frontiers in Genetics

ORIGINAL RESEARCH
published: 03 January 2022
doi: 10.3389/fgene.2021.815160

Genomic Insights Into the Population History and Biological Adaptation of Southwestern Chinese Hmong–Mien People

Yan Liu[1,2†], Jie Xie[1†], Mengge Wang[3,4†], Changhui Liu[3], Jingrong Zhu[5], Xing Zou[6], Wenshan Li[7], Lin Wang[8], Cuo Leng[7], Quyi Xu[3], Hui-Yuan Yeh[9*], Chuan-Chao Wang[10,11,12*], Xiaohong Wen[1*], Chao Liu[3,4*] and Guanglin He[9,10,11,12*†]*

[1]School of Basic Medical Sciences, North Sichuan Medical College, Nanchong, China, [2]Medical Imaging Key Laboratory of Sichuan Province, North Sichuan Medical College, Nanchong, China, [3]Guangzhou Forensic Science Institute, Guangzhou, China, [4]Faculty of Forensic Medicine, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China, [5]Department of Anthropology and Ethnology, Xiamen University, Xiamen, China, [6]College of Medicine, Chongqing University, Chongqing, China, [7]College of Medical Imaging, North Sichuan Medical College, Nanchong, China, [8]College of Clinical Medicine, North Sichuan Medical College, Nanchong, China, [9]School of Humanities, Nanyang Technological University, Singapore, Singapore, [10]State Key Laboratory of Cellular Stress Biology, National Institute for Data Science in Health and Medicine, School of Life Sciences, Xiamen University, Xiamen, China, [11]Department of Anthropology and Ethnology, Institute of Anthropology, School of Sociology and Anthropology, Xiamen University, Xiamen, China, [12]State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen, China

Hmong–Mien (HM) -speaking populations, widely distributed in South China, the north of Thailand, Laos, and Vietnam, have experienced different settlement environments, dietary habits, and pathogenic exposure. However, their specific biological adaptation remained largely uncharacterized, which is important in the population evolutionary genetics and Trans-Omics for regional Precision Medicine. Besides, the origin and genetic diversity of HM people and their phylogenetic relationship with surrounding modern and ancient populations are also unknown. Here, we reported genome-wide SNPs in 52 representative Miao people and combined them with 144 HM people from 13 geographically representative populations to characterize the full genetic admixture and adaptive landscape of HM speakers. We found that obvious genetic substructures existed in geographically different HM populations; one localized in the HM clines, and others possessed affinity with Han Chinese. We also identified one new ancestral lineage specifically existed in HM people, which spatially distributed from Sichuan and Guizhou in the north to Thailand in the south. The sharing patterns of the newly identified homogenous ancestry component combined the estimated admixture times via the decay of linkage disequilibrium and haplotype sharing in GLOBETROTTER suggested that the modern HM-speaking populations originated from Southwest China and migrated southward in the historic period, which is consistent with the reconstructed phenomena of linguistic and archeological documents. Additionally, we identified specific adaptive signatures associated with several important human nervous system biological functions. Our pilot work emphasized the importance of anthropologically informed sampling and deeply genetic structure reconstruction via whole-genome sequencing in

OPEN ACCESS

Edited by:
Jianye Ge,
University of North Texas Health Science Center, United States

Reviewed by:
Liming Li,
Princeton University, United States
Peng Chen,
Nanjing Medical University, China

***Correspondence:**
Mengge Wang
menggewang2021@163.com
Hui-Yuan Yeh
hyyeh@ntu.edu.sg
Chuan-Chao Wang
wang@xmu.edu.cn
Xiaohong Wen
xhongwen@sina.com
Chao Liu
liuchaogzf@163.com
Guanglin He
Guanglinhescu@163.com

†These authors have contributed equally to this work and share first authorship

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

Received: *15 November 2021*
Accepted: *03 December 2021*
Published: *03 January 2022*

Citation:
Liu Y, Xie J, Wang M, Liu C, Zhu J, Zou X, Li W, Wang L, Leng C, Xu Q, Yeh H-Y, Wang C-C, Wen X, Liu C and He G (2022) Genomic Insights Into the Population History and Biological Adaptation of Southwestern Chinese Hmong–Mien People. Front. Genet. 12:815160. doi: 10.3389/fgene.2021.815160

the next step in the deep Chinese Population Genomic Diversity Project (CPGDP), especially in the regions with rich ethnolinguistic diversity.

Keywords: Chinese Population Genetic Diversity Project (CPGDP), biological adaptation, genome-wide SNPs, genetic admixture model, HM people

1 INTRODUCTION

The Yungui Plateau and surrounding regions are the most ethnolinguistically diverse regions of China with a population size of approximately 0.205 billion (2020 census), which is the home to many ethnic groups, including the major population of Han Chinese and minorities of Hmong–Mien (HM), Tai–Kadai (TK), and Tibeto-Burman (TB). This region is a mountainous and rugged area, consisting of Sichuan, Chongqing, Guizhou, Yunnan and most parts of Tibet Autonomous Region, which is characterized by the Sichuan Basin in the northeast, the karstic Yunnan–Guizhou Plateau in the east, and the Hengduan Mountains in the west, and the majority of the region is drained by the Yangtze River. Historical records documented that portions of Southwest China were incorporated as unequivocal parts of greater China since at least the end of the third century BCE (Herman, 2018), and this region was largely dominated and incorporated into the Chinese domain by the time of the Ming dynasty (Harper, 2007). It has been suggested that the Nanman tribes were ancient indigenous people who inhabited in inland South and Southwest China (Yu and Li, 2021). The Nanman referred to various ethnic groups and were probably the ancestors of some present-day HM, TK, and non-Sinitic Sino-Tibetan (ST) groups living in Southwest China. Generally, Southwest China exhibits a unique panorama of geographic, cultural, ethnic, linguistic, and genetic diversity. However, the complete picture of genetic diversity of ethnolinguistically diverse populations in this region remained uncharacterized.

During the past decade, paleogenomic studies have transformed our knowledge of the population history of East Asians (Fu et al., 2013; Ning et al., 2019; Ning et al., 2020; Yang et al., 2020; Liu et al., 2021a; Wang et al., 2021a; Wang et al., 2021e; Mao et al., 2021). A recent archaeological study of the early Holocene human cranium from Guizhou (Zhaoguo M1) supported that regionalization of morphological variability patterns between Neolithic northern and southern East Asians could trace back to at least 10,000 years ago (ya) (Zhang et al., 2021). However, our knowledge about the demographic history of populations in Southwest China is limited due to the lack of ancient DNA data and sparse sampling of modern people in genome-wide SNP or whole-genome studies (Wang et al., 2020; Chen et al., 2021b; Bin et al., 2021; Liu et al., 2021c; Wang et al., 2021c). A series of recent genome-wide SNP studies demonstrated that southwestern Han Chinese showed a closer affinity with northern East Asian sources relative to indigenous populations and were well fitted via the admixture of ancient millet farmers from the Yellow River basin (YRB) and rice farmers from the Yangtze River basin (Wang et al., 2020; Wang et al., 2021b; Liu et al., 2021c; Wang et al., 2021c). Genetic findings focused on the culturally unique Hui people in this region also have proved that cultural diffusion has played an important role in the formation of the Hui people, and southwestern Huis could be modeled as a mixture of major East Asian ancestry and minor western Eurasian ancestry (Wang et al., 2020; Liu et al., 2021c). He et al. further obtained genomic information from 131 TB-speaking Tujia individuals from Southwest/South Central China and found the strong genetic assimilation between Tujia people and central Han Chinese, which provided evidence that massive population movements and genetic admixture under language borrowing have facilitated the formation of the genetic structure of Tujia people (He et al., 2021a). The patterns of the population structure of TK groups revealed the genetic differentiation among TK people from Southwest China and showed that YRB millet farmers and Yangtze River rice farmers contributed substantially to the gene pool of present-day inland TK people (Bin et al., 2021; Wang et al., 2021b). Chen et al. recently analyzed genome-wide SNP data of 26 Mongolic-speaking Mongolians and 55 Tungusic-speaking Manchus from Guizhou and found that southwestern Mongolic/Tungusic groups had a stronger genetic affinity with southern East Asians than with northern Altaic groups (Chen et al., 2021b). It is remarkable, however, no specific genome-wide studies have been published to shed new light on the population structure of HM groups from Southwest China.

Currently, HM groups mainly dwell in South China (including South Central, Southwest, and Southeast China) (He et al., 2019; Xia et al., 2019; Zhang et al., 2019; Huang et al., 2020) and Vietnam and Laos and Thailand in mainland Southeast Asia (Liu et al., 2020; Kutanan et al., 2021). The history of the HM language family is obscure, which has been passed down mainly through oral legends and myths, for which few written historical records exist. Hence, linguistic, genetic, and paleogenomic studies are crucial for reconstructing the demographic history of HM groups (Xia et al., 2019; Huang et al., 2020; Liu et al., 2020; Kutanan et al., 2021; Wang et al., 2021e). Wang et al. successfully obtained genomic material from 31 ancient individuals from southern China (Guangxi and Fujian) ranging from ~12,000 to 10,000 to 500 ya and identified HM-related ancestry represented by the ~500-year-old GaoHuaHua population (Wang et al., 2021e). Recent findings based on the Neolithic genomes from Southeast Asia have found that at least five waves of southward migrations from China have participated in the formation of modern patterns of genetic and ethnolinguistic diversity of Southeast Asians (Lipson et al., 2018; Mccoll et al., 2018; Larena et al., 2021), which were respectively associated with the dispersal of Neolithic Austroasiatic (AA) dispersal, Bronze Age and Iron Age coastal Austronesian (AN) and inland TK dissemination, and historic HM and Sino-Tibetan spread. Recent studies focused on the genetic information of HM groups from South Central China demonstrated that HM-related ancestry was

phylogenetically closer to the ancestry of Neolithic mainland Southeast Asians and modern AA groups than to AN (Xia et al., 2019). Huang et al. analyzed genome-wide SNP data of HM groups from Guangxi (Southeast China) and found that HM-related ancestry maximized in the western Hmong groups (Miao_Longlin and Miao_Xilin) (Huang et al., 2020). Findings of the human genetic history of mainland Southeast Asia also confirmed that the observed heterogeneity in HM people was derived from multiple ancestral sources during the extensive population movements and interactions (Liu et al., 2020; Kutanan et al., 2021). Therefore, systematic genome-wide studies focusing on the genetic history of the southwestern Chinese HM groups and their genetic relationship with the publicly available ancient East Asians will provide additional insights into the genetic makeup of HM groups from South China.

The Miao people are the largest of the HM-speaking populations and the fourth largest of the 55 ethnic minorities in China. The Miao are a group of linguistically related people mainly living in mountainous areas of South China. Xuyong is a county in the southeastern of Sichuan province, which borders Guizhou to the south and Yunnan to the west. Here, we generated new genome-wide data of the 52 northernmost HM-speaking Miao individuals from Xuyong, Sichuan, and co-analyzed newly generated data with publicly available genome-wide data of present-day and ancient East Eurasians leveraging shared alleles and haplotypes. We first aimed to 1) study the structure of genetic variations of Sichuan Miao people and explore the genetic relationship between Sichuan Miao and other geographically different HM-speaking people, such as Miao, She, Gejia, Dongjia, Hmong, Dao, and Xijia from China and Southeast Asia. 2) We then explored the genetic relationship between Miao people and other ethnolinguistically different East Asians and published spatiotemporally different East Asians based on the sharing alleles in the descriptive and qualitative analyses. 3) Based on the sharing alleles and haplotypes, we additionally reconstructed the demographic history of Miao people in the context of the modern geographically close ancestral source candidates and genetically related ancient surrogate populations. 4) Based on the cross-population signatures of natural selection and enrichment analysis, we finally explored the genetic adaptative history of the Chinese Miao people.

2 METHODS AND MATERIALS

2.1 Sample Collection, Genotyping, and Data Merging

All 52 newly genotyped individuals were collected from three geographically different populations in Sichuan (Baila, Hele, and Jiancao). The Oragene DN salivary collection tube was used to collect salivary samples. This study was approved via the Ethical Board of North Sichuan Medical College and followed the rules of the Helsinki Declaration. Informed consent was obtained from each participating volunteer. To keep a high representative of our included samples, the included subjects should be indigenous people and lived in the sample collection place for at least three generations. We genotyped 717,227 SNPs using the Infinium Global Screening Array (GSA) version 2 in the Miao people following the default protocols, which included 661,133 autosomal SNPs and the remaining 56,096 SNPs localized in X-/Y-chromosome and mitochondrial DNA. We used PLINK (version v1.90) (Chang et al., 2015) to filter-out raw SNP data based on the missing rate (mind: 0.01 and geno: 0.01), allele frequency (--maf 0.01), and p values of the Hardy–Weinberg exact test (--hwe 10^{-6}). We used the King software to estimate the degrees of kinship among 52 individuals and remove the close relatives within the three generations (Tinker and Mather, 1993). We finally merged our data with publicly available modern and ancient reference data from Allen Ancient DNA Resource (AADR: https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data) using the mergeit software. Besides, we also merged our new dataset with modern population data from China and Southeast Asia and ancient population data from Guangxi, Fujian, and other regions of East Asia (Yang et al., 2020; Mao et al., 2021; Wang et al., 2021a; Wang et al., 2021e) and finally formed the merged 1240K dataset and the merged HO dataset (**Supplementary Table S1**). In the merged higher-density Illumina dataset used for haplotype-based analysis, we merged genome-wide data of the Miao with our recent publication data from Han, Mongolian, Manchu, Gejia, Dongjia, Xijia, and others (Chen et al., 2021a; He et al., 2021b; Liu et al., 2021b; Yao et al., 2021).

2.2 Frequency-Based Population Genetic Analysis

2.2.1 Principal Component Analysis

We performed principal component analysis (PCA) in three population sets focused on a different scale of genetic diversity. Smartpca package in EIGENSOFT software (Patterson et al., 2006) was used to conduct PCA with an ancient sample projected and no outlier removal (numoutlieriter: 0 and lsqproject: YES). East-Asian-scale PCA included 393 TK people from 6 Chinese populations and 21 Southeast populations, 144 HM individuals from 7 Chinese populations and 6 Southeast populations, 968 Sinitic people from 16 Chinese populations, 356 TB speakers from 18 northern and 17 southern populations, 248 AA people from 20 populations, 115 AN people from 13 populations, 304 Trans-Eurasian people from 27 populations from North China and Siberia, and 231 ancient individuals from 62 groups. Chinese-scale PCA was conducted based on the genetic variations of Sinitic, northern TB and TK people in China, ancient populations from Guangxi, and all 16 HM-speaking populations. A total of twenty-three ancient samples from 9 Guangxi groups were projected (Wang et al., 2021e). The third HM-scale PCA included 15 modern populations (Vietnam Hmong populations shown as outliers) and two Guangxi ancient populations.

2.2.2 ADMIXTURE

We performed model-based admixture analysis using the maximum likelihood clustering in ADMIXTURE (version

1.3.0) software (Alexander et al., 2009) to estimate the individual ancestry composition. Included populations in the East-Asian-scale PCA analysis and Chinese-scale PCA analysis were used in the two different admixture analyses with the respective predefined ancestral sources ranging from 2 to 16 and 2 to 10. We used PLINK (version v1.90) to prune the raw SNP data into unlinked data via pruning for high-linkage disequilibrium (--indep-pairwise 200 25 0.4). We estimated the cross-validation error using the results of 100 times ADMIXTURE runs with different seeds, and the best-fitted admixture model was regarded being possessed the lowest error.

2.2.3 Phylogeny Modeling With TreeMix

We used PLINK (version v1.90) to calculate the pairwise Fst genetic distance between studied Sichuan Miao (SCM) and other modern and ancient references and also estimated the allele frequency distribution of included populations in the TreeMix analyses. Both modern and ancient populations were used to construct the maximum-likelihood–based phylogenetic relationship with population splits and migration events using TreeMix v.1.13 (Pickrell and Pritchard, 2012).

2.2.4 Outgroup-f_3-Statistics and Admixture-f_3-Statistics

We assessed the potentially existed admixture signatures in SCM via the admixture-f_3-statistics in the form of f_3 (source1, source2; Miao_Baila/Jiancao/Hele), which was calculated using qp3Pop (version 435) package in the ADMIXTOOLS software (Patterson et al., 2012). The target populations with the observed negative f_3 values and Z-scores less than -3 were regarded as mixed populations with two surrogates of ancestral populations related to source1 and source2. Following this, similar to the quantitation of the genetic similarities and differences as pairwise Fst, we assessed the genetic affinity between studied populations and other reference populations via the outgroup-f_3-statistics in the form of f_3(Reference source, studied Miao; Mbuti).

2.2.5 Pairwise qpWave Tests

We calculated p-values of the rank tests of all possible population pairs among HM-speaking populations and other geographically close modern and ancient reference populations using qpWave in the ADMIXTOOLS package (Patterson et al., 2012) to test their genetic evolutionary relationships and genetic homogeneity. Here, we used a set of distant outgroup sets, which included Mbuti, Ust_Ishim, Kostenki14, Papuan, Australian, Mixe, MA1, Jehai, and Tianyuan. The obtained pairwise matrix of the p values was visualized and presented in a heatmap using the pheatmap package.

2.2.6 Admixture Modeling Using qpAdm

We further assessed the relative ancestral source and corresponding admixture proportion of Chinese HM-speaking and surrounding Han Chinese populations using a two-way–based admixture model in the qpAdm (version 634) in the ADMIXTOOLS package (Patterson et al., 2012). One of the studied populations combined with two predefined ancestral modern and ancient sources was used as the left populations, and the aforementioned pairwise-based outgroups were used as the right populations along with two additional parameters (allsnps: YES; details: YES).

2.2.7 Demographic Modeling With qpGraph

We used the R package of ADMIXTOOLS 2 (Patterson et al., 2012) to explore the best-fitted phylogenetic topology with admixture events and mixing proportions with the Mbuti, Onge, Loschbour, Tianyuan, Baojianshan, Qihe, GaoHuaHua, and Longshan as the basic representative genetic lineages for molding the formation of modern SCM. A "rotating" scheme of adding other modern and ancient populations was used to explore other genetic ancestries that would improve the qpGraph-based admixture models. One model with the predefined admixture events ranging from 0 to 5 was run 50 times, and we then chose the best models based on the Z-scores and best-fitted scores. We also replaced the Longshan people with the upper Yellow River Lajia people as the northern ancestral lineage and ran all aforementioned admixture models.

2.2.8 Linkage Disequilibrium Estimation

We estimated the decay of linkage disequilibrium in SCM using all possible population pairs of modern East Asians as surrogate populations in ALDER 1.0 (Loh et al., 2013). Two additional parameters were used here: jackknife: YES and mindis: 0.005.

2.3 Haplotype-Based Population Genetic Analysis

2.3.1 Segmented Haplotype Estimation

We used SHAPEIT software (Segmented HAPlotype Estimation & Imputation Tool) to phase our dense SNP data with the default parameters (--burn 10 --prune 10 --main 30) (Delaneau et al., 2012). Pairwise sharing IBD segments were calculated using Refined-IBD software (16May19. ad5. jar) with the length parameter as 0.1 (Browning and Browning, 2013).

2.3.2 Chromosome Painting

We ran ChromoPainterv2 software (Lawson et al., 2012) to paint the target SCM and sampled surrogate northern and southern East Asians using all-phased populations as the surrogate populations, which was regarded as the full analysis. We also removed the SCM and their most close genetic relatives (Gejia, Dongjia, and Xijia) in the set of surrogates and painted all target and surrogate populations once again, which was regarded as the regional analysis. We then combined all chunk length output files of 22 chromosomes as the final dataset of sharing chunk length.

2.3.3 FineSTRUCTURE Analysis

We identified the fine-scale population substructure using fineSTRUCTURE (version 4.0) (Lawson et al., 2012). Perl scripts of convertrecfile.pl and impute2chromopainter.pl were used to prepare the input phase data and recombination data. fineSTRUCTURE, ChromoCombine, and ChromoPainter were combined in the four successive steps of analyses with the parameters (-s3iters 100000 -s4iters 50,000 -s1minsnps 1000 -s1indfrac 0.1). The estimated coancestry was used to run PCA analysis and phylogenetic relationships at the individual-level and population-level.

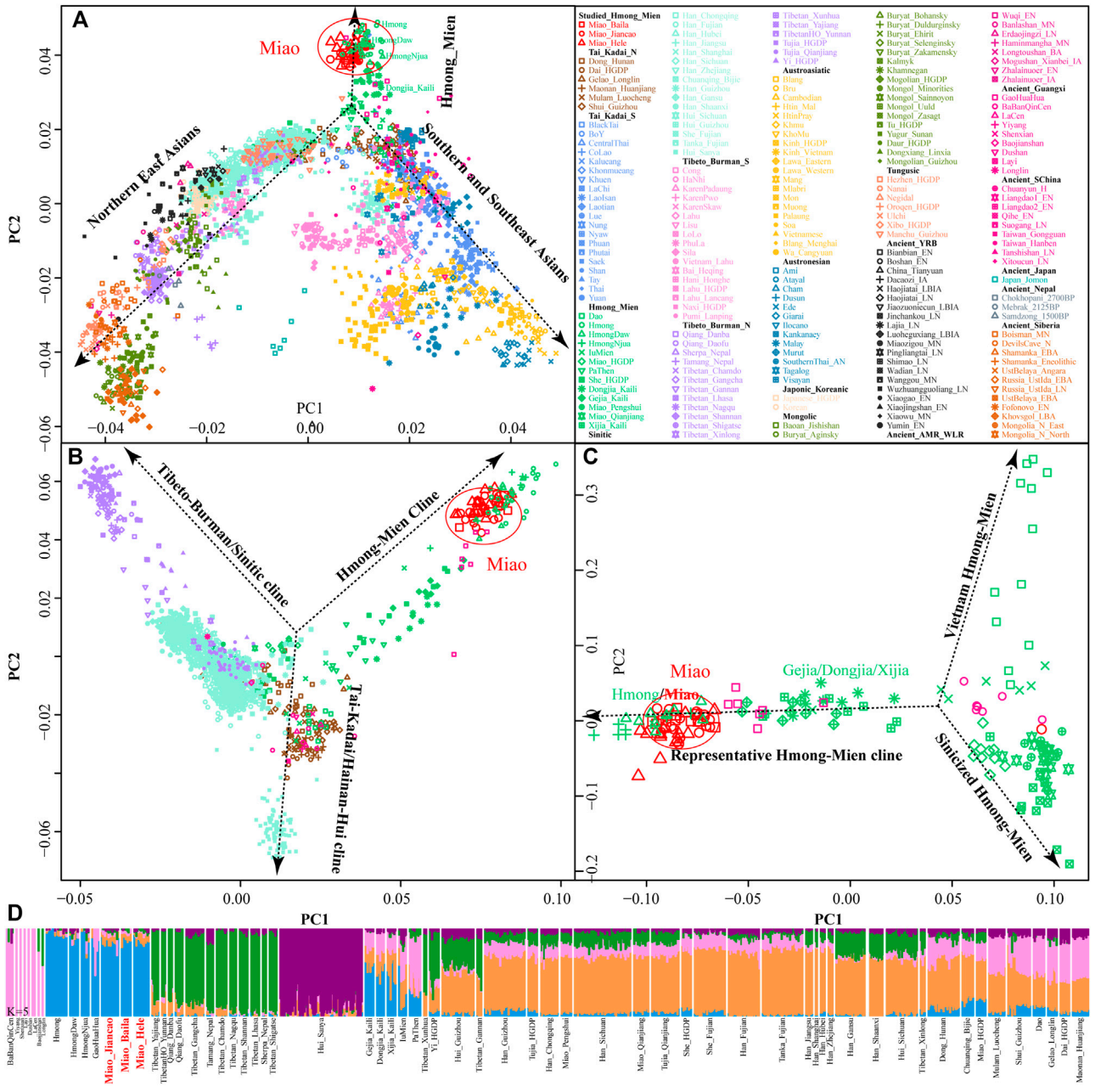

FIGURE 1 | Genetic affinity of HM people in the context of modern and ancient eastern Eurasians. **(A–C)**, Principal component analyses focused on the genetic diversity from East Asian, South Chinese populations, and HM-speaking populations. Included East Asian ancient populations were projected onto the modern genetic background. Populations were color-coded based on geographical and linguistic categories. **(D)** Model-based ADMIXTURE results with five predefined ancestral sources showed the ancestral cluttering pattern and their individual ancestral proportion.

2.3.4 GLOBETROTTER-Based Admixture Estimation

We ran the R program of GLOBETROTTER (Hellenthal et al., 2014) to further identify, date, and describe the admixture events of the target SCM. Both painting samples and copy vectors estimated in the ChromoPainterv2 were used as the basal inputs in the GLOBETROTTER-based estimation. We first ran it to infer admixture proportions, dates, and sources with two specifically predefined parameters (prop.ind: 1; bootstrap. num: 20), and we then reran it with 100 bootstrap samples to estimate the confidence interval of the admixture dates.

2.3.5 Natural Selection Indexes of XPEHH and iHS Estimation

We calculated the integrated haplotype score (iHS) and cross-population extended haplotype homogeneity (XPEHH) using the R package of REHH (Gautier et al., 2017). Here, both northern

Han Chinese from Shaanxi and Gansu provinces and southern Han Chinese from Sichuan, Chongqing, and Fujian provinces were used as the reference in the XPEHH estimation.

2.3.6 Gene Enrichment Analysis

The online tool of Metascape (Zhou et al., 2019) was used to annotate the potentially existed natural selection signatures in the iHS and XPEHH values.

3 RESULTS

3.1 Newly Identified HM Genetic Cline in the Context of East Asian Populations

We genotyped 52 genome-wide SNP data in three SCM populations (Baila, Jiancao, and Hele) and found that five samples possessed close sibship with other samples. After removing relatives, we merged our data with the human origin dataset in AADR (merged HO dataset) to explore the genetic diversity of SCM and their genetic relationship with modern and ancient Eurasian populations. East-Asian-scale PCA results showed three genetic clines (**Figure 1A**), which included the northern East Asian cluster (Altaic and northern ST speakers) and the southern East Asian and Southeast Asian cluster (AA, AN, TK, and southern TB) and the newly identified HM genetic cline. Interestingly, our newly studied three SCM populations separated from other Chinese populations and clustered closely with geographically distant Hmong people from North Vietnam (Hmong) and Thailand (Hmong Daw and Hmong Njua), suggesting their strong genetic affinity and potentially existing common origin history. Dao and Iu Mien clustered closely with TK people, and Miao and She people from Chongqing and other southern China were overlapped with geographically close Han people, which suggested the massive population interaction between HM people and their neighbors. Other HM people, including Geijia, Dongjia, and Xijia in Guizhou, and Pa Then in Vietnam were localized between three genetically different HM genetic lineages.

Focused on the genetic diversity of ST and TK people in China and all studied and reference HM populations, we used a panel of 65 populations and identified three primary directions in the first two dimensions represented by ST, HM, and Hainan Hui people [(top right, top left, and bottom, respectively), **Figure 1B**]. We found that ~500-year-old prehistoric Guangxi GaoHuaHua was localized closely with SCM, but ~1500-year-old BaBanQinCen overlapped with Chinese TK people and HM Dao. Additionally, we explored the finer-scale population relationship within geographically different Miao populations and found that Vietnam Hmong separated from other populations along PC2. After removing this outlier of Hmong, PCA patterns also showed three different genetic clades among the remaining sixteen HM populations, which were represented by the representative HM cline, Sinicized HM, and Vietnam HM [(right, top left, and bottom left, respectively), **Figure 1C**]. These identified population stratifications among HM-speaking populations were confirmed via pairwise Fst genetic distances among 29 Chinese populations based on the Illumina-based dataset (**Supplementary Table S2**) and among 65 populations based on the merged HO dataset (**Supplementary Table S3**). Genetic differences estimated via Fst values showed that SCM had a close genetic relationship with Guizhou HM people (Gejia, Dongjia, and Xijia), followed by geographically different ST groups, northern Mongolic Mongolian, and southern AA populations (Blang and Wa). Results from the lower-density HO dataset not only confirmed the general patterns of genetic affinity between SCM and East Asians reported in the Illumina dataset but also directly identified that SCM possessed the genetic affinity with Hmong people from Vietnam and Thailand among modern reference populations, with GaoHuaHua (Miao_Baila: 0.1398; Miao-Jiancao: 0.1394; Miao_Hele: 0.1419) among ancient Guangxi references.

3.2 Ancestral Composition of HM-Speaking Populations

Consistent with the identified unique genetic cluster of SCM people, we expectedly observed one dominant unique ancestry component in HM-speaking populations (blue ancestry in **Figure 1D**). HM-specific ancestry maximized in Vietnam and Thailand Hmong people as well as existed in SCM and GaoHuaHua with a higher proportion. Different from the gene pool of HM people in Southeast Asia, SCM and ~500-year-old GaoHuaHua people harbored more ancestry related to 1500-year-old historic Guangxi people (pink ancestry). Furthermore, SCM harbored more genetic influence from Sinitic-related populations (orange and purple ancestries) relative to the GaoHuaHua people. A similar pattern was observed in Guizhou populations but with different ancestry proportions, in which Guizhou HM people harbored higher pink and orange ancestries and smaller blue ancestry. This observed pattern of the ancestry composition suggested that Guizhou and Sichuan HM-speaking populations absorbed additional gene flow from northern East Asians when they experienced extensive population movement and interaction. Indeed, other Miao people from Chongqing and She and Miao in the HGDP project possessed similar ancestry composition with neighboring Hans, which supported the stronger extent of admixture between proto-HM and incoming southward Han's ancestor. The admixture signatures in the f_3(East Asians, Miao_Baila; Miao_Jiancao) confirmed that Jiancao Miao was an admixed population and harbored additional genetic materials from northern East Asians (negative Z-scores in LateXiongnu (-3.798), LateXiongnu_han (-3.506), and Han_Shanxi (-3.076)) and southern East Asians (-3.443 in Li_Hainan) (**Supplementary Table S4**). However, no statistically significant negative f_3-values have been identified in the targets of the other two SCM groups. Evidence from the ancient genomes has suggested that prehistoric Guangxi GaoHuaHua people were the temporally direct ancestor of modern Guangxi Miao people (Wang et al., 2021e). However, only marginal negative f_3-values were observed in Jiancao Miao, as f_3(GaoHuaHua, Pumi_Lanping; Miao_Jiancao) = -1.228*SE, although we observed a close cluster relationship in the PCA and ADMIXTURE.

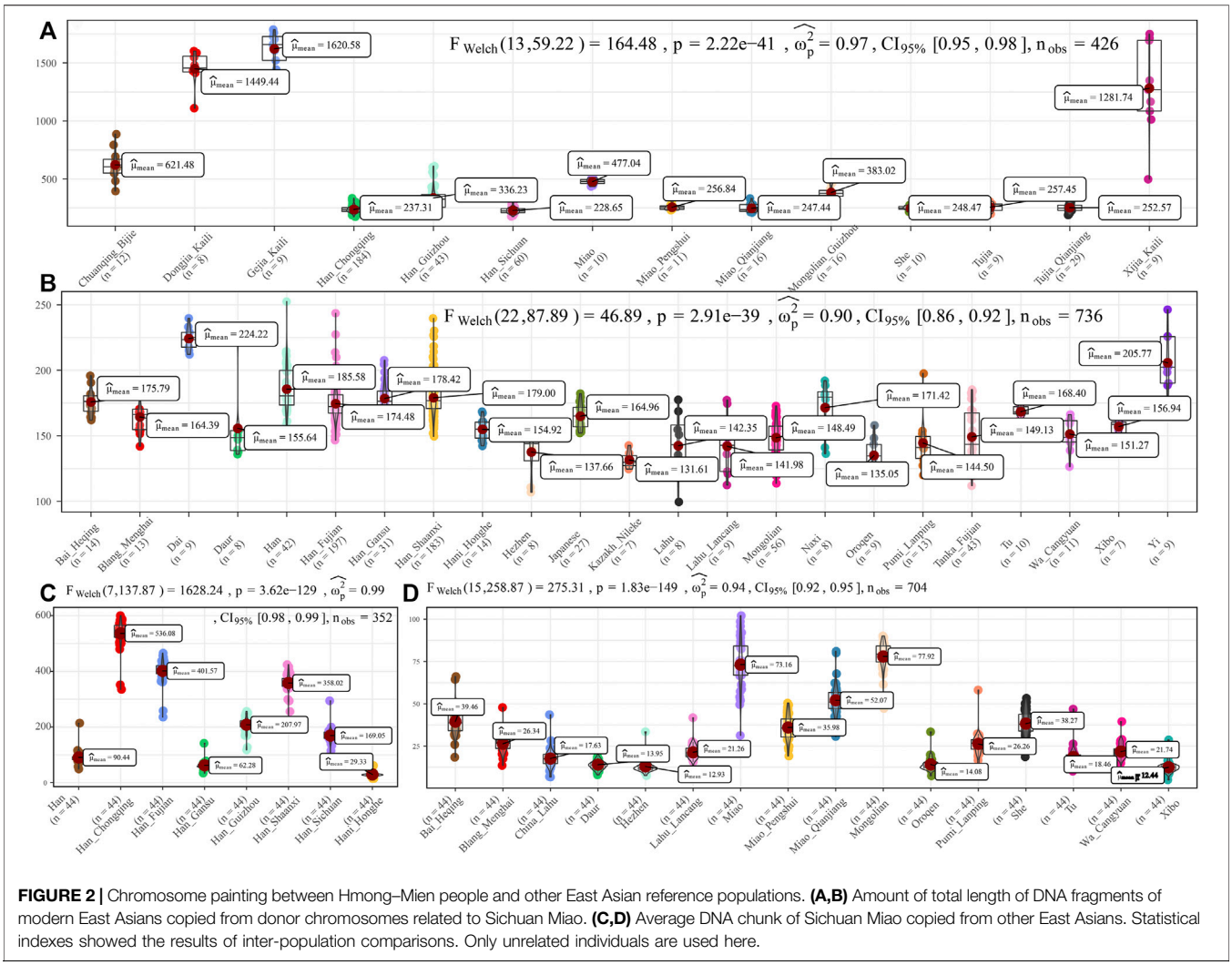

FIGURE 2 | Chromosome painting between Hmong–Mien people and other East Asian reference populations. **(A,B)** Amount of total length of DNA fragments of modern East Asians copied from donor chromosomes related to Sichuan Miao. **(C,D)** Average DNA chunk of Sichuan Miao copied from other East Asians. Statistical indexes showed the results of inter-population comparisons. Only unrelated individuals are used here.

To further characterize the admixture landscape of SCM and other East Asian representative populations based on the sharing haplotypes, we used SCM as the surrogate of the ancestral source and painted all other sampled East Asian populations using ChromoPainter. We found Guizhou HM populations (Gejia, Dongjia, and Xijia) copied the longest DNA chunk from SCM with the total copied chunk length over 1,287.74 centimorgan (**Figure 2A**). SCM also contributed much genetic material to geographically close Miao, Han, and Chuanqing groups (over 237.31 centimorgan) and donated relatively less ancestry to northern Altaic- and southern AA- and TB-speaking populations, including the Wa, Pumi, Lahu, and Bai in geographically close Yunnan Province (**Figure 2B**). Following this, we explored the extent to which other putative East Asian surrogates contributed to the formation of the SCM people. We used other non-HM people as the ancestral surrogate to paint the SCM people, and we found southern Han Chinese donated much ancestry to targeted Miao (**Figure 2C**), even higher than that of southern Miao and She and other southern East Asian indigenous populations (**Figure 2D**), which provided supporting evidence for genetic interactions between HM and southern Sinitic people. Collectively, the ancestral sources related to SCM people served as one unique ancestral proxy that contributed much genetic ancestry to modern East Asians, especially for the HM people.

Although the genetic affinity between SCM and Sinitic Han Chinese was identified, finer-scale population structure inferred from the fineSTRUCTURE showed that SCM possessed a similar pattern sharing ancestry with Guizhou HM people and formed one specific HM branch (**Figure 3**). The inferred PCA patterns based on the sharing haplotypes showed that SCM separated from other Han Chinese and Yunnan AA and TB people and had a close relationship with Guizhou HM people (**Figures 3A–C**). Clustering patterns based on the sharing DNA fragments among population-level and individual-level (**Figures 3D, E**) further confirmed the genetic differentiation between HM people and Sinitic people, which is consistent with the genetic affinity observed in the shared IBD matrix. Additionally, we used the GLOBETROTTER to identify, date, and describe the admixture status of SCM. We first conducted the regional analysis, in which meta-SCM was used as the targeted populations and other East Asians except to Guizhou HM people used as the surrogates. The best-guess conclusion was an unclear signal, which provided

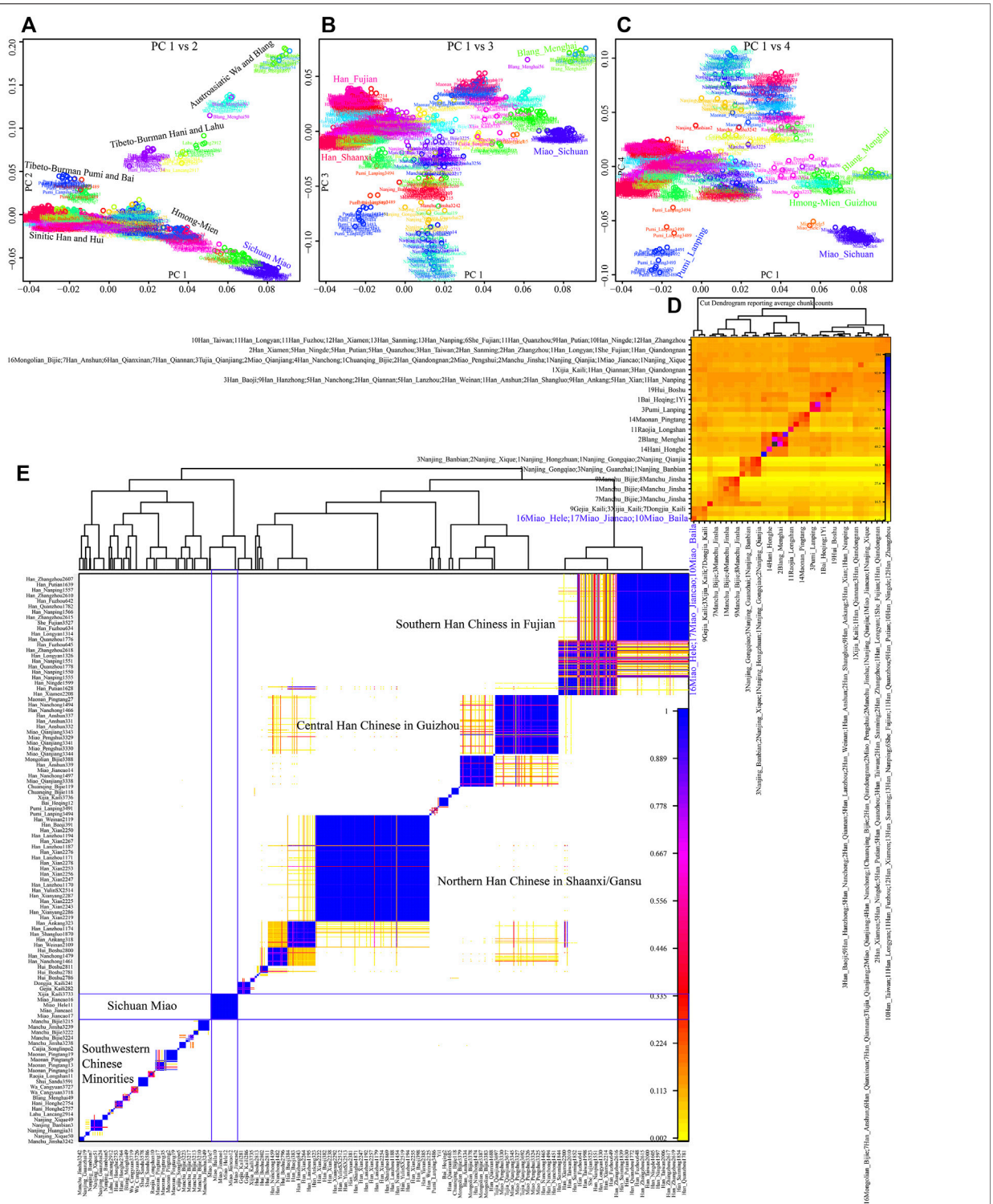

FIGURE 3 | Fine-scale population genetic structure based on the shared haplotype data. **(A–C)** PCA results based on the coancestry matrix showed a genetic relationship among modern East Asians. The color showed the re-classification of the homogenous population label. **(D,E)** Clustering patterns of individual-level and population-level East Asians based on the pairwise coincidence matrixes.

evidence for their unique population history of SCM. Thus, we second performed full analysis to characterize three SCM people conditional on all other sampled East Asian populations as ancestral proximity. We identified recent admixture events in all three geographically different targets. A one-date admixture model for Baila Miao suggested that it was formed via recent

admixture events in seven generations ago with one source related to Jiancao Miao (0.86) and the other source related to Sichuan Han (0.14). A similar admixture model was identified in Hele Miao people, in which the identified one-date model showed that a recent admixture event occurred five generations ago with major ancestry sources related to Jiancao Miao (0.84) and the minor source related to Guizhou Han (0.16). We found a two-date-two-way admixture model best fitted the genetic admixture history of Jiancao Miao. The ancient admixture events occurred 86 generations ago with the Guizhou Gejia as the minor source proximity (0.48) and Baila Miao as the major source proximity (0.52). A recent admixture occurred five generations ago with Baila Miao as the major donor (0.83) and Guizhou Han as the minor donor (0.17). We further estimated the admixture times using ALDER using three SCMs as the targets and all other modern East Asians as the ancestral sources to test the decay of linkage disequilibrium (**Supplementary Table S5**). When we used Guizhou HM people as one of the sources, both population compositions from northern and southern East Asians can produce statistically significant admixture signatures with the admixture times ranging from 22.35+/−6.92 (Maonan) to 160.58 ± 70.32 (Xijia), which also provided supporting clues for the complex ancient admixture events for different ancestral sources.

3.3 Genetic Admixture and Continuity of HM-Specific Ancestry at the Crossroads of East and Southeast Asia in the Past 1,500 Years

To further explore the geographic distribution of our identified HM-dominant ancestry and further constrain the formed time range, we conducted a series of formal tests to validate our predefined phylogenetic topologies. Shared genetic drift inferred from outgroup-f_3-statistics in the form of f_3(SCMs, modern East Asians; Mbuti) suggested that SCM shared a closest genetic relationship with Guizhou HM people, followed by TK people in South China and geographically close Han based on the merged 1240K dataset (**Supplementary Table S6**). The genetic affinity between SCM and Hmong people in Vietnam and Thailand was directly evidenced via the observed largest outgroup-f_3-values in the merged HO dataset, suggesting HM-specific ancestry widely distributed in Sichuan, Guizhou, Guangxi, Vietnam, and Thailand. Focused on the ancient reference populations, we found that historic Guangxi GaoHuaHua people were on the top list for the shared genetic drift (0.3324 for Baila Miao, 0.3317 for Hele Miao, and 0.3304 for Jiancao Miao). 1500-year-old Guangxi BaBanQinCen, the proposed direct ancestor of modern Tai-Kadai people (Wang et al., 2021e) and Iron Age Taiwan Hanben, the proposed ancestor of modern Austronesian people (Wang et al., 2021a) also possessed a strong genetic affinity with SCM, suggesting the possibility of their common origin history, and possibly originated from South China. These patterns of genetic affinity among spatiotemporally different southern East Asians were consistent with the shared characteristics attested by cultural, linguistic, and archeological documents.

To further explore the genetic relationship between ancient Guangxi populations and modern ethnolinguistic populations, we conducted pairwise qpWave analysis among 16 HM populations, five Guangxi ancient groups (GaoHuaHua, BaBanQinCen, Baojianshan, Dushan, and Longlin), seven TK-, 16 Sinitic-, and 18 TB-speaking populations (**Figure 4**). We found genetic homogeneity existed within populations from geographically and linguistically close populations, especially in TB, Sinitic, and HM. Here, we only observed strong genetic affinity within geographically diverse HM people and found genetic heterogeneity between historic Guangxi populations and modern HM people. Considering different admixture models identified among three SCM populations, we performed symmetrical f_4-statistics in the form of f_4(SCM1, SCM2; reference populations, Mbuti) (**Supplementary Table S7**). We also identified the differentiated evolutionary history among them; Jiancao Miao shared more alleles with Guizhou HM people than Miao people from Baila and Hele and Jiancao Miao also shared more northern East Asian ancestry related to the other two Miao populations. The results from another version of symmetrical f_4-statistics in the form of f_4(reference1, reference2; SCM, Mbuti) first confirmed the strong genetic affinity between SCM people and other HM people, as most negative f_4-values identified in f_4(reference1, HM; SCM, Mbuti) (**Supplementary Table S8**). All 126 tested f_4(Reference, GaoHuaHua; SCM, Mbuti) values were negative, and 123 out of 126 were statistically significant, which suggested the SCM shared more ancestry and a closer genetic relationship with GaoHuaHua relative to other modern and ancient East Asians. We also tested f_4(Reference, SCM; GaoHuaHua, Mbuti) (**Supplementary Table S9**) and found GaoHuaHua shared more alleles with SCM than all reference populations. These observed results were consistent with the hypothesis of SCM people being the descendants or their relatives of historic Guangxi GaoHuaHua. We also tested f_4(GaoHuaHua, SCM; reference, Mbuti) and found additional gene flow from ancestral sources related to late Neolithic populations from the YRB, as observed negative f_4-values in f_4(GaoHuaHua, Miao_Hele; Han_Gansu, Mbuti) = −3.78*SE or f_4(GaoHuaHua, Miao_Baila; China_Upper_YR_LN, Mbuti) = −3.252*SE. Indeed, we previously observed admixture signatures in Jiancao Miao in admixture-f_3(GaoHuaHua, northern East Asians; Jiancao Miao), which suggested SCM shared major ancestry from GaoHuaHua and also experienced additional genetic admixture from northern East Asians.

Focused on the deeper temporal population dynamics, we next tested the genetic relationship between SCM and ~1500-year-old BaBanQinCen using the same strategies (**Supplementary Table S9**). Positive results in f_4(Dongjia/Maonan/China_SEastAsia_Coastal_LN/Guangxi_1500BP, SCM; BaBanQinCen, Mbuti) showed that BaBanQinCen shared more derived alleles with late Neolithic and Iron Age Fujian populations and other spatiotemporally close Guangxi historic populations. Statistically significant values in f_4(BaBanQinCen, SCM; reference, Mbuti) further confirmed that BaBanQinCen did

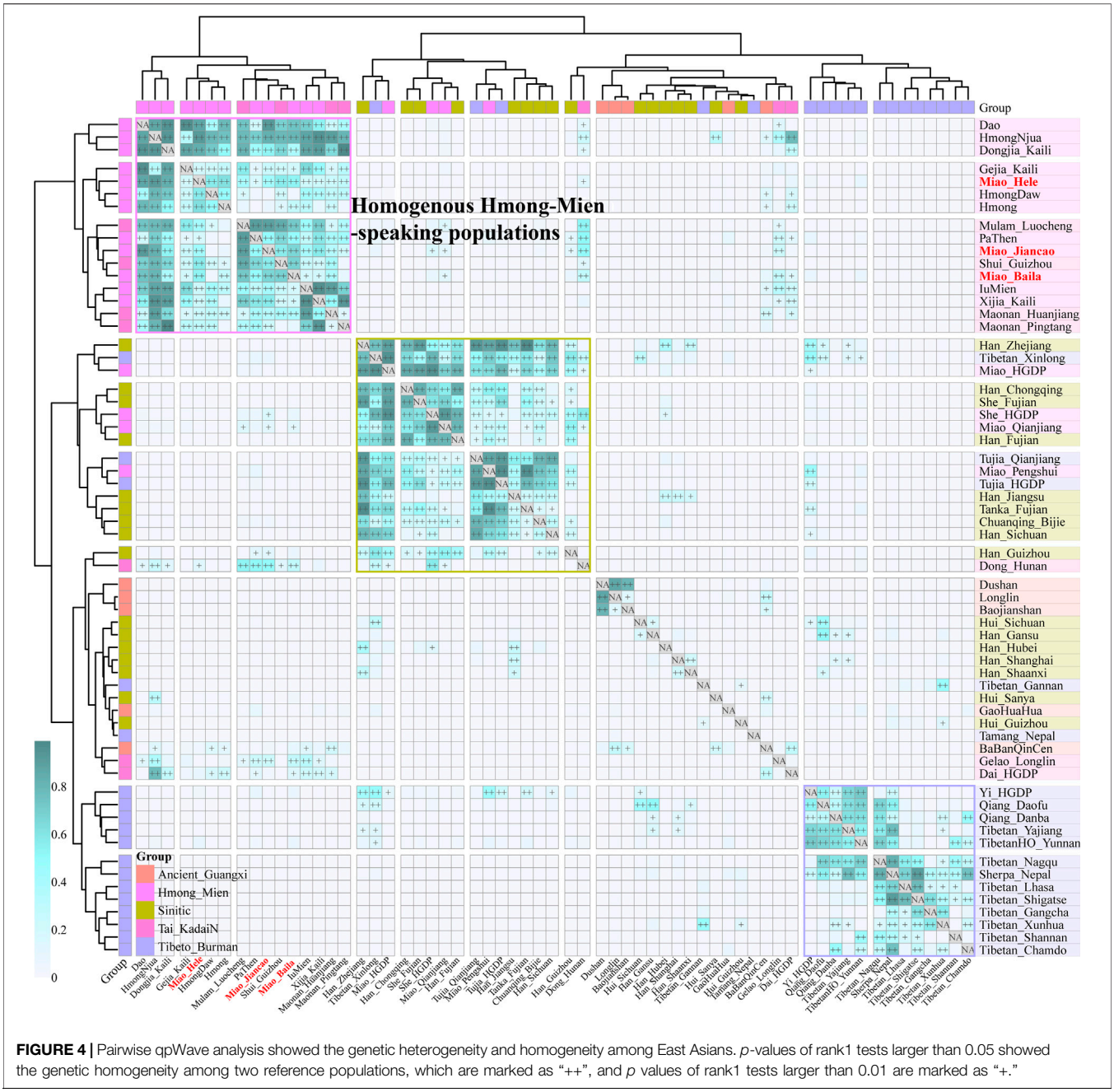

FIGURE 4 | Pairwise qpWave analysis showed the genetic heterogeneity and homogeneity among East Asians. *p*-values of rank1 tests larger than 0.05 showed the genetic homogeneity among two reference populations, which are marked as "++", and *p* values of rank1 tests larger than 0.01 are marked as "+."

not form a clade with SCM and shared more alleles with pre-Neolithic Amur River people (AR14K), Neolithic-to-Iron Age Fujian populations, and indigenous Guangxi prehistoric populations (Baojianshan and Dushan) than SCM, which was further supported via the f_4-statistics focused on other ~1500-year-old Guangxi populations (Guangxi_1500BP) and Taiwan Hanben. But, SCM shared more genetic influence from northern East Asians than ~1500-year-old Guangxi people. Compared with other Guangxi prehistoric populations [f_4(Longlin, Baojianshan, and Dushan, reference; SCM, Mbuti)], SCM shared much ancestry with ancient northern East Asians, southern Fujian, and modern East Asian ancestry. Compared with SCM, prehistoric Guangxi populations shared much Neolithic to Iron Age Fujian and Guangxi ancestries. We also tested the genetic relationship between SCM and YRB farmers using asymmetric-f_4-statistics and found YRB millet farmers shared more alleles with SCM people than with early Asians and southern Fujian and Fujian ancient populations. As expected, SCM harbored many HM-related alleles or ancient Fujian and Guangxi ancestries compared with millet farmers. Generally, formal test results demonstrated that SCM possessed the strongest genetic affinity with ~500-year-old Guangxi GaoHuaHua people and additionally obtained genetic influx from northern East Asians recently.

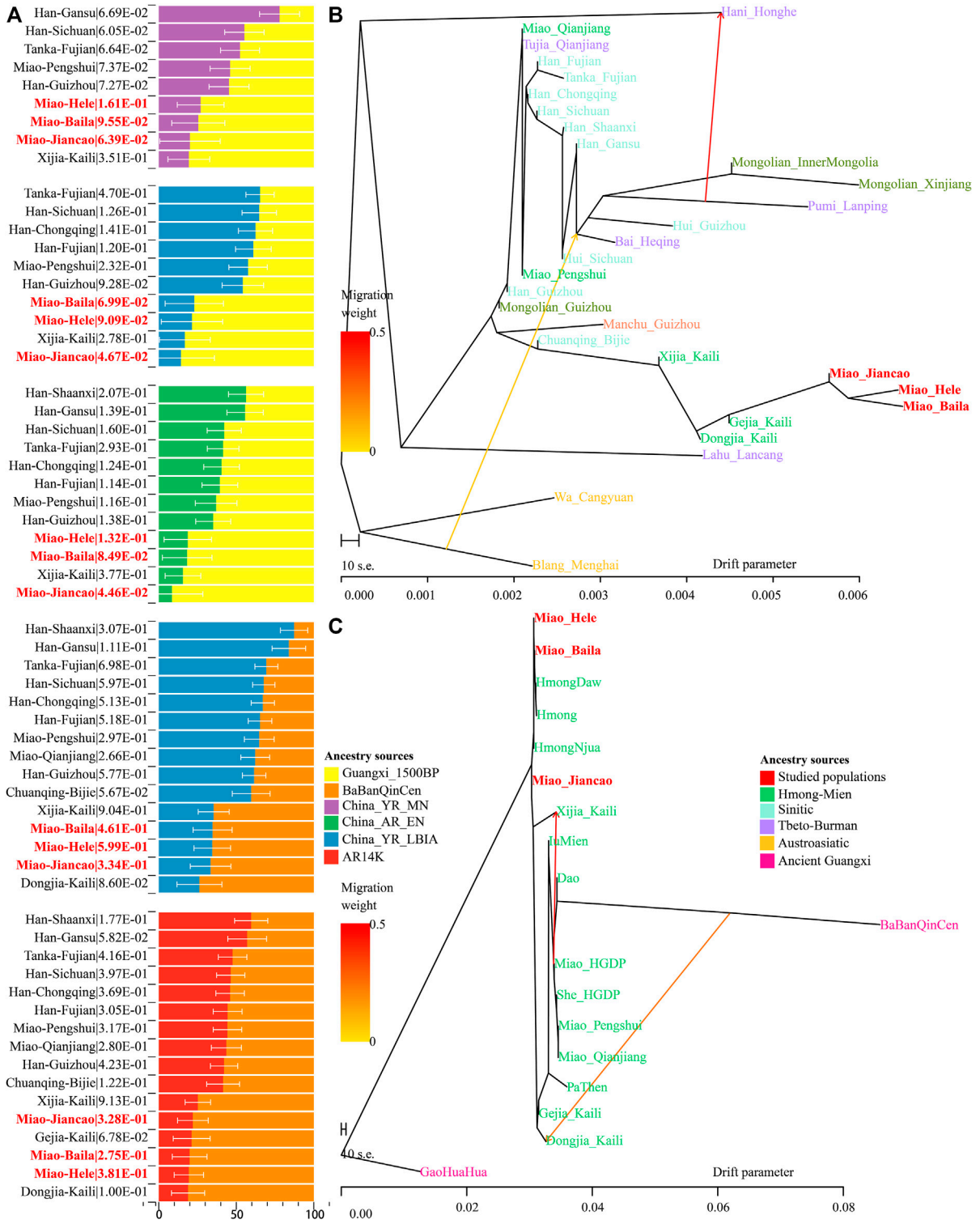

FIGURE 5 | Results of qpAdm models and TreeMix-based phylogenies. **(A)** Two-way admixture models showed ancestry comparison in different ancestral source pairs. **(B,C)** TreeMix-based phylogenetic tree with two migration events showed the genetic relationship between East Asians.

3.4 Admixture Evolutionary Models

A close genetic relationship between Guangxi historic populations and SCM has been evidenced in our descriptive analyses and quantitative *f*-statistics. We further conducted two-way qpAdm models with two Guangxi ancient populations as the southern surrogates and four northern ancient populations from YRB and Amur River as the northern ancestral sources to estimate the ancestral composition of SCM and their ethnically and geographically close populations (**Figure 5A**). When we used BaBanQinCen as the source, we tested the two-way admixture models: proportion of ancestry contribution of historic Guangxi population ranged from 0.811 ± 0.107 in Kali Dongjia to 0.404 ± 0.107 in Shaanxi Hans in the AR14K-BaBanQinCen model and spanned from 0.738 ± 0.145 to 0.127 ± 0.088 in Shaanxi Hans in the China_YR_LBIA-BaBanQinCen model. SCM derived 0.780–0.806 ancestry from historic Guangxi ancestry in the former model and 0.653–0.666 ancestry from it in the latter model (**Figure 5A**). We also confirmed that the unique gene pool of SCM derived from major ancestry from Guangxi and minor ancestry from North East Asians via the additional two qpAdm admixture models with early Neolithic Amur River Hunter-Gatherer and middle Neolithic-to-Iron Age YRB farmers as the northern sources.

Until now, to explore the population genetic diversity of Chinese populations and provide some pilot works supporting the initiation of the Chinese Population Genome Diversity Project (CPGDP) based on the deep whole-genome sequencing on anthropologically informed sampling populations, we have genotyped the array-based genome-wide SNP data in 29 ethnolinguistically different populations. We reconstructed phylogenetic relationships between three studied SCM populations and 26 other Chinese populations from ST, Altaic, AA, and HM (**Figure 5B**). We identified that branch clusters were consistent with the linguistic categories and geographical division. Tibetan Lahu and Hani clustered closely with AA Blang and Wa, and other populations were clustered as the northern and southern East Asian branches. The southern branches consisted of our newly studied Miao and Guizhou HM people and Guizhou Chuanqing and Manchu. The northern branch comprised Mongolic, TB, and Sinitic people. We found that two Chongqing Miao populations clustered closely with the northern branch, suggesting much genetic material mixed from surrounding Han Chinese populations. We also identified regional population gene flow events from ethnically different populations, such as gene flow events from Pumi to Hani and from Blang to common ancestral lineage of Bai, Pumi, and Mongolian. To directly reconstruct the phylogenies between the HM population and historic Guangxi populations, we merged 16 HM-speaking populations with GaoHuaHua and BaBanQinCen and found two separated branches respectively clustered closely with GaoHuaHua and BaBanQinCen (**Figure 5C**). A close phylogenetic relationship among SCM, Guangxi GaoHuaHua, Guizhou Gejia, Dongjia, and Xijia, and Vietnam and Thailand Hmong further supported the common origin of geographically different HM people.

We finally reconstructed the deep population admixture history of HM-speaking populations using the qpGraph model with population splits and admixture events. We used the ancestral lineage of Mbuti in Africa, Loschbour in western Eurasia, Onge in South Asia, and Tianyuan in East Asia as the basal deep early continental lineages. We used Baojianshan in the early Neolithic period and GaoHuaHua in the historic time from Guangxi and Qihe in the early Neolithic in Fujian as southern East Asian lineages and used Neolithic YRB millet farmers as the northern East Asian lineages. In our first best-fitted model (**Figure 6A**), we added additional late Fujian Xitoucun and Tanshishan from the late Neolithic period, we found GaoHuaHua could be fitted as major ancestry related to upper Yellow River Qijia people (0.52) and minor ancestry related to late Neolithic Fujian people (0.48). However, SCM derived much more ancestry from northern East Asians (0.82) in this model, suggesting additional northern East Asian gene flow influenced the genetic formation of modern HM-speaking populations. In the second best-fitted model (**Figure 6B**), we added hunter-gatherer lineage from the Mongolian Plateau (Bosiman) and found Xuyong Miao could be fitted as 0.86 ancestry from GaoHuaHua and the remaining ancestry from Qijia people (0.14). The third best-fitted model (**Figure 6C**) with adding Australian lineage also replicated the shared major ancestry between GaoHuaHua and Xuyong Miao. In the final version of the qpGraph model (**Figure 6D**), we added the American indigenous lineages, in which Miao was fitted as 0.37 ancestry from western Eurasian and 0.63 ancestry from East Asians. Xuyong Miao was modeled as a similar ancestry composition as the third model. Here, we should be cautious about the differences in the topologies of the early deep lineages when different populations were added to our basal models. The detailed true phylogenetic relationship should be further explored and reconstructed via denser spatiotemporally different early Asian population sequencing data. But, the consistent pattern of Miao's genetic profiles of major ancestry from GaoHuaHua and minor ancestry from northern East Asia was obtained from four different admixture models, suggesting it is valuable to illuminate the simple model of the formation of modern SCM.

3.5 Uniparental Founding Lineages

We obtained high-resolution uniparental maternal and paternal lineages in SCM (**Supplementary Table S10**). We identified four dominant maternal founding lineages in SCM [(B5a1c1 (0.3462), F1g1 (0.1346), B4a (0.0769), and F1a (0.0769)]. We also identified two paternal founding lineages [(O2a2a1a2a1a2 (0.3913) and O2a1c1a1a1a1a1a1b (0.1739)] in SCM, which is consistent with the hypothesis of the primary ancestry of Miao originated from southern Chinese indigenes. In detail, we observed 10 terminal paternal lineages among 23 males and 17 terminal maternal lineages in 52 females. Compared with geographically close Chongqing Han populations, we found a significant difference in the frequency of major lineages between Chongqing Han and Sichuan Miao (**Figure 7**).

3.6 Natural Selection Signatures and Their Biological Adaptation

Genetic studies have identified many biologically adaptive genes or pathways in ethnolinguistically diverse populations. Evolutionary adaptative mutations could be accumulated and generated as longer extended haplotype homozygosity with their increase of allele

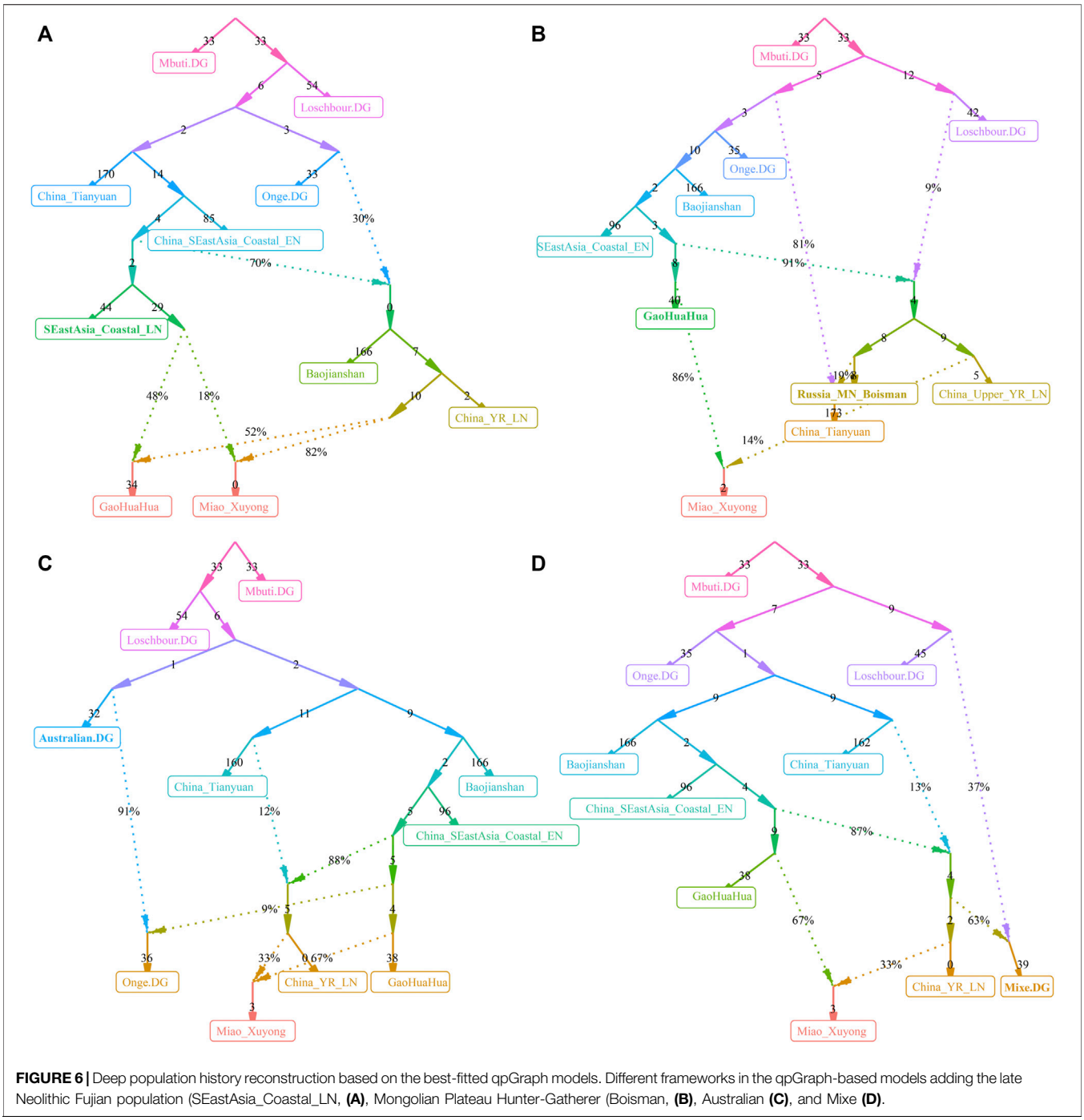

FIGURE 6 | Deep population history reconstruction based on the best-fitted qpGraph models. Different frameworks in the qpGraph-based models adding the late Neolithic Fujian population (SEastAsia_Coastal_LN, **(A)**, Mongolian Plateau Hunter-Gatherer (Boisman, **(B)**, Australian **(C)**, and Mixe **(D)**.

frequency of the initial mutations. We scanned for candidates of the positive selections using iHS and XPEHH in SCM. We first calculated XPEHH values for Miao using northern Han as a reference population and identified obvious candidates in chromosomes 1-3, 9, 20, and 22 (**Figure 8A**). Chromosome 1 showed selection signals in the vicinity of the *neuroblastoma breakpoint family member 9/10* (NBPF 9/10) locus, reflecting well-known signals associated with susceptibility of the neuroblastoma. We further identified a strong selection signal implicating *polypeptide N-acetylgalactosaminyltransferase 13* (GALNT13) and *potassium voltage-gated channel subfamily J member 3* (KCNJ3) located in chromosome 3. The former one is expressed in all neuroblastoma cells and encodes a glycosyltransferase enzyme responsible for the synthesis of O-glycan. The latter one encodes G proteins in the potassium channel and is associated with susceptibility candidates for schizophrenia (Yamada et al., 2012). We also identified four top candidate genes in chromosome 3, including the *abhydrolase domain containing 10* (ABHD10), *RNA-binding motif single-stranded interacting protein 3* (RBMS3), *RBMS3 antisense RNA 3*

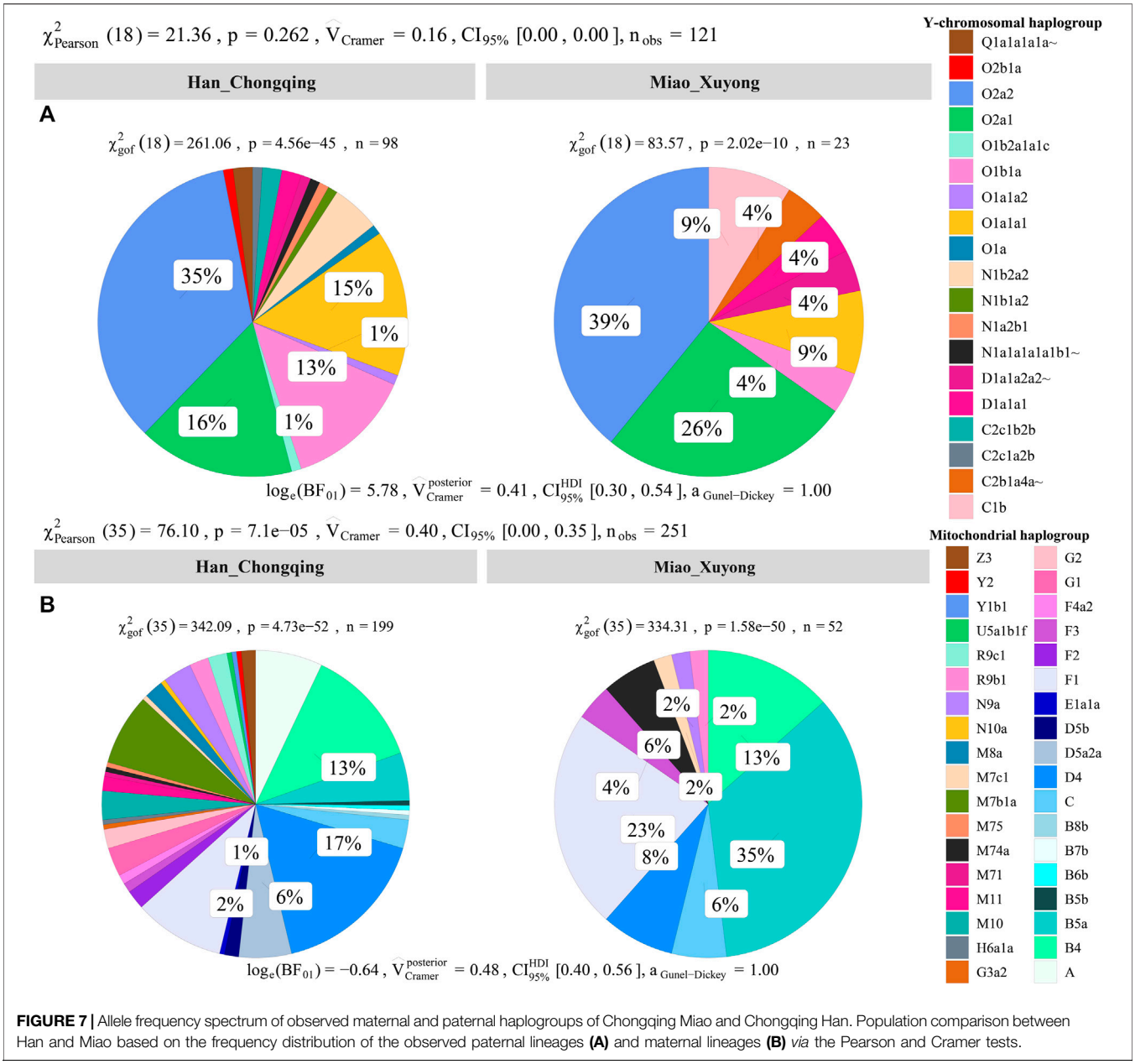

FIGURE 7 | Allele frequency spectrum of observed maternal and paternal haplogroups of Chongqing Miao and Chongqing Han. Population comparison between Han and Miao based on the frequency distribution of the observed paternal lineages **(A)** and maternal lineages **(B)** *via* the Pearson and Cramer tests.

(RBMS3-AS3), and *transgelin 3* (TAGLN3). ABHD10 is one of the important members of the AB hydrolase superfamily and is associated with enzymes for deglucuronidation of mycophenolic acid acyl-glucuronide (Iwamura et al., 2012). RBMS3 encodes protein-binding Prx1 mRNA in a sequence-specific manner *via* binding poly(A) and poly(U) oligoribonucleotides and controls Prx1 expression and indirectly collagen synthesis (Fritz and Stefanovic, 2007). It also served as the tumor-suppressor gene associated with lung squamous cell carcinoma and esophageal squamous cell carcinoma (Li et al., 2011). TAGLN3 encodes a cytoskeleton-associated protein and is reported to possess an association with schizophrenia (Ito et al., 2005). Chromosome 8 shows a selection signal of *myotubularin-related protein 7* (MTMR7), which was localized at and associated with the susceptibility of Creutzfeldt–Jakob risk. Three top genes were identified in chromosome 9, which included *contactin-associated protein-like 3B* (CNTNAP3B), *phosphoglucomutase 5 pseudogene 2* (PGM5P2), and *SWI/SNF-related, matrix-associated, actin-dependent regulator of chromatin, subfamily A, member 2* (SMARCA2). SMARCA2 encodes the protein-controlled coactivator participating in transcriptional activation and vitamin D-coupled transcription regulation. Genetic evidence has shown the association between its genetic polymorphisms and the susceptibility of schizophrenia (Sengupta et al., 2006), Nicolaides–Baraitser syndrome (Van Houdt et al., 2012), and lung cancer (Oike et al., 2013). *ADAM metallopeptidase domain 12* (ADAM12) situates in chromosome 10, and ADAM12 encodes trans-membrane metalloproteinase, which can secrete glycoproteins that are

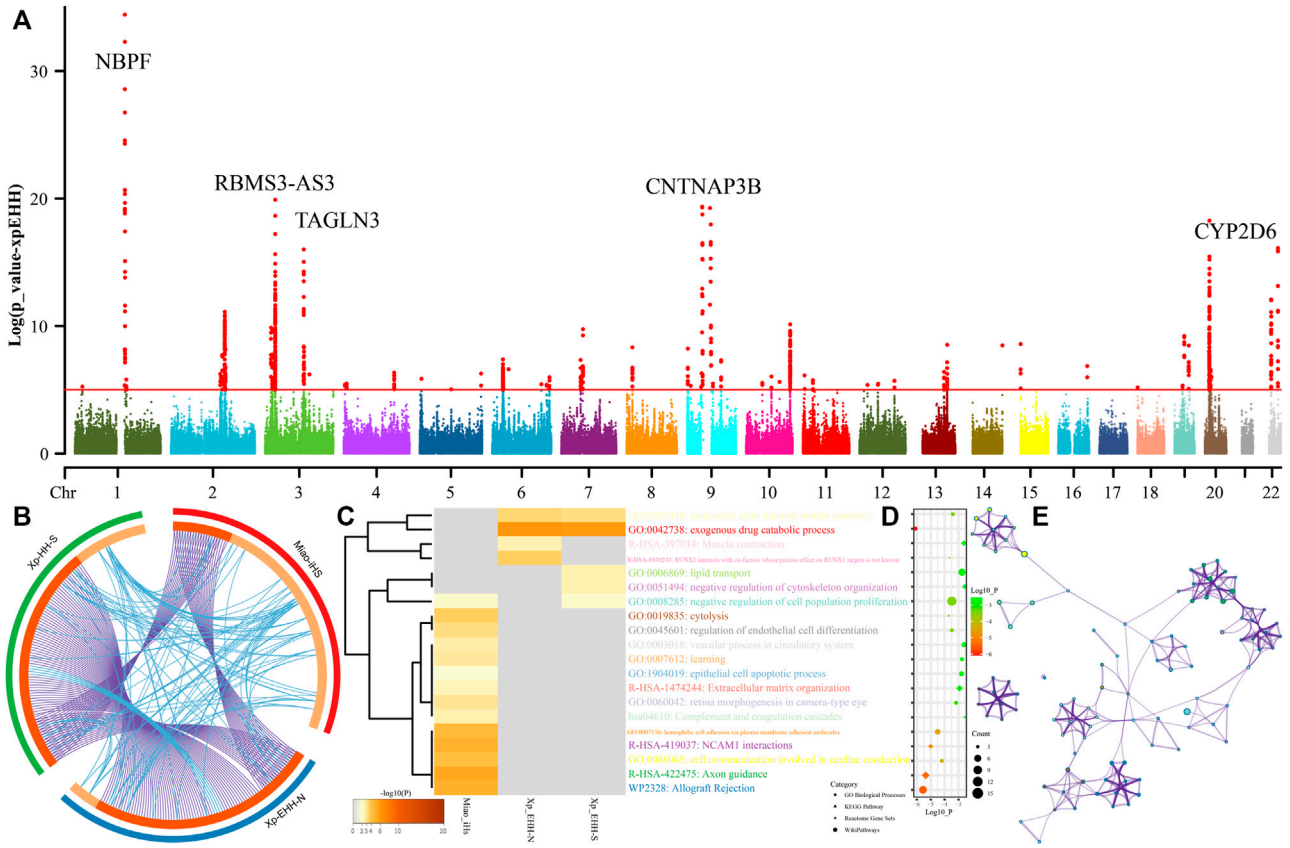

FIGURE 8 | Manhattan showed the natural selection signatures and enrichment analysis. **(A)** *p*-values of XPEHH in the Miao population using northern Han as the reference population. **(B)** Overlap among three gene lists based on gene-level and shared-term level, where blue curves link genes that belong to the same enriched ontology term. The inner-circle represents gene lists, where hits are arranged along the arc. Genes that hit multiple lists are colored in dark orange, and genes unique to a list are shown in light orange. **(C)** Heatmap of top twenty enriched terms across three input gene lists, colored by *p*-values. **(D)** Top 20 clusters with their representative enriched terms. **(E)** Network of enriched terms colored by cluster-ID.

involved in cell–cell interaction, fertilization, and muscle development. We also identified natural selection signatures in *cytochrome P450 family 2 subfamily A member 6* (CYP2A6), *isthmin 1* (ISM1), and *cytochrome P450 family 2 subfamily D member 6* (CYP2D6).

We further calculated another set of XPEHH scores using southern Han Chinese as the reference population and iHS scores in the SCM populations. To explore the biological functions of all possible naturally selected genes (102 loci in iHS-based, 93 XPEHH_N-based, and XPEHH_S-based), we made enrichment analysis based on three sets of identified natural-selection genes. Loci with *p*-values of XPEHH scores larger than 5 and normalized iHS scores larger than 3.3 were used in the enrichment analysis *via* the Metascape. Overlapping loci observed among three gene candidate lists showed the more common gene candidates inferred from XPEHH and less overlapping loci between XPEHH-based loci and iHS-based loci (**Figure 8B**). A heatmap based on *p*-values of enrichment pathways (**Figures 8C–E**) showed that all three ways identified the candidate genes associated with *metabolic process (GO:0008152)*, *response to stimulus (GO:0050896)*, *cellular process (GO:0009987)*, *regulation of biological process (GO:0050789)*, *biological adhesion (GO:0022610)*, and *developmental process (GO:0032502)*. The results from the iHS also showed other top-level gene ontology biological processes, which included immune system process (GO:0002376), biological regulation (GO:0065007), positive regulation of *biological process (GO:0048518)*, *behavior (GO:0007610)*, *signaling (GO:0023052)*, *multicellular organismal process (GO:0032501)*, *locomotion (GO:0040011)*, *negative regulation of biological process (GO:0048519) and localization (GO:0051179)*, the detailed enriched terms, pathways, and processes enrichment analysis and their networks of top twenty clusters showed in do not reveal the previously reported naturally selected loci-associated pigmentation, alcohol metabolism, and other common adaptive signals (EDAR et al.) of East Asians (Mao et al., 2021).

4 DISCUSSION

4.1 Unique Genetic History of HM-Speaking Populations

Genetic diversity and population history of East Asians have been comprehensively explored and reconstructed in the past 20 years *via* lower-density genetic markers (STRs, SNPs, and InDels) and higher-

density array-based genome-wide SNPs and whole-genome sequencing data, which advanced our understating of the origin, diversification, migration, admixture, and adaptation of Chinese populations (Chen et al., 2009; Consortium et al., 2009; Xu et al., 2009; Cao et al., 2020; Wang et al., 2021a). As we all know that the International Human Genome Organization (HUGO) initiated the broader Human Genome Diversity Project (HGDP) in 1991. The HGDP aimed at illuminating the structure of genomes and population genetic relationships among worldwide populations via initial array-based genome-wide SNPs and recent whole-genome sequencing (Bergstrom et al., 2020). A similar work of the CHGDP was publicly reported in 1998 (Cavalli-Sforza, 1998), in which Chu et al. first comprehensively reported genetic relationships and general population stratification based on STR data (Chu et al., 1998). Six years later, Wen et al. illuminated that demic diffusion of northern East Asians contributed to the formation of the genetic landscape of modern Han Chinese populations and their sex-biased admixture processes via uniparental markers (Y-chromosome SNPs/STRs and mitochondrial SNPs) (Wen et al., 2004). The next important step occurred around 2009, and several genetic analyses based on genome-wide SNPs, including mapping Asian genetic diversity reported by the HUGO Pan-Asian SNP consortium, have identified population stratification among linguistically different Asian populations and genetic differentiation between northern and southern Han Chinese populations (Chen et al., 2009; Consortium et al., 2009; Xu et al., 2009). However, these studies had limitations of the lower resolution of used marker panel or limited representative samples from the ethnolinguistic region of China. Recently, large-scale genetic data from the Taiwan Biobank, China Metabolic Analytics Project (ChinaMAP), and other low-coverage sequencing projects (Chiang et al., 2018; Liu et al., 2018; Cao et al., 2020; Lo et al., 2021) have reconstructed fine-scale genetic profiles of the major populations in China and reconstructed a detailed framework of the population evolutionary history. Cao et al. identified seven population clusters along with geographically different administrative divisions (Li et al., 2021), which is consistent with our recently identified differentiated admixture history of geographically different Han Chinese populations possessing major ancestry related to northern East Asians and additional gene influx from neighboring indigenous populations (He et al., 2021a; He et al., 2021b; Liu et al., 2021b; Wang et al., 2021c; Yao et al., 2021). Genetic studies focused on ethnolinguistic Chinese regions further identified different genetic lineages in modern East Asians, TB lineage in the Tibetan Plateau, Tungusic lineage in the Amur River Basin, and AA and AN lineage in South China and Southeast Asia (Siska et al., 2017; Wang et al., 2021a). Recent ancient genomes also identified differentiated ancestral sources that existed in East Asia since the early Neolithic, including Guangxi, Fujian, Shandong, Tibet, and Siberia ancestries (Yang et al., 2020; Mao et al., 2021; Wang et al., 2021a; Wang et al., 2021e). However, many gaps of Southwest Chinese indigenous populations needed to be completed in the Chinese HGDP-based anthropological sampling and Trans-Omics for Precision of Medicine of the Chinese population (CPTOPMed). Large-scale genomic data from ethnolinguistically different populations may be provided new insights into the population history and medical utilization in the precision medication for East Asians such as the UK10K and TOPMed (Wang et al., 2021d; Taliun et al., 2021).

To comprehensively provide a complete picture of the genetic diversity of China and make comprehensive sampling and sequencing strategies in the next whole-genome sequencing projects, it is necessary to explore the basal pattern genetic background using the small sample size and array genotyping technology. As our part of the initial pilot work in the CPGDP based on anthropologically informed sampling, we reported genome-wide SNP data of 55 SCM samples from three geographically diverse populations. Our analysis reveals the key features of the landscape of southwestern HM lineage, including the identified unique HM cline in East-Asian-scale PCA and population stratification in regional-scale-PCA, the observed dominant specific ancestry in geographically distant HM people, the estimated strong genetic affinity among HM people *via* the Fst, outgroup-f_3-statistics, and f_4-statistics. We further confirmed that stronger genetic affinity within HM people via the sharing patterns of DNA fragments in the IBD, chromosome painting, and fineSTRUCTURE as well as the attested close-clustered pattern in TreeMix-based phylogeny and close phylogenetic relationships between HM people and 500-year-old GaoHuaHua people. Admixture models based on the two-way models further found the dominant 1500-year-old Guangxi historic ancestry in modern HM people. These observed genetic affinities between HM people from Sichuan, Guizhou, Vietnam, and Thailand suggested that all modern HM people possessed a common origin. Combining previous cultural, linguistic, and archaeogenetic evidence, the most originated center of modern HM people is the Yungui Plateau in Southeast China. We also found that Miao from Chongqing and HGDP and She people shared more ancestry with Han Chinese populations, suggesting some HM people also obtained much genetic material with southward Han Chinese populations. Compared with historic Guangxi populations (BaBanQinCen and GaoHuaHua), SCM shared much derived ancestry with northern East Asians, suggesting that the persistent southward gene flow from northern East Asians influenced the modern genetic profile of HM people. Based on the admixture times dated via GLOBETROTTER and ALDER, complex population migration and admixture events occurred in the historic and prehistoric proto-HM people. Spatiotemporal analysis between modern HM people and their genetic evolutionary relationship with surrounding modern ethnolinguistically diverse populations as well as the genetic relationship between ancient Yellow River millet farmers and Fujian and Guangxi ancient populations suggested that HM people originated from the crossroad region of Sichuan and Guizhou provinces. Modern HM people may have remained the most representative ancestry of ancient Daxi and Shijiahe people in the middle Yangtze River basin, which needed to be validated directly *via* ancient genomes in this region.

4.2 Specific Genomic Patterns of Natural Selection Signatures

Ethnically different populations undergoing historical differences in the pathogen exposure may remain as different patterns of the allele frequency spectrum and extended haplotype homozygosity under

natural selection processes. We identified different natural selection candidates (NBPF9, RBMS3-AS3, CNTNAP3B, NBPF10, CYP2D6, TAGLN3, ISM1, RBMS3, KCNJ3, ADAM12, GALNT13, PGM5P2, CYP2A6, MTMR7, and SMARCA2) associated with several different biological functions (metabolic process, response to stimulus, cellular process, and regulation of biological processes) in Miao people compared with other East Asians. Denisovan archaic high-altitude adaptive introgression signals were observed in Tibetans (EPAS1 and EGLN1), which is not observed in HM people with obvious natural selection signatures (Yi et al., 2010). More Denisovan archaic adaptive introgression signals related to immune function (TNFAIP3, SAMSN1, CCR10, CD33, DDX60, EPHB2, EVI5, IGLON5, IRF4, JAK1, ROBO2, PELI2, ARHGEF28, BANK1, LRRC8C and LRRC8D, and VSIG10L) and metabolism (DLEU1, WARS2, and SUMF1) (Choin et al., 2021) were identified in Austronesian and Oceanian populations. But, we only observed immune-related Denisovan introgression signals in the DCC gene situated in chromosome 18, which underwent the natural selection evidenced via a higher iHS score (3.5517 in rs17755942, 3.4758 in rs1237775, 3.3540 in rs16920, and 3.3299 in rs79301210) in SCM. Choin et al. also reported Neanderthal adaptive introgression genes in Oceanians, including dermatological or pigmentation phenotypes (OCA2, LAMB3, TMEM132D, SLC36A1, KRT80, FANCA, and DBNDD1), metabolism (LIPI, ZNF444, TBC1D1, GPBP1, PASK, SVEP1, OSBPL10, and HDLBP), immunity (IL10RA, TIAM1, and PRSS57), and neuronal development (SIPA1L2, TENM3, UNC13C, SEMA3F, and MCPH1) (Choin et al., 2021). However, our analysis based on the XPEHH scores only identified one Neanderthal introgression immunity signal (CNTN5) and one pigmentation phenotype signal (PTCH1). CNTN5 harbored high XPEHH scores (>2.1313) ranging from 99577624 to 99616124 in chromosome 11 with the highest values of 4.2829 in rs7111400. Loci situated from 98209156 to 98225683 in PTCH1 in chromosome 9 also possessed higher EPEHH scores in HM people with the highest values in missense mutation rs357564 (4.5412). ALDH2 and ADH1B were reported to possess a strong association with alcohol metabolism (Taliun et al., 2021); however, the highest XPEHH absolute scores in HM people were less than 0.5937 for ALDH2 and 1.6013 for ADH1B. Five selection-candidate genes of CTNNA2, LRP1B, CSNK1G3, ASTN2, and NEO15 were evidenced to have undergone natural selection in Taiwan Han populations (Lo et al., 2021); however, only LRP1B associated with lipid metabolisms was evidenced and replicated in HM people. The observed differentiated patterns of the genomic selection process in HM people are consistent with their reconstructed unique population history and specific living environments in Southwest China. Thus, further whole-genome sequencing in the CPGDP based on the sampling of larger sample size in Southwest China would provide deep insights into the adaptation history of HM people.

5 CONCLUSION

Taken together, we provided genome-wide SNP data from SCM and directly evidenced their genetic affinity with the southmost Thailand and Vietnam Hmong and ancient 500-year-old Guangxi GaoHuaHua people. We identified HM-specific ancestry components spatially distributed ranged from the middle Yangtze River basin to Southeast Asia and temporally distributed at least since 500 years ago. These results provided direct evidence that supported a model in which HM-speaking populations originated from the ancient Baiyue in the middle Yangtze River basin and experienced a recent southward migration from Sichuan and Guizhou to Vietnam and Thailand. Additionally, unique patterns of naturally -selected signatures in SCM have identified many candidate genes associated with important neural system biological processes and pathways, which do not support the possibility of recent large-scale admixture occurring between HM people and surrounding Han Chinese. If these phenomena occurred, genetic changes can produce shifts in the allele frequency spectrum of pre-existing mutations and trend to show a consistent pattern of the selected signals.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants reviewed and approved by this project was inspected and approved by the Medical Ethics Committee of North Sichuan Medical College. The patients/ participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

GH, MW, H-YY, C-CW, XW, and CL conceived the idea for the study. YL, JX, MW, CL, JZ, XZ, WL, LW, CL, and QX performed or supervised the wet laboratory work. GH, JX, and MW analyzed the data. GH, JX, and MW wrote and edited the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by the Project funded by the China Postdoctoral Science Foundation (2021M691879), the Opening project of Medical Imaging Key Laboratory of Sichuan Province (MIKLSP202104), the Science and Technology Program of Guangzhou, China (2019030016), the "Double First Class University Plan" key construction project of Xiamen University (the origin and evolution of East Asian populations and the spread of Chinese civilization), the National Natural Science Foundation of China (NSFC 31801040), the Nanqiang Outstanding Young Talents Program of Xiamen University (X2123302), the Major Project of National Social Science Foundation of China (20&ZD248), and the European Research Council (ERC) grant to Dan Xu

(ERC-2019-ADG-883700-TRAM). S. Fang and Z. Xu from the Information and Network Center of Xiamen University are acknowledged for the help with the high-performance computing.

ACKNOWLEDGMENTS

We thank Wibhu Kutanan in Khon Kaen University and Mark Stoneking and Dang Liu in Max Planck Institute for Evolutionary Anthropology for sharing genome-wide SNP data from Vietnam, Thailand, and Laos.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.815160/full#supplementary-material

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast Model-Based Estimation of Ancestry in Unrelated Individuals. *Genome Res.* 19, 1655–1664. doi:10.1101/gr.094052.109

Bergström, A., Mccarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., et al. (2020). Insights into Human Genetic Variation and Population History from 929 Diverse Genomes. *Science* 367. doi:10.1126/science.aay5012

Bin, X., Wang, R., Huang, Y., Wei, R., Zhu, K., Yang, X., et al. (2021). Genomic Insight into the Population Structure and Admixture History of Tai-Kadai-Speaking Sui People in Southwest China. *Front. Genet.* 12, 735084. doi:10.3389/fgene.2021.735084

Browning, B. L., and Browning, S. R. (2013). Improving the Accuracy and Efficiency of Identity-By-Descent Detection in Population Data. *Genetics* 194, 459–471. doi:10.1534/genetics.113.150029

Cao, Y., Li, L., Li, L., Xu, M., Feng, Z., Sun, X., et al. (2020). The ChinaMAP Analytics of Deep Whole Genome Sequences in 10,588 Individuals. *Cell Res* 30, 717–731. doi:10.1038/s41422-020-0322-9

Cavalli-Sforza, L. L. (1998). The Chinese Human Genome Diversity Project. *Proc. Natl. Acad. Sci.* 95, 11501–11503. doi:10.1073/pnas.95.20.11501

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of Larger and Richer Datasets. *GigaSci* 4, 7. doi:10.1186/s13742-015-0047-8

Chen, J., He, G., Ren, Z., Wang, Q., Liu, Y., Zhang, H., et al. (2021a). Genomic Insights into the Admixture History of Mongolic- and Tungusic-Speaking Populations from Southwestern East Asia. *Front. Genet.* 12, 685285. doi:10.3389/fgene.2021.685285

Chen, J., He, G., Ren, Z., Wang, Q., Liu, Y., Zhang, H., et al. (2021b). Genomic Insights into the Admixture History of Mongolic- and Tungusic-Speaking Populations from Southwestern East Asia. *Front. Genet.* 12, 685285. doi:10.3389/fgene.2021.685285

Chen, J., Zheng, H., Bei, J.-X., Sun, L., Jia, W.-h., Li, T., et al. (2009). Genetic Structure of the Han Chinese Population Revealed by Genome-wide SNP Variation. *Am. J. Hum. Genet.* 85, 775–785. doi:10.1016/j.ajhg.2009.10.016

Chiang, C. W. K., Mangul, S., Robles, C., and Sankararaman, S. (2018). A Comprehensive Map of Genetic Variation in the World's Largest Ethnic Group-Han Chinese. *Mol. Biol. Evol.* 35, 2736–2750. doi:10.1093/molbev/msy170

Choin, J., Mendoza-Revilla, J., Arauna, L. R., Cuadros-Espinoza, S., Cassar, O., Larena, M., et al. (2021). Genomic Insights into Population History and Biological Adaptation in Oceania. *Nature* 592, 583–589. doi:10.1038/s41586-021-03236-5

Chu, J. Y., Huang, W., Kuang, S. Q., Wang, J. M., Xu, J. J., Chu, Z. T., et al. (1998). Genetic Relationship of Populations in China. *Proc. Natl. Acad. Sci.* 95, 11763–11768. doi:10.1073/pnas.95.20.11763

Consortium, H. P.-A. S., Abdulla, M. A., Ahmed, I., Assawamakin, A., Bhak, J., Brahmachari, S. K., et al. (2009). Mapping Human Genetic Diversity in Asia. *Science* 326, 1541–1545. doi:10.1126/science.1177074

Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A Linear Complexity Phasing Method for Thousands of Genomes. *Nat. Methods* 9, 179–181. doi:10.1038/nmeth.1785

Fritz, D., and Stefanovic, B. (2007). RNA-binding Protein RBMS3 Is Expressed in Activated Hepatic Stellate Cells and Liver Fibrosis and Increases Expression of Transcription Factor Prx1. *J. Mol. Biol.* 371, 585–595. doi:10.1016/j.jmb.2007.06.006

Fu, Q., Meyer, M., Gao, X., Stenzel, U., Burbano, H. A., Kelso, J., et al. (2013). DNA Analysis of an Early Modern Human from Tianyuan Cave, China. *Proc. Natl. Acad. Sci.* 110, 2223–2227. doi:10.1073/pnas.1221359110

Gautier, M., Klassmann, A., and Vitalis, R. (2017). rehh2.0: a Reimplementation of the R Packagerehhto Detect Positive Selection from Haplotype Structure. *Mol. Ecol. Resour.* 17, 78–90. doi:10.1111/1755-0998.12634

Harper, D. (2007). China's Southwest. *Lonely Planet.*

He, G. L., Li, Y. X., Wang, M. G., Zou, X., Yeh, H. Y., Yang, X. M., et al. (2021a). Fine-scale Genetic Structure of Tujia and central Han Chinese Revealing Massive Genetic Admixture under Language Borrowing. *J. Syst. Evol.* 59, 1–20. doi:10.1111/jse.12670

He, G. L., Li, Y. X., Wang, M. G., Zou, X., Yeh, H. Y., Yang, X. M., et al. (2021b). Fine-scale Genetic Structure of Tujia and central Han Chinese Revealing Massive Genetic Admixture under Language Borrowing. *J. Syst. Evol.* 59, 1–20. doi:10.1111/jse.12670

He, G., Wang, Z., Zou, X., Wang, M., Liu, J., Wang, S., et al. (2019). Tai-Kadai-speaking Gelao Population: Forensic Features, Genetic Diversity and Population Structure. *Forensic Sci. Int. Genet.* 40, e231–e239. doi:10.1016/j.fsigen.2019.03.013

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., et al. (2014). A Genetic Atlas of Human Admixture History. *Science* 343, 747–751. doi:10.1126/science.1243518

Herman, J. (2018). "Empire and Historiography in Southwest China," in *Oxford Research Encyclopedia of Asian History.* doi:10.1093/acrefore/9780190277727.013.129

Huang, X., Xia, Z.-Y., Bin, X., He, G., Guo, J., Lin, C., et al. (2020). *Genomic Insights into the Demographic History of Southern Chinese.*

Ito, M., Depaz, I., Wilce, P., Suzuki, T., Niwa, S.-i., and Matsumoto, I. (2005). Expression of Human Neuronal Protein 22, a Novel Cytoskeleton-Associated Protein, Was Decreased in the Anterior Cingulate Cortex of Schizophrenia. *Neurosci. Lett.* 378, 125–130. doi:10.1016/j.neulet.2004.12.079

Iwamura, A., Fukami, T., Higuchi, R., Nakajima, M., and Yokoi, T. (2012). Human α/β Hydrolase Domain Containing 10 (ABHD10) Is Responsible Enzyme for Deglucuronidation of Mycophenolic Acid Acyl-Glucuronide in Liver. *J. Biol. Chem.* 287, 9240–9249. doi:10.1074/jbc.m111.271288

Kutanan, W., Liu, D., Kampuansai, J., Srikummool, M., Srithawong, S., Shoocongdej, R., et al. (2021). Reconstructing the Human Genetic History of Mainland Southeast Asia: Insights from Genome-wide Data from Thailand and Laos. *Mol. Biol. Evol.* 38, 3459–3477. doi:10.1093/molbev/msab124

Larena, M., Sanchez-Quinto, F., Sjödin, P., Mckenna, J., Ebeo, C., Reyes, R., et al. (2021). Multiple Migrations to the Philippines during the Last 50,000 Years. *Proc. Natl. Acad. Sci. USA* 118, e2026132118. doi:10.1073/pnas.2026132118

Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of Population Structure Using Dense Haplotype Data. *Plos Genet.* 8, e1002453. doi:10.1371/journal.pgen.1002453

Li, L., Huang, P., Sun, X., Wang, S., Xu, M., Liu, S., et al. (2021). The ChinaMAP Reference Panel for the Accurate Genotype Imputation in Chinese Populations. *Cel Res.* doi:10.1038/s41422-021-00564-z

Li, Y., Chen, L., Nie, C.-j., Zeng, T.-t., Liu, H., Mao, X., et al. (2011). Downregulation of RBMS3 Is Associated with Poor Prognosis in Esophageal Squamous Cell Carcinoma. *Cancer Res.* 71, 6106–6115. doi:10.1158/0008-5472.can-10-4291

Lipson, M., Cheronet, O., Mallick, S., Rohland, N., Oxenham, M., Pietrusewsky, M., et al. (2018). Ancient Genomes Document Multiple Waves of Migration in Southeast Asian Prehistory. *Science* 361, 92–95. doi:10.1126/science.aat3188

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

Liu, S., Huang, S., Chen, F., Zhao, L., Yuan, Y., Francis, S. S., et al. (2018). Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. *Cell* 175, 347–359. doi:10.1016/j.cell.2018.08.016

Liu, Y., Mao, X., Krause, J., and Fu, Q. (2021a). *Insights into Human History from the First Decade of Ancient Human Genomics*.

Liu, Y., Yang, J., Li, Y., Tang, R., Yuan, D., Wang, Y., et al. (2021b). Significant East Asian Affinity of the Sichuan Hui Genomic Structure Suggests the Predominance of the Cultural Diffusion Model in the Genetic Formation Process. *Front. Genet.* 12, 626710. doi:10.3389/fgene.2021.626710

Liu, Y., Yang, J., Li, Y., Tang, R., Yuan, D., Wang, Y., et al. (2021c). Significant East Asian Affinity of the Sichuan Hui Genomic Structure Suggests the Predominance of the Cultural Diffusion Model in the Genetic Formation Process. *Front. Genet.* 12, 626710. doi:10.3389/fgene.2021.626710

Lo, Y.-H., Cheng, H.-C., Hsiung, C.-N., Yang, S.-L., Wang, H.-Y., Peng, C.-W., et al. (2021). Detecting Genetic Ancestry and Adaptation in the Taiwanese Han People. *Mol. Biol. Evol.* 38, 4149–4165. doi:10.1093/molbev/msaa276

Loh, P.-R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J. K., Reich, D., et al. (2013). Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. *Genetics* 193, 1233–1254. doi:10.1534/genetics.112.147330

Mao, X., Zhang, H., Qiao, S., Liu, Y., Chang, F., Xie, P., et al. (2021). The Deep Population History of Northern East Asia from the Late Pleistocene to the Holocene. *Cell* 184, 3256–3266. doi:10.1016/j.cell.2021.04.040

Mccoll, H., Racimo, F., Vinner, L., Demeter, F., Gakuhari, T., Moreno-Mayar, J. V., et al. (2018). The Prehistoric Peopling of Southeast Asia. *Science* 361, 88–92. doi:10.1126/science.aat3628

Ning, C., Li, T., Wang, K., Zhang, F., Li, T., Wu, X., et al. (2020). Ancient Genomes from Northern China Suggest Links between Subsistence Changes and Human Migration. *Nat. Commun.* 11, 2700. doi:10.1038/s41467-020-16557-2

Ning, C., Wang, C.-C., Gao, S., Yang, Y., Zhang, X., Wu, X., et al. (2019). Ancient Genomes Reveal Yamnaya-Related Ancestry and a Potential Source of Indo-European Speakers in Iron Age Tianshan. *Curr. Biol.* 29, 2526–2532. doi:10.1016/j.cub.2019.06.044

Oike, T., Ogiwara, H., Tominaga, Y., Ito, K., Ando, O., Tsuta, K., et al. (2013). A Synthetic Lethality-Based Strategy to Treat Cancers Harboring a Genetic Deficiency in the Chromatin Remodeling Factor BRG1. *Cancer Res.* 73, 5508–5518. doi:10.1158/0008-5472.can-12-4593

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., et al. (2012). Ancient Admixture in Human History. *Genetics* 192, 1065–1093. doi:10.1534/genetics.112.145037

Patterson, N., Price, A. L., and Reich, D. (2006). Population Structure and Eigenanalysis. *Plos Genet.* 2, e190. doi:10.1371/journal.pgen.0020190

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

Sengupta, S., Xiong, L., Fathalli, F., Benkelfat, C., Tabbane, K., Danics, Z., et al. (2006). Association Study of the Trinucleotide Repeat Polymorphism within SMARCA2 and Schizophrenia. *BMC Genet.* 7, 34. doi:10.1186/1471-2156-7-34

Siska, V., Jones, E. R., Jeon, S., Bhak, Y., Kim, H. M., Cho, Y. S., et al. (2017). Genome-wide Data from Two Early Neolithic East Asian Individuals Dating to 7700 Years Ago. *Sci. Adv.* 3, e1601877. doi:10.1126/sciadv.1601877

Taliun, D., Harris, D. N., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., et al. (2021). Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program. *Nature* 590, 290–299. doi:10.1038/s41586-021-03205-y

Tinker, N. A., and Mather, D. E. (1993). KIN: Software for Computing Kinship Coefficients. *J. Hered.* 84, 238. doi:10.1093/oxfordjournals.jhered.a111330

Van Houdt, J. K. J., Nowakowska, B. A., Sousa, S. B., Van Schaik, B. D. C., Seuntjens, E., Avonce, N., et al. (2012). Heterozygous Missense Mutations in SMARCA2 Cause Nicolaides-Baraitser Syndrome. *Nat. Genet.* 44, 445S441–449. doi:10.1038/ng.1105

Wang, C.-C., Yeh, H.-Y., Popov, A. N., Zhang, H.-Q., Matsumura, H., Sirak, K., et al. (2021a). Genomic Insights into the Formation of Human Populations in East Asia. *Nature* 591, 413–419. doi:10.1038/s41586-021-03336-2

Wang, M., He, G., Zou, X., Chen, P., Wang, Z., Tang, R., et al. (2021b). Reconstructing the Genetic Admixture History of Tai-Kadai and Sinitic People: Insights from Genome-wide Data from South China. *J. Genet. Genomics*.

Wang, M., Yuan, D., Zou, X., Wang, Z., Yeh, H.-Y., Liu, J., et al. (2021c). Fine-scale Genetic Structure and Natural Selection Signatures of Southwestern Hans Inferred from Patterns of Genome-wide Allele, Haplotype, and Haplogroup Lineages. *Front. Genet.* 12, 727821. doi:10.3389/fgene.2021.727821

Wang, Q., Zhao, J., Ren, Z., Sun, J., He, G., Guo, J., et al. (2020). Male-Dominated Migration and Massive Assimilation of Indigenous East Asians in the Formation of Muslim Hui People in Southwest China. *Front. Genet.* 11, 618614. doi:10.3389/fgene.2020.618614

Wang, Q., Dhindsa, R. S., Carss, K., Harper, A. R., Nag, A., Tachmazidou, I., et al. (2021d). Rare Variant Contribution to Human Disease in 281,104 UK Biobank Exomes. *Nature* 597, 527–532. doi:10.1038/s41586-021-03855-y

Wang, T., Wang, W., Xie, G., Li, Z., Fan, X., Yang, Q., et al. (2021e). Human Population History at the Crossroads of East and Southeast Asia since 11,000 Years Ago. *Cell* 184, 3829–3841. doi:10.1016/j.cell.2021.05.018

Wen, B., Li, H., Lu, D., Song, X., Zhang, F., He, Y., et al. (2004). Genetic Evidence Supports Demic Diffusion of Han Culture. *Nature* 431, 302–305. doi:10.1038/nature02878

Xia, Z.-Y., Yan, S., Wang, C.-C., Zheng, H.-X., Zhang, F., Liu, Y.-C., et al. (2019). *Inland-coastal Bifurcation of Southern East Asians Revealed by Hmong-Mien Genomic History*.

Xu, S., Yin, X., Li, S., Jin, W., Lou, H., Yang, L., et al. (2009). Genomic Dissection of Population Substructure of Han Chinese and its Implication in Association Studies. *Am. J. Hum. Genet.* 85, 762–774. doi:10.1016/j.ajhg.2009.10.015

Yamada, K., Iwayama, Y., Toyota, T., Ohnishi, T., Ohba, H., Maekawa, M., et al. (2012). Association Study of the KCNJ3 Gene as a Susceptibility Candidate for Schizophrenia in the Chinese Population. *Hum. Genet.* 131, 443–451. doi:10.1007/s00439-011-1089-3

Yang, M. A., Fan, X., Sun, B., Chen, C., Lang, J., Ko, Y.-C., et al. (2020). Ancient DNA Indicates Human Population Shifts and Admixture in Northern and Southern China. *Science* 369, 282–288. doi:10.1126/science.aba0909

Yao, H., Wang, M., Zou, X., Li, Y., Yang, X., Li, A., et al. (2021). New Insights into the fine-scale History of Western-Eastern Admixture of the Northwestern Chinese Population in the Hexi Corridor via Genome-wide Genetic Legacy. *Mol. Genet. Genomics* 296, 631–651. doi:10.1007/s00438-021-01767-0

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., et al. (2010). Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science* 329, 75–78. doi:10.1126/science.1190371

Yu, X., and Li, H. (2021). Origin of Ethnic Groups, Linguistic Families, and Civilizations in China Viewed from the Y Chromosome. *Mol. Genet. Genomics* 296, 783–797. doi:10.1007/s00438-021-01794-x

Zhang, H., He, G., Guo, J., Ren, Z., Zhang, H., Wang, Q., et al. (2019). Genetic Diversity, Structure and Forensic Characteristics of Hmong-Mien-speaking Miao Revealed by Autosomal Insertion/deletion Markers. *Mol. Genet. Genomics* 294, 1487–1498. doi:10.1007/s00438-019-01591-7

Zhang, Y., Lu, H., Zhang, X., Zhu, M., He, K., Yuan, H., et al. (2021). An Early Holocene Human Skull from Zhaoguo Cave, Southwestern China. *Am. J. Phys. Anthropol.* 175, 599–610. doi:10.1002/ajpa.24294

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets. *Nat. Commun.* 10, 1523. doi:10.1038/s41467-019-09234-6

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.