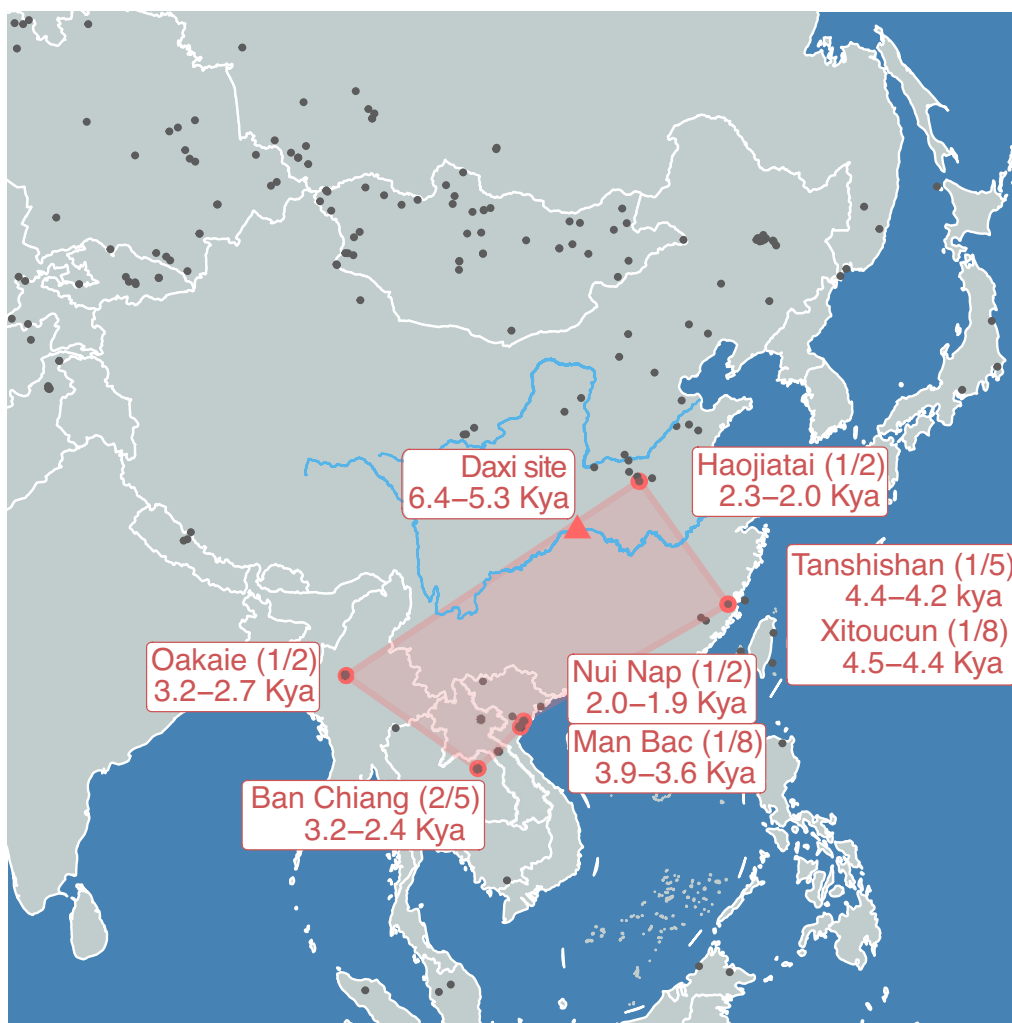
Fig. 5 A map for tracking the population migration history of the HM population. The black dots in the image are all publicly available ancient DNA samples. The red triangle denotes the Daxi site in the middle reaches of the Yangtze River. The red dot shows the eight ancient DNA samples that share the most genetic drift with the reconstructed aHM population. The numbers in parentheses represent the total number of samples in the relic and the number of samples associated with aHM. These ancient DNA samples came from people who mainly lived in southern East Asia for about 3000–4000 years, which is later than the time of the most recent common ancestor of present-day HM populations. These ancient DNA samples reflect the footprints of population diffusion after the divergence of HM populations. Ancient DNA samples showed that the ancestors of the HM population reached the China–Indochina Peninsula in the South and the Yellow River Basin in the North

Asian populations (Fig. 6b). It cannot be explained by the archaic infiltration or the founder effect (Fig. 6b). In addition, rs72474224 showed a strong selection signal in HM populations (Fig. 6c). This prevalence of the disease-risk allele in natural populations suggests the existence of pleiotropy.

To overcome the power loss in detecting natural selection due to the potential cancellation of signatures resulting from recent gene flow, we further analyzed population-specific ancestral fragments based on the reconstructed ancestral genomes. We calculated the F_{ST} of each variant between the reconstructed aHM and the Han Chinese population (Additional file 1: Table S18).

Despite there being some missing variants in the reconstructed ancestral genomes, we successfully collected a total of about 10 million SNVs shared by the ancestral and present-day populations. It turned out that this strategy increased the power for detection (Additional file 2: Fig. S19). Altogether 2371 (37.39%) of a total of 6341 extra SNVs underlying natural selection were identified as significant eQTL in the GTEx database, which is again higher than the average density of eQTL in the whole genome (15.47%). The genes covered by these additional variants were also found to be significantly enriched in multiple functional pathways such as protein binding (FDR $P = 1.5 \times 10^{-33}$) and plasma membrane

Gao *et al. BMC Biology*   (2024) 22:59
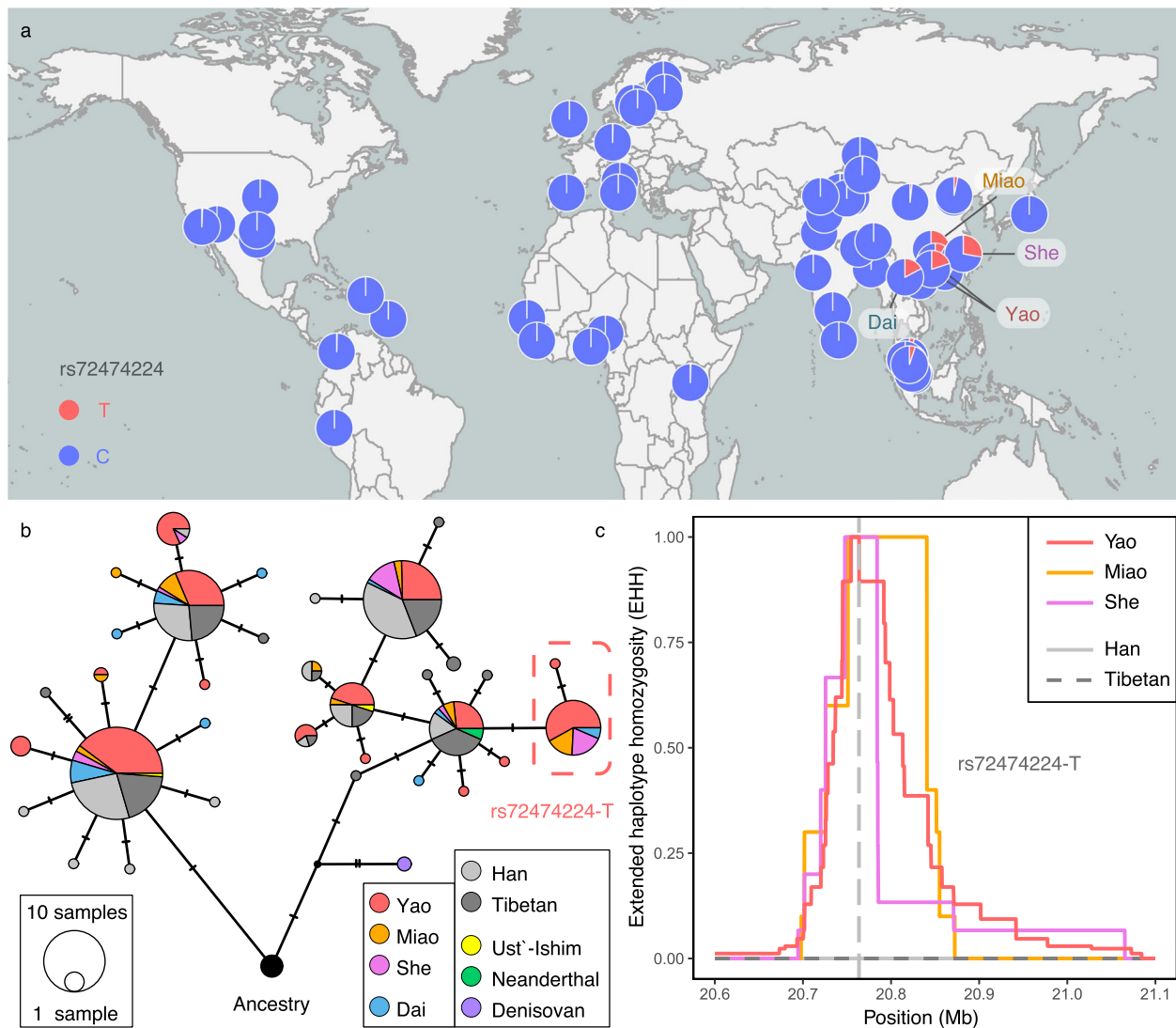
Page 10 of 22

Fig. 6 rs72474224 showed a specific positive selection signal in populations of Southern East Asia, including HM. **a** Distribution of allele frequency of rs72474224-T in the context of worldwide populations. **b** Haplotype network was constructed by PopART using the 39 SNVs located in *GJB2*. The rs72474224-T only appears on the haplotypes in the red box. **c** Extended haplotype homozygosity (EHH) of rs72474224-T. This allele showed positive selection signals in HM populations

(FDR $P = 1.6 \times 10^{-26}$) (Additional file 1: Table S19). The most significantly differentiated variants ($F_{ST:}$ 0.853) were located in the *AF146191.4* gene (lincRNA) on chromosome 4. Another significantly differentiated variant was located in the *PLCB2* gene on chromosome 15 ($F_{ST:}$ 0.847), which encodes a phosphodiesterase and participates in the signal transduction pathway of the type 2 taste receptor [41].

We also used the iHS method to identify the selection signals of the HM population. We analyzed the present-day HM population (consisting of 10 Yao, 10 Miao, and 10 She) (Additional file 2: Fig. S20a) and the reconstructed aHM population (Additional file 2: Fig. S20b) separately, and detected possible selection variants of 133,611 (Additional file 1: Table S20) and 146,604 (Additional file 1: Table S21) according to the threshold of | iHS |> 2. The intersection variants of the two populations are 9,950. In the present-day HM population, the most significant signal gene is *LINC00552*, a lncRNA. Other genes include *ADI1, HLA-DQA1,* and *SFTPA1*. In the aHM population, the most significant signal gene is *RBFOX2*. Other genes include *DDX1, SLC4A8, SMARCC1,* and *CHD9*. We also compared the selection signals found by F_{ST} and iHS methods, but the signal variants found by different methods have

Gao *et al. BMC Biology* (2024) 22:59

Page 11 of 22

differences and few intersections (Additional file 2: Fig. S21).

We also applied four methods, F_{ST} (Additional file 2: Fig. S22), iHS (Additional file 2: Fig. S23), XPEHH (Additional file 2: Fig. S24), and Tajima's D, to detect the natural selection signals of Yao population, and determined candidate selection signal regions based on 100 kb segmentation. For the results, most of the signal segments identified in each method were unique on their own, and only a small number of signals were shared with other methods (Additional file 2: Fig. S25). Thus, we found 6 sharing segments in the 3 most commonly used methods (F_{ST}, iHS, and XPEHH). These 6 shared signal segments included chr13:99400001–99800000 (Two protein-coding genes: *SLC15A1*, plays an important role in the uptake and digestion of dietary proteins [42]; *DOCK9*, associated with irregular astigmatism and corneal ectasia [43]), chr14:106000001–106100000 (Two protein-coding genes: *IGHA2* and *IGHG4*, both of them involve immunoglobulin heavy chains [44, 45]) and chr19:57500001–57600000 (No protein-coding gene. Only a pseudogene *RPL7AP69*). Furthermore, using a more relaxed threshold can help us find two signal segments sharing in all four methods including chr1:161500001–161600000 (Two protein-coding genes: *FCGR3A* and *FCGR3B*, both of them involve low affinity immunoglobulin gamma Fc region receptor [46, 47]) and chr6:32500001–32600000 (Two protein-coding genes: *HLA-DRB1* and *HLA-DQA1*, both of them involve HLA class II histocompatibility antigen [48]).

Taking advantage of the new approach based on the reconstructed ancestral genomes, we applied more stringent criteria for screening the natural selection signals: (i) candidate adaptive alleles only exist in HM populations; (ii) statistically significant difference in allele frequency (>0.3) between the aHM and the Han Chinese population; (iii) significant enrichment of adaptive alleles in the aHM genomes (see "Methods"). We identified 2779 variants (Additional file 1: Table S22), 395 of which were eQTL. Notably, the 175 genes regulated by these QTL were significantly enriched in the external side of the plasma membrane (FDR $P = 1.7 \times 10^{-13}$), interferon-gamma-mediated signaling pathway (FDR $P = 8.2 \times 10^{-12}$), ER to Golgi transport vesicle membrane (FDR $P = 1.7 \times 10^{-11}$), and MHC class II protein complex (FDR $P = 4.8 \times 10^{-10}$) (Additional file 1: Table S23). In addition, among the 2779 variants, we identified a differential variant aggregation region of 102 kb, and 151 variants in this region showed a frequency difference of 70% or larger. The two genes involved in this region, *SLCO1B3*, and *SLCO1B7*, are members of the liver-specific organic anion transporter family; they encode transmembrane receptors, play a key role in bile acid and bilirubin transport, and participate in bile salt recycling [49–51].

We also found that the rare variants with strong effects played a role in the adaptation of the Yao population to the environment. Based on the allele frequency, we determined the ancestral allele in the East Asian population and then found new mutations in the Yao population. Seven protein-coding genes (*RASSF5*, *MYH3*, *ADCY9*, *DENND48*, *TANGO6*, *SBNO2*, and *BEGAIN*) were specifically enriched with strong effects on recently derived alleles in the Yao population (Additional file 1: Table S24). However, these alleles were not detected in the Han population. We annotated 42 rare variants specifically carried by the Yao individuals on these 7 genes for conservatism (GERP) and pathogenicity (CADD). We found 32 out of the 42 mutations with GREP ≥ 2 or CADD ≥ 10, especially 27 of them with GREP ≥ 4 or CADD ≥ 15, indicating the potential functions of these rare mutations. We also found that these strong effects in rare variants showed familial aggregation, that is, 4 of the 24 pairs of related samples shared at least one strong effect in rare variants in these genes. However, this proportion was only 35/3136 in the unrelated sample pairs ($P = 1.76 \times 10^{-4}$; Additional file 2: Text S8, Additional file 1: Table S25). The enrichment of rare mutations in these genes suggests that there may be positive selection at the gene level in the Yao population. Multiple rare variants with a strong effect from one gene were dispersed in different individuals (Fig. 7), which greatly improves the carrying rate of mutated genes in the Yao population. However, the positive selection acting on these genes did not increase the frequency of each allele, and they were missing in the Han Chinese population. Therefore, we speculate that there might be negative selection forces as well acting on these sites. The functional annotation also indicated the potential pleiotropy of these genes. For example, *SBNO2* is related to bone homeostasis [52] and also participates in the pro-inflammatory cascade [53]. *MYH3* is related to muscle organ development [54] and also to the Freeman-Sheldon syndrome [55]. *RASSF5* is related to wound healing [56] and also to multiple human cancers [57–61]. The various effects of these pleiotropic genes could be driven by different evolutionary forces.

In addition, we identified HLA types in the HM population (Additional file 1: Table S26). The HLA-B * 15:02 allele has been reported to be strongly associated with severe skin allergic reactions caused by carbamazepine, and the East Asian population has a high allele frequency [62]. In our data, the frequency of the HLA-B * 15:02 allele in the Yao population (14.52%, 18/124) is higher than that in Miao (2/20), She (0/20), Dai (1/18), Han
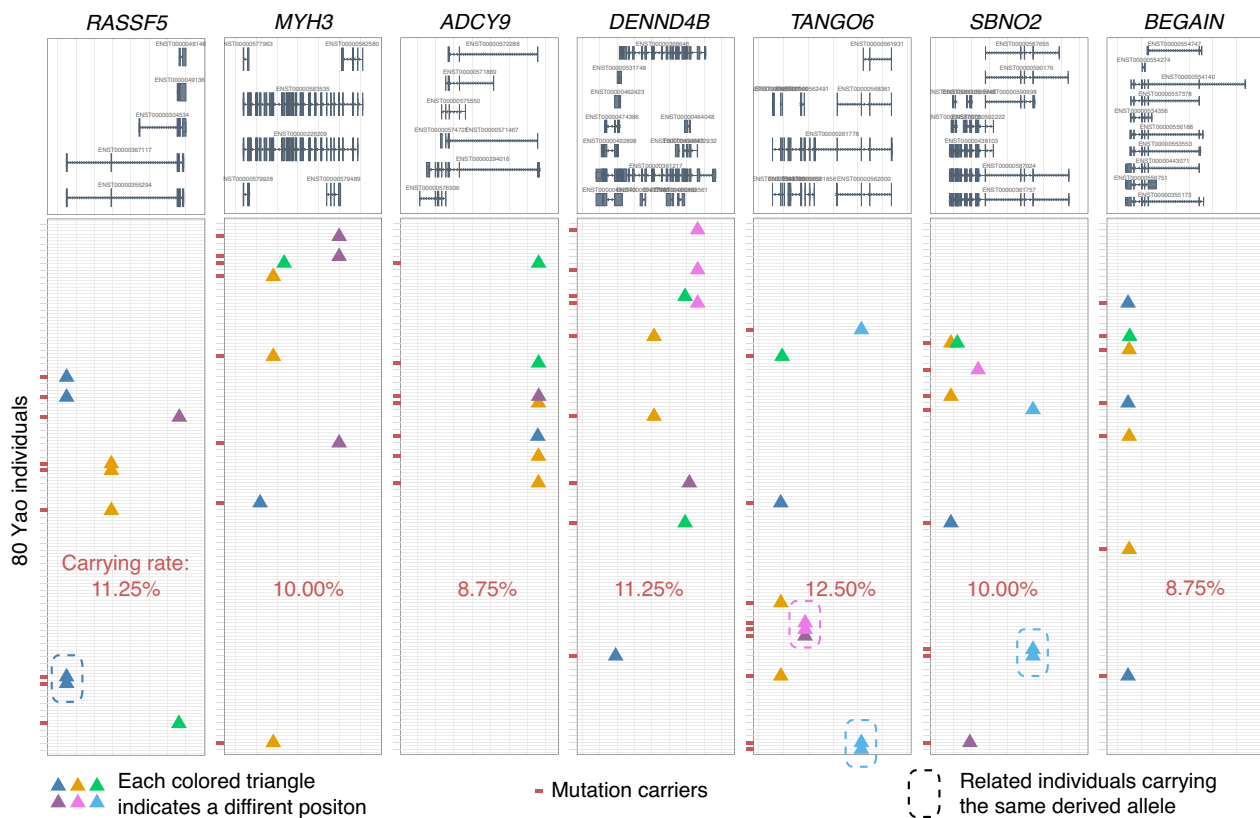
Gao *et al. BMC Biology*   (2024) 22:59

Page 12 of 22

Fig. 7 Rare derived alleles with strong effects were dispersed in different individuals. The top shows all transcripts of each gene, and each line below is a Yao individual. Triangles of different colors indicate that rare derived alleles are located in different loci. The box indicates that two related individuals share the same rare variants at this locus. The individual marked with a short red line indicates that the individual carries at least one variant with strong effects on the gene. The proportion of individuals carrying strong effect variants on each gene is the carrying rate

(2/80), and Tibetan (0/76) populations (Additional file 2: Fig. S26, Additional file 1: Table S27). This guides the use of related drugs. However, the driving force of HLA-B * 15:02 allele fluctuation in the population needs to be studied.

Discussion
In this study, we collected 80 samples of the Yao population and performed a deep whole-genome sequence. We identified 504,927 novel variants and 71.56 MB of Yao-specific ancient infiltration fragments which expanded the gene pool of modern humans. In particular, we have also found mutations such as rs72474224 ((p.Val37Ile, *GJB2*) that are associated with deafness in the HM population are in high frequency (Fig. 6), indicating the medical potential of genetic research on minority populations. Although the sample size may lead to a rough estimation of the allele frequency in the Miao and She populations, the high allele frequency observed in three independent populations is sufficient to support our conclusion.

We inferred the paternal and maternal haplogroups of the HM population, as well as estimated the time to the MRCA of the HM-specific paternal haplogroup

O-N5 (4288 years ago, Fig. 1e), which is approximately 1800 years earlier than the time of the MRCA of O-N5 previously reported (2500 years ago) [7]. This may benefit from the whole-genome deep sequencing. In addition, we found that in the three populations of Yao, Miao, and She, the main paternal and maternal haplogroups were different (Additional file 1: Table S8, S11), which may reflect the complex population history after the divergence of the HM population. In addition, the frequency of O-M7 (upstream of O-N5) in the ancient population of the Daxi site is 5/16, which is similar to the frequency in the present-day She (3/7) and Miao (2/7) populations, but the frequency of O-M7 in the Yao population is extremely low (2/44). Daxi site may be a direct ancestor of the Miao population or the O-M7 lineage may have been diluted in the Yao population. However, the lack of autosomal variation at the Daxi site makes it difficult to draw an exclusive conclusion.

Our analyses including PCA (Fig. 1a), ancestry (Fig. 1d), linguistic similarity (Fig. 1f), and Y chromosome haplogroups (Fig. 1e) showed that Yao was first separated from the HM population. However, we also noticed that the F_{ST} between Miao and She is larger than that

Gao *et al. BMC Biology*   (2024) 22:59

Page 13 of 22

between them and Yao (Fig. 2a). We speculate that F_{ST} is more sensitive to recent population isolation. For example, Ami and Atayal, two indigenous peoples in Taiwan Island, have extremely high F_{ST} with other populations (Additional file 1: Table S12). Therefore, we evaluated the recent gene flow between the HM subgroups and the surrounding populations with F_{ST}. In addition, we found that there were differences in the divergence time or the time of the MRCA obtained based on different data and methods. Y chromosome data estimates that the MRCA of the HM population is 4288 years ago (Fig. 1e), while MSMC and MSMC-IM estimate that the divergence time between Yao and Miao, as well as between Yao and She, is 5800 years ago, and the divergence time between Miao and She is 7197 years ago (Additional file 2: Fig. S6a). The genetic traces on the Y chromosome can be disrupted by recent population events, such as the bottleneck leading to the loss of ancient haplogroups, and recent gene flow leading to the insertion of new haplogroups, all of which may lead to underestimation or overestimation of the time of the MRCA. Therefore, the time estimation from the Y chromosome is only used as auxiliary evidence. The divergence pattern of MSMC is consistent with the results of F_{ST}, and we speculate that it may also be affected by recent gene flow or isolation, such as the admixture between HM and Han resulting in underestimated divergence time (Additional file 2: Fig. S6b). In the MSMC results, the divergence time between Yao and She is consistent with that between Yao and Miao, which is consistent with the model where Yao first separated from the HM population. Therefore, we speculate that this is a reliable time for the MRCA for the HM population. As for the divergence time between Miao and She, which is inferred to be 7197 years ago, we speculate that this may be related to admixture with the Han Chinese population and recent isolation.

Our further analysis of genetic and linguistic data supported the consistent relationship within the HM group, which indicated a substantial genetic basis of the HM language family. In particular, the She population was found to have closer ties with the Miao population from the perspective of both genetics and linguistics. At present, the two views on the formation time of the HM language family are 2500 years ago [63] and 4243 years ago [64]. The time of 4243 years is similar to the time of the MRCA of the HM population inferred based on the Y chromosome (4288 years ago), but later than the time of the MRCA of HM speakers inferred based on the autosome. (5800 years ago). In addition, from this perspective, the formation time of the Hmong (Miao) language branch is 2777 years ago [64], which is later than our inferred time based on the Y chromosome (3418 years ago). Therefore, we suggested that the formation time of the HM language

family was underestimated owing to the population admixture after divergence. We also observed considerable gene flow to the HM population from the Tai–Kadai and the Han populations, implying a complex population history of different language speakers.

Based on the comparison of the ancestral components of autosomes and X-chromosomes, we found that there is a sex-biased admixture in present-day HM populations, with more admixture between southern males and northern females (Additional file 2: Fig. S10). However, this is not contradictory to the common belief that the population migration southward was male-driven. The main reason is that the sex bias of population migration is different from the preference for spouse selection. Some ethnic historical records indicate that most southern ethnic minorities did not intermarry with outsiders until nearly a hundred years ago, as can be seen from the ancestral components on the autosomes. In the past, there was a tradition among the Yao ethnic group that women did not marry other ethnic groups, while a minority would marry women from other ethnic groups, which may lead to a high proportion of northern ancestral components on the X chromosome.

In palaeontological research, the ancestral genome has been reconstructed by searching the homologous blocks of the extant genome under the divergence model [65, 66]. However, this method is not suitable for reconstructing the ancestors of closely related human populations in a complex model with admixture due to recombination and replacement. We thus developed an approach to reconstruct the ancestral genomes of a certain ethnic group based on present-day population genome data. This strategy facilitates the identification of the ancestral genomic segments from the present-day populations and overcomes the limitations of ancient DNA data or the unavailability of ethnic information from ancient samples. It can show the genomic diversity of ancestral populations. Moreover, we found that the method is robust. In most genomic regions, there is no significant sequence difference between different populations, while a small number of ancestral inference errors in these regions are unlikely to affect the results significantly. However, we would make it clear that the reconstructed ancestral genomes may not represent the true individuals that have existed in real history in the current version; rather, they are the proxy genomes representing an ancient gene pool of relatively isolated populations with less influence from recent gene flow owing to the massive migration of human populations.

With this approach, we gained further insights into the genetic origins and admixture history of the HM populations. The Yangtze River Basin is one of the cradles of agricultural civilization in East Asia. There is evidence

that rice agriculture has been developing in the lower reaches of the Yangtze River for 6000–8000 years. We estimated that the differentiation of the HM populations was about 5800 years, which is also consistent with the origin and development time of agricultural civilization in the Yangtze River Basin. There is also evidence of the establishment of rice agriculture in the middle reaches of the Yangtze River during the Daxi period about 5300–6400 years ago. Ancient DNA samples found in the Daxi site indicated that the Daxi culture is related to the HM ancestral population. We also found evidence that the HM ancestors lived in the Yangtze River Basin and extended northward and southward through the study of other ancient DNA samples. The high consistency between geographical and genetic coordinates implied that the HM population has settled in the current area for a long time (Fig. 1b,c). All the evidence suggests that the ancestral HM population was one of the early groups that developed with agricultural civilization in the Yangtze River Basin. The genetic components of the HM ancestral population found in present-day East Asian populations, especially southern East Asian populations, also support the founder effect of HM ancestors.

In this study, we also revealed an admixture pattern of "mutual gene flow" among several major populations in East Asia, including the Han Chinese, HM, and Tai–Kadai populations. Our model is different from that proposed in previous studies, which typically assumes that two or more ancestral populations are admixed into a new population (A + B—> C). In our model, the two ancestral populations have a short contract or a small proportion of infiltration such that the two ancestral populations obtain a certain proportion of genetic components from each other (A + B—> A' + B'). According to our current research, this admixture pattern of mutual infiltration may have played a major role in the formation of most East Asian populations.

Despite there being more published ancient DNA data available, there is an obvious gap between ancient samples and present-day populations owing to the lack of ethnic label in the ancient samples. Frequent human dispersals and prevalent gene flow have prevented most studies from establishing reliable links between geographic information and ethnic information. In this study, we attempted to label 968 ancient DNA samples by calculating the shared genetic drift between each sample and present-day populations. This effort narrowed the gap to some degree and established a rough link between ancient samples and present-day populations, which is expected to facilitate tracing the genetic origins and admixture history of present-day populations.

We use multiple methods to detect potential natural selection signals on the genome of HM and Yao populations, and different methods complement each other based on different principles and assumptions (Additional file 2: Fig. S21, S25). Although the significance of some selected signals is currently unclear, it provides a reference for other related studies. In addition, we also found that the reconstructed aHM genome is very helpful for detecting selection signals. It can weaken the impact of recent gene flow, highlighting some masked signal variants. However, from the results, it can be seen that the selection signals based on reconstructed ancestral genome detection are more complementary than substitutive compared to using present-day population genomes. We speculate that this may reflect the natural selection in different historical periods.

We found that the cancellation of natural selection signatures could result in a power loss of commonly used methods for detecting selection. Several such cases have been identified in the analysis of Yao genome data. The cancellation could be due to multiple effects of a gene being affected by positive and negative selection forces at a certain time point, or it could be due to environmental changes in history that have caused a gene to be successively affected by positive and negative selection forces at different time points. This type of alleles might be in relatively low frequency in the population and do not show remarkable differentiation between different populations. These atypical selection signals of the low allele frequency and cancellation of natural selection can be only recognized when population data are available and carefully analyzed. Therefore, we suggest that much more attention be paid to the low-frequency variants in the analysis of diverse natural populations, especially ethnic minority groups.

Conclusions

In this study, we investigated the genetic diversity and local adaptation of the HM populations. We developed a method for reconstructing the ancestral genome based on genomes of present-day populations, and we further demonstrated that this method is robust and helpful for the study of genetic structure, population history, and local adaptation. Our results support that the three main HM populations, Miao, Yao, and She, shared the most recent common ancestor about 5800 years ago, with their origins dating back to the middle reaches of the Yangtze River. The genetic history of the HM populations is complex. The Yao population first diverged from the HM populations. The She population has experienced a longer period of population isolation after separating from the Miao population. All three HM groups have been influenced by gene flow from the surrounding populations. For example, Yao received gene flow from Zhuang, while Miao received gene flow from Tujia. Each of the three

Gao *et al. BMC Biology*   (2024) 22:59

Page 15 of 22

HM groups received a slightly different gene flow from the Han Chinese population. In the study of local adaptation, we found that a risk variant rs72474224 in the *GJB2* gene is associated with deafness and underlying positive selection in the HM populations. We speculate that this gene may have other unknown effects on the evolution of the HM populations.

Methods
Samples
In total, 80 Yao blood samples were collected from the main inhabited area of the Yao population, Guangxi Province, in southern China. Informed consent was obtained from all the individuals who participated in this study.

Data compilation

Resource	Source	Identifier
Deposited data		
Yao80	This manuscript	xushua@fudan.edu.cn
AAGC	[67]	https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA000246
HGDP	[68]	https://www.ebi.ac.uk/ena/browser/view/PRJEB6463
SGDP	[14]	https://www.ebi.ac.uk/ena/browser/view/PRJEB9586
1KGP	[69]	https://www.internationalgenome.org/
Human origin	[70]	http://www.ebi.ac.uk/ena/data/view/PRJEB6272
ADP	[71]	https://ega-archive.org/datasets/EGAD00010001491
GTEx v7		https://gtexportal.org/
Clinvar	[39]	https://www.ncbi.nlm.nih.gov/clinvar/
PGG.SNV	[40]	https://www.pggsnv.org/
Software and algorithms		
bwa	[72]	https://github.com/lh3/bwa
picard	[73]	http://broadinstitute.github.io/picard/
samtools	[74]	https://github.com/samtools/samtools
GATK	[75]	https://gatk.broadinstitute.org/hc/en-us
bcftools	[74]	https://github.com/samtools/bcftools
king	[76]	https://www.chen.kingrelatedness.com
vcftools	[77]	https://github.com/vcftools/vcftools
Archaicseeker2	[15]	https://github.com/Shuhua-Group/ArchaicSeeker2.0

Resource	Source	Identifier
FlashPCA2	[78]	https://github.com/gabraham/flashpca
ADMIXTURE	[79]	http://software.genetics.ucla.edu/admixture/
AdmixTools	[23]	https://github.com/DReichLab/AdmixTools/
AdmixTools 2	[20]	https://github.com/uqrmaie1/admixtools/
HaploGrep	[80]	https://github.com/seppinho/haplogrep-cmd
Y-LineageTracker	[81]	https://github.com/Shuhua-Group/Y-LineageTracker
BEAST	[82]	http://beast.community/
SHAPEIT2	[83]	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
ChromoPainter	[21]	https://github.com/sahwa/ChromoPainterV2
GLOBETROTTER	[19]	https://people.maths.bris.ac.uk/~madjl/finestructure/globetrotter.html
hap-ibd	[22]	https://github.com/browning-lab/hap-ibd
MSMC	[16]	https://github.com/stschiff/msmc
MSMC-IM	[17]	https://github.com/wangke16/MSMC-IM
VEP	[84]	https://github.com/Ensembl/ensembl-vep
Selscan	[85]	https://github.com/szpiech/selscan
KOBAS	[86]	http://kobas.cbi.pku.edu.cn/
PopART	[87]	http://popart.otago.ac.nz/index.shtml
REHH	[88]	https://gitlab.com/oneoverx/rehh/
Data analysis code	This manuscript	https://github.com/Shuhua-Group/Construct-ancestral-genome https://doi.org/10.5281/zenodo.10499683

Whole-genome sequencing and data processing
Each 1–3 µg of DNA from blood was sheared into segments of 200–800 bp with the Covaris system. DNA segments were then treated according to the Illumina DNA sample preparation protocol. The DNA library was sequenced on the Illumina X10 platform using 150 bp paired-end reads, and the sequencing depth of clean data exceeded $30\times$ for each sample.

Reads of each sample were merged and mapped to the human reference genome (b37) using bwa mem version 0.7.10. The Picard (version 1.117) was used to mark duplicate reads after mapping. Then, we executed local indel realignment and base quality recalibration using GATK

Gao *et al. BMC Biology*   (2024) 22:59

Page 16 of 22

(version 3.6). Variant calling was executed through the *haplotypecaller* command of GATK, using the gvcf mode. The whole-genome sequencing data of other reference populations used in this study adopted the same analysis process, including 9 Dai, 10 Miao, and 10 She from the Human Genome Diversity Project (HGDP) [13] and Simons Genomic Diversity Project (SGDP) [14] and 40 Han, 33 Tibetan, and 5 Sherpa from the Asian Admixed Genomes Consortium (AAGC) [67].

We integrated these 80 Yao samples and other reference populations mentioned above by a joint call and implemented strict quality control through VQSR and a universal mask. Finally, autosomal biallelic SNVs were retained for downstream analysis (Panel 1). For the Y chromosome and mtDNA, the output of joint variant calling was used directly without VQSR and a universal mask. To further expand the reference populations, we integrated Panel 1 with global populations from the Affymetrix Human Origins genotyping dataset [70] and the East Asian population from the Asian Diversity Project (ADP) [71] by retaining the common variants, which were used in the analysis of variant density insensitivity (Panel 2).

Genetic relationship
We calculated the genetic relationship of all Yao samples from Panel 1 using the KING (version 2.2, related model) and removed at least one sample for each related pair to ensure that no pair had a relationship of the third degree or closer after filtering.

For 30 reconstructed genomes of the HM ancestral population, we performed the relationship inference, and the results show that no pair had a relationship of the third degree or closer.

PCA
PCA was performed in three different contexts, including Eurasia, East Asia, and Southern East Asia context. First, target populations were extracted from the Panel 2 dataset using the bcftools [74] (version 1.10.2) for each context, and the variants with a missing rate ≤ 1% and minor allele frequency ≥ 5% were retained. These informative candidate variants were downsampled by vcftools [77] (version 0.1.15) according to the principle that the physical distance between any two variants is not less than 50 kb, which can eliminate the linkage disequilibrium (LD) between loci. Finally, 33,576 (Eurasia contexts), 32,443 (East Asia contexts), and 32,308 (southern East Asia) SNPs were used in PCA. The selected variants underwent PCA by FlashPCA [78] (version 2.0). In addition, in the results of the East Asian context, we performed linear fitting of PC1 and PC2 with the latitude and longitude, respectively.

When performing PCA on the reconstructed ancestral population, to balance the sample size, we randomly selected 10 reconstructed samples and integrated them with the Panel 2 dataset in the Eurasian context. Considering the missing rate of reconstructed ancestral genomes, we filtered out the variants with a missing rate greater than 1% in any population. Then, as in the previous process, the variants with a minor allele frequency < 5% in all samples were filtered out and downsampled according to a 50-kb physical interval for LD. Finally, 24,740 SNPs were used. FlashPCA was used to perform PCA.

ADMIXTURE
We used the Panel 2 dataset. The selected target population includes all East Asian people and Mbuti is the representative of African ancestry, with French as the representative of European ancestral, Mala as the representative of South Asian ancestral, Onge as the representative of Negrito, and Nganasan as the representative of Siberian ancestral. The data preprocessing was similar to PCA. The variants with a missing rate ≤ 1% and a minor allele frequency ≥ 5% were retained, but the standard of downsampling was a 10-kb interval, which helped to improve the resolution. Finally, 73,785 SNPs were used. We ran the ADMIXTURE [79] software according to the gradient of the assumed number of ancestral populations from 2 to 15. Then, we ran an additional 20 times to detect fluctuations in the admixture model. Only one-time fluctuation occurs at $K=5$, and other Admixture models are consistent (Fig. S4).

When inferring the ancestral components of the reconstructed HM ancestral population with the present-day populations, we randomly selected 10 of the reconstructed 30 ancestral genomes and integrated them with samples used in the above ADMIXTURE analysis. The only difference from the above data preprocessing is that only variants with a missing rate = 0 were retained. In total, 32,956 SNPs were used.

Pairwise FST
To quantify the genetic affinity between Yao and other populations more accurately, we used the Panel 1 dataset, 13,198,356 SNPs in total. To avoid the bias caused by the sample size, according to the sample size of the population with the smallest sample size, we randomly selected nine samples from each population each time and calculated the pairwise F_{ST} between two populations 100 times [89].

To research the recent population admixture, we used the Panel 2 dataset to infer the degree to of the two target populations were affected by other populations after the divergence of the two target populations. Here, we developed a new method. We calculated the relative difference (RD) of genomic affinity between the two target populations (A, B) and other reference populations (X). The formula is as follows:

$$relative\ difference = \frac{F_{ST}(A,X) - F_{ST}(B,X)}{F_{ST}(A,X) + F_{ST}(B,X)}$$

If the relative difference is positive, the greater it is, the more the genetic communication between population X and population B. If the relative difference is negative, the smaller it is, the more the genetic communication between population X and population A.

The variants from the Panel 2 dataset with a missing rate $\leq 1\%$ and a minor allele frequency $\geq 10\%$ were retained and downsampled according to a 10-kb physical interval. In total, 66,640 SNPs were used to calculate F_{ST}.

Outgroup F3
To infer the genetic relationship between the three present-day HM populations (Yao, Miao, and She) and other East Asian populations, we calculated the shared genetic drift between each HM population and all other populations with the outgroup F3 method. Based on the Panel 2 dataset, we filtered out the variants with a missing rate $> 1\%$ in all samples or MAF $< 5\%$ in all populations and downsampled according to the physical interval of no less than 10 kb between any two variants to eliminate LD. After data processing, 79,802 SNPs are reserved. The outgroup F3 was performed using the qp3pop command in the ADMIXTOOLS software package [23] with Mbuti as the outgroup population.

When evaluating the quality of the reconstructed HM ancestral genomes, we extracted 10 reconstructed HM ancestral samples and merged them with the Panel2 dataset. Only the variants without any missing genotype in all samples and MAF more than 10% in any single population were retained, and the variants were downsampled using a 10-kb physical interval. Finally, 36,361 SNPs are reserved. We calculated the shared genetic drift of each HM-related population including three present-day HM populations (Yao, Miao, She) and reconstructed ancestral populations (aHM) with other East Asian reference populations using the outgroup F3 method.

To identify the ancient DNA samples most closely related to aHM, we labeled these ancient samples according to the classification of present-day populations. We integrated 968 publicly available ancient samples, using the Panel 2 dataset and reconstructed aHM data

as the reference populations, and filtered out variants with a missing rate $> 20\%$ in any one reference population. Finally, 106,153 SNPs are reserved. However, due to the high missing rate, the number of observable sites of ancient DNA samples ranges from several hundred to several hundred thousand. We calculated the shared genetic drift of each ancient sample with all reference populations. We found that 11 ancient samples shared the most genetic drift with the reconstructed aHM compared to other reference populations. Among them, three samples were filtered out owing to low coverage, and finally, only eight ancient samples were retained for further analysis.

Y chromosome and mtDNA haplogroups
To construct the maternal and paternal genealogy, we classified the mtDNA and Y chromosome haplogroups of the Panel1 datasets. The mtDNA haplogroups were classified using HaploGrep2 [80] based on PhyloTree17 [90]. The Y chromosome haplogroups were classified using Y-LineageTracker [81] based on the ISOGG Y-DNA phylogenetic tree 2019–2020 (https://isogg.org).

To estimate the age of NRY haplogroups, 106 samples with sufficient coverage and depth were used to construct the NRY phylogenetic tree and calculate the age of haplogroups. We performed Bayesian evolutionary analyses using BEAST v.2.6.0 [82] with the GTR model under the strict clock to construct the phylogenetic tree and estimate the coalescent times of NRY haplogroups and their sub-lineages. The results are visualized in FigTree v1.4.4. The age of the NRY haplogroup CT-M168 (71,760 years, 95% CI = 69,777–73,799)[91] was used for the age estimation of all NRY haplogroups.

Linguistic distance
The linguistic distance matrix among the 12 HM languages was derived from Deng's work [92]. For calculating the linguistic distance, we first obtained the similarity matrix of HM languages, which is measured by the sharing proportions of lexical cognates between every two language samples. Second, we transformed the similarity matrix into the distance matrix using the following equation:

$$D = (-\log_{10}(s)) \times 100$$

where D is the distance value between two languages, and s is the similarity value between two languages.

qpGraph
To visualize the population history model of present-day HM populations and verify the reliability of our model,

Gao *et al. BMC Biology*   (2024) 22:59

Page 18 of 22

we constructed an admixture graph using qpGraph in the admixtools2 package [20]. Here, we used the Panel 2 dataset and filtered out variants with a missing rate > 0 in all samples and an MAF < 0.1 in all populations. Finally, 75,381 SNPs are reserved. According to the results of the above analyses, we designed the basic skeleton of the HM population's origin model and refined the model based on the information provided by admixtools2 such as the score, branch length, and admixture ratio.

Haplotype inference

To determine the haplotypes of each sample, we filtered out all the variants with a missing rate greater than 5% and used SHAPEIT2 [83] to phase all the samples from Panel 1. We used a genetic map obtained from HapMap II [93] and adopted 0.5 Mb for the window size, as recommended for sequence data.

Local ancestry inference

To identify the HM-specific ancestral segments in the genome of present-day HM populations, we performed local ancestral inference by ChromoPainter [21] (V2). Based on the phased data of the Panel 1 dataset, we used 9 Dai, 9 Han, 9 Tibetan, and 9 present-day HM samples as the donors to paint the other HM samples. Considering the divergent order of the HM populations and the recent population admixture, we used nine Yao samples to represent the HM population as the donor when we painted the genomes of the Miao and She populations, and we used nine She samples to represent the HM population as the donor when we painted the Yao samples. According to the default parameters, each genome was painted 10 times. These results were used to reconstruct the aHM genome.

Reconstruction of ancestral genomes

Firstly, based on local ancestral inference, we constructed a candidate haplotype library for HM ancestors. For sample size balance, there are 10 samples each for Miao, Yao, and She in the library. To fully reflect the genetic diversity of the population, sample selection has comprehensively considered the results of IBD and PCA. Candidate haplotypes in the library require support from at least half of the 10 ancestral inference results. Assembly starts from a starting site and randomly selects one from the candidate haplotypes passing through this site. Then extend back along this haplotype for no more than 6000 bp. When a breakpoint was encountered or the expansion reached 6000 bp, we repeat the above steps to randomly select the next candidate haplotype to continue extending (sampling without replacement). To avoid the diversity

deviation caused by the low coverage of ancestral fragments in local areas, we only selected genome regions covered by at least 12 candidate haplotypes (20% of the sample size in the library) to reconstruct the ancestral genome. Finally, we reconstructed 60 aHM haplotypes and assembled 30 aHM individual genomes. Assembly of each genome is through replacement sampling.

IBD

We adopted the default parameters of hap-ibd software for IBD analysis based on the phased Panel 1 dataset with the HapMap II genetic map. We compared the HM ancestral fragment from local ancestry inference and a homologous fragment from IBD analysis at the haplotype level for 30 HM samples used in the reconstruction of ancestral genomes. We designed the contingency table (Additional file 2: Text S9) and tested whether there was a significant correlation between the above two fragments through Fisher's exact test.

Divergence time

We applied multiple sequentially Markovian coalescent (MSMC, version 1.1.0) analyses [16] to infer the divergence time from the high-coverage genomes. Two samples were selected from each population when the divergence time was to be estimated for the two populations. We first used bamCaller.py for each sample to generate the mask file, which gives the regions in which the genome of that individual was covered sufficiently. In addition to the mask for each sample, we also used the additional mappability mask for GRCh37, which gives all regions in which short sequencing reads can be uniquely mapped. Rather than phasing the data from bamCaller.py directly, we used the phased Panel 1 dataset. We calculated the absolute estimation by assuming a slow mutation rate of $\sim 1.25 \times 10^{-8}$ per base per human generation for a generation time of 29 years, as the results based on the slow mutation rate agree better with the palaeoanthropological record and with the estimates from mtDNA [94–96]. Then, further analysis was conducted according to the recommended parameters of MSMC-IM (-beta 1e-8,1e-6 –printfittingdetails –plotfittingdetails –xlog). The divergence time between Han and HM was estimated by their Ne inferred by MSMC. We calculated the mean value and standard deviation of Ne for each period of three HM populations. If Han's Ne was not within the range of adding or subtracting three times the standard deviation of HM in the same period, Han and HM were considered divergent. We chose the earliest time range of their latest divergence and calculated the median of this time range as the final divergence time.

Selection

To identify variants underlying natural selection in HM populations, we used the Han Chinese population as the reference population. The loci with a missing rate of more than 10% in any of the five groups were filtered out, including 40 Han, 10 She, 10 Miao, and 62 unrelated Yao samples from two sampling sites. We calculated the F_{ST} of each locus using 40 Han samples and 30 modern HM populations (10 Yao, 10 Miao, and 10 She) used in reconstructing the ancestral genome, and the algorithm balanced the difference in sample size [89]. The calculated results were sorted from large to small according to F_{ST}, and the loci with a large F_{ST} (top 0.1%) were considered to be affected by natural selection. These loci were annotated by the Ensembl Variant Effect Predictor (VEP) [97] (GRCh37 ensemble92), the GTEx database v7 (https://gtexportal.org/), and the Clinvar database [39] (GRCh37.20210302). Functional enrichment analysis was performed by KOBAS [86] (http://kobas.cbi.pku.edu.cn/). Using the same method, we also found significantly differentiated loci between the 30 reconstructed aHM genomes and 40 Han samples. Extended haplotype homozygosity (EHH) [98] was estimated with an R package, REHH [88]. A haplotype network was constructed by PopART [99] using all of the 39 SNVs located in the *GJB2* gene from the phased Panel 1 dataset. GERP++scores [100] were obtained from the *PGG*.SNV database [40] (https://www.pggsnv.org/).

To search HM-specific selection signals, we developed an approach with more stringent criteria. First, the allele frequency in the Han Chinese population should be fixed. Second, the difference in allele frequency should be greater than or equal to 0.3 between the reconstructed aHM and the Han Chinese population. Third, the allele frequency must be distributed in a gradient in the following order: the HM ancestral population, the present HM population, and the Han Chinese population. SNVs involved in the top 0.1% F_{ST} between the reconstructed HM ancestral population and the Han Chinese population but not in the top 1% F_{ST} between the present HM population and the Han Chinese population were considered extra SNVs.

Rare variants of strong effects

We focused on MAF ≤ 20% SNVs with moderate or high biological effects among 40 Han genomes. An allele with a higher frequency than the other allele was considered a major ancestral allele in East Asian populations. An allele with a lower frequency was considered a recently derived allele in East Asian populations. We calculated the proportion of individuals carrying at least one derived SNV for each gene by population. The highest frequency of a recently derived allele in each gene in the Yao population was listed in the last column (Additional file 1: Table S24). In addition, genes carrying multiple dispersed rare derived alleles of strong effect SNVs were searched according to the following standards: (1) no derived allele in the Han Chinese population; (2) more than 10% of samples carrying at least one derived allele; (3) the frequency of the highest derived allele should be less than 5% in Yao population; (4) genes carry at least two SNVs.

Statistical analysis

All details of the statistics applied are provided in the Supplemental Information. We have implemented Fisher's exact test three times in R through the *ecx* command.

Abbreviations

aHM	HM ancestral
EHH	Extended haplotype homozygosity
eQTL	Expression quantitative trait loci
HM	Hmong–Mien
IBD	Identity-by-descent
mtDNA	Mitochondrial DNA
MRCA	The most recent common ancestor
PC	Principal component
PCA	Principal component analysis
ROH	Run of homozygosity
SNP	Single-nucleotide polymorphism
VEP	The Ensembl Variant Effect Predictor
VQSR	Variant Quality Score Recalibration

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-024-01838-9.

Additional file 1: Table S1. Quality control information of whole genome sequencing of Yao samples. **Table S2.** Total length of Archaic segments and Ancestry segments at the individual level among HM. **Table S3.** Higheast Altai Neanderthal segments in 80 Yao samples. **Table S4.** Higheast Denisova segments in 80 Yao samples. **Table S5.** Unique Altai Neanderthal segments in 80 Yao samples. **Table S6.** Unique Denisova segments in 80 Yao samples. **Table S7.** Y-haplogroups inferred by Y-LineageTracker software based on Panel 1 dataset. **Table S8.** The distribution of Y-haplogroups in sequencing data samples. **Table S9.** Hmong-Mien language distence matrix. **Table S10.** mtDNA-haplogroups inferred by Haplogrep2 software based on Panel 1 dataset. **Table S11.** The distribution of mtDNA-haplogroups in sequencing data samples. **Table S12.** Pairwise-F_{ST} between two populations based on Panel 2 dataset in the context of East Asia. **Table S13.** Inferring the origin of the South China components in the Han Chinese population through f4. **Table S14.** 8 ancient DNA samples sharing the most genetic drift with the Hmong-Mien ancestral population. **Table S15.** The F3 results of 8 ancient DNA samples sharing the most genetic drift with the Hmong-Mien ancestral population. **Table S16.** Top 0.1% signal of the pairwise F_{ST} of HM and Han. **Table S17.** KOBAS pathway analysis on the genes in which highly differentiated SNVs between HM and Han are located. **Table S18.** Top 0.1% signal of the pairwise F_{ST} of aHM and Han. **Table S19.** KOBAS pathway analysis on the genes in which newly identified signal SNVs of aHM are located. **Table S20.** IHS results of the HM population group. **Table S21.** IHS results of the aHM population group. **Table S22.** SNVs distributed in a gradient among aHM, HM and Han. **Table S23.** KOBAS pathway analysis on the genes that regulated by eQTLs from HM special SNVs. **Table S24.** Proportion of carrying recent derived alleles. **Table S25.** 3-degree or closer relationship pairs in 80 Yao samples. **Table S26.** HLA types of all samples in Panel 1 dataset. **Table S27.** Frequency statistics of each HLA type in each population.

Additional file 2. This file includes Supplemental Text and Supplemental Figures S1 to S28.

Gao *et al. BMC Biology*   (2024) 22:59

Page 20 of 22

Availability of data and materials
The script for reconstructing the ancestral genomes from this work can be found on GitHub and Zenodo, https://github.com/Shuhua-Group/Construct-ancestral-genome (https://doi.org/10.5281/zenodo.10499683). The release of the variants of 80 Yao samples by this work is permitted by The Ministry of Science and Technology of the People's Republic of China (permission no. 2022BAT1948) at the National Omics Data Encyclopedia (https://www.biosino.org/node) with accession number OEZ014103. All data generated or analyzed during this study are included in this published article, its supplementary information files, and publicly available repositories.

Declarations

Ethics approval and consent to participate
All procedures performed in studies involving human participants were approved by the Biomedical Research Ethics Committee of Shanghai Institutes for Biological Sciences (No. ER-SIBS-261408) and were in accordance with the 1964 Helsinki Declaration, its later amendments, or comparable ethical standards. Informed consent was obtained from all individual participants included in the study. The personal identifiers of all samples, if any existed, were stripped off before sequencing and analysis.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

References
1. Handel Z. Review of Ratliff (2010): Hmong-Mien language history. Diachronica. 2012;29(3):385–98.
2. Wen B, Li H, Gao S, Mao X, Gao Y, Li F, et al. Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. Mol Biol Evol. 2005;22(3):725–34.
3. Li H, Huang Y, Mustavich LF, Zhang F, Tan JZ, Wang LE, et al. Y chromosomes of prehistoric people along the Yangtze River. Hum Genet. 2007;122(3–4):383–8.
4. Cai X, Qin Z, Wen B, Xu S, Wang Y, Lu Y, et al. Human migration through bottlenecks from Southeast Asia into East Asia during last glacial maximum revealed by Y chromosomes. PLoS ONE. 2011;6(8):e24282.
5. Consortium HPAS, Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, et al. Mapping human genetic diversity in Asia. Science. 2009;326(5959):1541–5.
6. Yang M, He G, Ren Z, Wang Q, Liu Y, Zhang H, et al. Genomic insights into the unique demographic history and genetic structure of five hmong-mien-speaking miao and yao populations in Southwest China. Front Ecol Evol. 2022;10:849195.
7. Xia ZY, Yan S, Wang CC, Zheng HX, Zhang F, Liu YC, et al. Inland-coastal bifurcation of southern East Asians revealed by Hmong-Mien genomic history. 2019:730903.
8. Liu Y, Xie J, Wang M, Liu C, Zhu J, Zou X, et al. Genomic insights into the population history and biological adaptation of Southwestern Chinese Hmong-Mien people. Front Genet. 2022;12:2654.
9. Huang X, Xia ZY, Bin X, He G, Guo J, Adnan A, et al. Genomic insights into the demographic history of the Southern Chinese. Front Ecol Evol. 2022;10:853391.
10. Luo T, Wang R, Wang C-C. Inferring the population structure and admixture history of three Hmong-Mien-speaking Miao tribes from southwest China based on genome-wide SNP genotyping. Ann Hum Biol. 2021;48(5):418–29.
11. Tan H, Wang R, Wang C-C. Fine-scale genetic profile and admixture history of two hmong-mien–speaking miao tribes from Southwest China inferred from genome-wide data. Hum Biol. 2022;93(3):179–99.
12. Wang T, Wang W, Xie G, Li Z, Fan X, Yang Q, et al. Human population history at the crossroads of East and Southeast Asia since 11,000 years ago. Cell. 2021;184(14):3829-41 e21.
13. Bergstrom A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020;367:6484.
14. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The simons genome diversity project: 300 genomes from 142 diverse populations. Nature. 2016;538(7624):201–6.
15. Yuan K, Ni X, Liu C, Pan Y, Deng L, Zhang R, et al. Refining models of archaic admixture in Eurasia with ArchaicSeeker 2.0. Nat Commun. 2021;12(1):6232.
16. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. 2014;46(8):919–25.
17. Wang K, Mathieson I, O'Connell J, Schiffels S. Tracking human population structure through time from whole genome sequences. PLoS Genet. 2020;16(3):e1008552.
18. Su B, Xiao J, Underhill P, Deka R, Zhang W, Akey J, et al. Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. Am J Hum Genet. 1999;65(6):1718–24.
19. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. Science. 2014;343(6172):747–51.
20. Maier R, Flegontov P, Flegontova O, Changmai P, Reich D. On the limits of fitting complex models of population history to genetic data. bioRxiv. 2022:2022.05.08.491072.
21. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012;8(1):e1002453.
22. Zhou Y, Browning SR, Browning BL. A fast and simple method for detecting identity-by-descent segments in large-scale data. Am J Hum Genet. 2020;106(4):426–37.
23. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012;192(3):1065–93.
24. Yang MA, Fan X, Sun B, Chen C, Lang J, Ko YC, et al. Ancient DNA indicates human population shifts and admixture in northern and southern China. Science. 2020;369(6501):282–8.
25. Lipson M, Cheronet O, Mallick S, Rohland N, Oxenham M, Pietrusewsky M, et al. Ancient genomes document multiple waves of migration in Southeast Asian prehistory. Science. 2018;361(6397):92–5.
26. Wang CC, Yeh HY, Popov AN, Zhang HQ, Matsumura H, Sirak K, et al. Genomic insights into the formation of human populations in East Asia. Nature. 2021;591(7850):413–9.
27. Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, et al. The formation of human populations in South and Central Asia. Science. 2019;365:6457.

28. Ning C, Li T, Wang K, Zhang F, Li T, Wu X, et al. Ancient genomes from northern China suggest links between subsistence changes and human migration. Nat Commun. 2020;11(1):2700.

29. Sikora M, Pitulko VV, Sousa VC, Allentoft ME, Vinner L, Rasmussen S, et al. The population history of northeastern Siberia since the Pleistocene. Nature. 2019;570(7760):182-+.

30. Sikora M, Seguin-Orlando A, Sousa VC, Albrechtsen A, Korneliussen T, Ko A, et al. Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. Science. 2017;358(6363):659–62.

31. Ning C, Wang CC, Gao S, Yang Y, Zhang X, Wu X, et al. Ancient genomes reveal yamnaya-related ancestry and a potential source of Indo-European Speakers in Iron Age Tianshan. Curr Biol. 2019;29(15):2526-32 e4.

32. Yang MA, Gao X, Theunert C, Tong H, Aximu-Petri A, Nickel B, et al. 40,000-year-old individual from asia provides insight into early population structure in Eurasia. Curr Biol. 2017;27(20):3202-8 e9.

33. Mao X, Zhang H, Qiao S, Liu Y, Chang F, Xie P, et al. The deep population history of northern East Asia from the late pleistocene to the holocene. Cell. 2021;184(12):3256-66 e13.

34. McColl H, Racimo F, Vinner L, Demeter F, Gakuhari T, Moreno-Mayar JV, et al. The prehistoric peopling of Southeast Asia. Science. 2018;361(6397):88–92.

35. Jeong C, Ozga AT, Witonsky DB, Malmstrom H, Edlund H, Hofman CA, et al. Long-term genetic stability and a high-altitude East Asian origin for the peoples of the high valleys of the Himalayan arc. Proc Natl Acad Sci U S A. 2016;113(27):7485–90.

36. Muhlemann B, Jones TC, Damgaard PB, Allentoft ME, Shevnina I, Logvin A, et al. Ancient hepatitis B viruses from the Bronze age to the medieval period. Nature. 2018;557(7705):418–23.

37. Kanzawa-Kiriyama H, Kryukov K, Jinam TA, Hosomichi K, Saso A, Suwa G, et al. A partial nuclear genome of the Jomons who lived 3000 years ago in Fukushima. Japan J Hum Genet. 2017;62(2):213–21.

38. Allen Ancient DNA Resource https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data", version 44.3.

39. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44(D1):D862–8.

40. Zhang C, Gao Y, Ning Z, Lu Y, Zhang X, Liu J, et al. PGG.SNV: understanding the evolutionary and medical implications of human single nucleotide variations in diverse populations. Genome Biol. 2019;20(1):215.

41. Fushan AA, Simons CT, Slack JP, Drayna D. Association between common variation in genes encoding sweet taste signaling components and human sucrose perception. Chem Senses. 2010;35(7):579–92.

42. Wang CY, Liu S, Xie XN, Tan ZR. Regulation profile of the intestinal peptide transporter 1 (PepT1). Drug Des Devel Ther. 2017;11:3511–7.

43. Karolak JA, Rydzanicz M, Ginter-Matuszewska B, Pitarque JA, Molinari A, Bejjani BA, et al. Variant c.2262A>C in DOCK9 leads to exon skipping in keratoconus family. Invest Ophthalmol Vis Sci. 2015;56(13):7687–90.

44. Kang S, Maeng H, Kim BG, Qing GM, Choi YP, Kim HY, et al. In situ identification and localization of IGHA2 in the breast tumor microenvironment by mass spectrometry. J Proteome Res. 2012;11(9):4567–74.

45. Jeraiby MA. Molecular basis of immunoglobulin heavy constant G4 gene (IGHG4)-related low serum IgG4 subclasses in Down syndrome. Saudi Med J. 2021;42(9):975–80.

46. Liu B, Maier LA, Hamzeh N, MacPhail K, Mroz MM, Liu H, et al. Polymorphism of FCGR3A gene in chronic beryllium disease. Genes Immun. 2019;20(6):493–9.

47. Alberici F, Bonatti F, Adorni A, Daminelli G, Sinico RA, Gregorini G, et al. FCGR3B polymorphism predicts relapse risk in eosinophilic granulomatosis with polyangiitis. Rheumatology (Oxford). 2020;59(11):3563–6.

48. Giotis ES, Carnell G, Young EF, Ghanny S, Soteropoulos P, Wang LF, et al. Entry of the bat influenza H17N10 virus into mammalian cells is enabled by the MHC class II HLA-DR receptor. Nat Microbiol. 2019;4(12):2035–8.

49. Liu S, Peng T, Wang Z, Li Y, Zhang H, Gui C. Effect of rare coding variants of charged amino acid residues on the function of human organic anion transporting polypeptide 1B3 (SLCO1B3). Biochem Biophys Res Commun. 2021;557:1–7.

50. Meyer Zu Schwabedissen HE, Seibert I, Grube M, Alter CL, Siegmund W, Hussner J. Genetic variants of are of relevance for the transport function of. Pharmacol Res. 2020;161:105155.

51. Chun SE, Thakkar N, Oh Y, Park JE, Han S, Ryoo G, et al. The N-terminal region of organic anion transporting polypeptide 1B3 (OATP1B3) plays an essential role in regulating its plasma membrane trafficking. Biochem Pharmacol. 2017;131:98–105.

52. Kim SK. Identification of 613 new loci associated with heel bone mineral density and a polygenic risk score for bone mineral density, osteoporosis and fracture. PLoS ONE. 2018;13(7):e0200785.

53. El Kasmi KC, Smith AM, Williams L, Neale G, Panopoulos AD, Watowich SS, et al. Cutting edge: A transcriptional repressor and corepressor induced by the STAT3-regulated anti-inflammatory signaling pathway. J Immunol. 2007;179(11):7215–9.

54. Karsch-Mizrachi I, Travis M, Blau H, Leinwand LA. Expression and DNA sequence analysis of a human embryonic skeletal muscle myosin heavy chain gene. Nucleic Acids Res. 1989;17(15):6167–79.

55. Bowman S, Noble G, Rahmani B, Mets M, Ralay Ranaivo H, Castelluccio V. A case of blepharophimosis: freeman sheldon syndrome. Ophthalmic Genet. 2021;43:1–4.

56. Fujita H, Fukuhara S, Sakurai A, Yamagishi A, Kamioka Y, Nakaoka Y, et al. Local activation of Rap1 contributes to directional vascular endothelial cell migration accompanied by extension of microtubules on which RAPL, a Rap1-associating molecule, localizes. J Biol Chem. 2005;280(6):5022–31.

57. Li S, Teng J, Li H, Chen F, Zheng J. The emerging roles of RASSF5 in human malignancy. Anticancer Agents Med Chem. 2018;18(3):314–22.

58. Guo W, Wang C, Guo Y, Shen S, Guo X, Kuang G, et al. RASSF5A, a candidate tumor suppressor, is epigenetically inactivated in esophageal squamous cell carcinoma. Clin Exp Metastasis. 2015;32(1):83–98.

59. Zhou XH, Yang CQ, Zhang CL, Gao Y, Yuan HB, Wang C. RASSF5 inhibits growth and invasion and induces apoptosis in osteosarcoma cells through activation of MST1/LATS1 signaling. Oncol Rep. 2014;32(4):1505–12.

60. Djos A, Martinsson T, Kogner P, Caren H. The RASSF gene family members RASSF5, RASSF6 and RASSF7 show frequent DNA methylation in neuroblastoma. Mol Cancer. 2012;11:40.

61. Lee CK, Lee JH, Lee MG, Jeong SI, Ha TK, Kang MJ, et al. Epigenetic inactivation of the NORE1 gene correlates with malignant progression of colorectal tumors. BMC Cancer. 2010;10:577.

62. Lo C, Nguyen S, Yang C, Witt L, Wen A, Liao TV, et al. Pharmacogenomics in Asian subpopulations and impacts on commonly prescribed medications. Clin Transl Sci. 2020;13(5):861–70.

63. Roger Blench LS, Alicia Sanchez-Mazas. The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics. Routledge. 2005.

64. Holman EW, Brown CH, Wichmann S, Müller A, Velupillai V, Hammarström H, et al. Automated dating of the world's language families based on lexical similarity. Curr Anthropol. 2011;52(6):841–75.

65. Avdeyev P, Jiang S, Aganezov S, Hu F, Alekseyev MA. Reconstruction of ancestral genomes in presence of gene gain and loss. J Comput Biol. 2016;23(3):150–64.

66. O'Connor RE, Romanov MN, Kiazim LG, Barrett PM, Farre M, Damas J, et al. Reconstruction of the diapsid ancestral genome permits chromosome evolution tracing in avian and non-avian dinosaurs. Nat Commun. 2018;9(1):1883.

67. Lu D, Lou H, Yuan K, Wang X, Wang Y, Zhang C, et al. Ancestral origins and genetic history of tibetan highlanders. Am J Hum Genet. 2016;99(3):580–94.

68. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. bioRxiv. 2019:674986.

69. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.

70. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014;513(7518):409–13.

71. Liu X, Lu D, Saw WY, Shaw PJ, Wangkumhang P, Ngamphiw C, et al. Characterising private and shared signatures of positive selection in 37 Asian populations. Eur J Hum Genet. 2017;25(4):499–508.

72. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997v2. 2013.

73. Broad. Picard Toolkit. 2019. https://broadinstitute.github.io/picard/.

74. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):008.
75. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8.
76. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867–73.
77. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
78. Abraham G, Qiu Y, Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics. 2017;33(17):2776–8.
79. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–64.
80. Weissensteiner H, Pacher D, Kloss-Brandstatter A, Forer L, Specht G, Bandelt HJ, et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. 2016;44(W1):W58-63.
81. Chen H, Lu Y, Lu D, Xu S. Y-LineageTracker: a high-throughput analysis framework for Y-chromosomal next-generation sequencing data. BMC Bioinformatics. 2021;22(1):114.
82. Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2014;10(4):e1003537.
83. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods. 2011;9(2):179–81.
84. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The ensembl variant effect predictor. Genome Biol. 2016;17(1):122.
85. Szpiech ZA, Hernandez RD. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol Biol Evol. 2014;31(10):2824–7.
86. Bu D, Luo H, Huo P, Wang Z, Zhang S, He Z, et al. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. Nucleic Acids Res. 2021;49(W1):W317–25.
87. Leigh JW, Bryant D. popart: full-feature software for haplotype network construction. Methods Ecol Evol. 2015;6(9):1110–6.
88. Gautier M, Vitalis R. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics. 2012;28(8):1176–7.
89. Weir BS, Hill WG. Estimating F-statistics. Annu Rev Genet. 2002;36:721–50.
90. van Oven M. PhyloTree build 17: growing the human mitochondrial DNA tree. Forensic Sci Int. 2015;5:e392–4.
91. Karmin M, Saag L, Vicente M, Wilson Sayres MA, Jarve M, Talas UG, et al. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res. 2015;25(4):459–66.
92. Xiaohua D. The Relationship and Classification of Sino-Tibetan Languages 2006.
93. International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449(7164):851–61.
94. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science. 2010;328(5978):636–9.
95. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012;488(7412):471–5.
96. Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. 2012;13(10):745–53.
97. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic Acids Res. 2021;49(D1):D884–91.
98. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419(6909):832–7.
99. Leigh JW, Bryant D. popart: full-feature software for haplotype network construction. Methods Ecol Evol. 2015;6(9):1110–6.
100. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6(12):e1001025.

Publisher's Note