A GENETIC PERSPECTIVE ON ASIAN POPULATIONS

EDITED BY: Wibhu Kutanan, Piya Changmai and Chuan-Chao Wang

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88976-470-9
DOI 10.3389/978-2-88976-470-9

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

A GENETIC PERSPECTIVE ON ASIAN POPULATIONS

Topic Editors:
Wibhu Kutanan, Khon Kaen University, Thailand
Piya Changmai, University of Ostrava, Czechia
Chuan-Chao Wang, Xiamen University, China

Citation: Kutanan, W., Changmai, P., Wang, C.-C., eds. (2022). A Genetic Perspective on Asian Populations. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88976-470-9

Table of Contents

frontiers | Frontiers in Genetics

EDITORIAL
published: 08 June 2022
doi: 10.3389/fgene.2022.883843

Editorial: A Genetic Perspective on Asian Populations

Wibhu Kutanan[1]*, *Piya Changmai*[2]* and *Chuan-Chao Wang*[3,4,5]*

[1]Department of Biology, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand, [2]Department of Biology and Ecology, Faculty of Science, University of Ostrava, Ostrava, Czechia, [3]Department of Anthropology and Ethnology, School of Sociology and Anthropology, Institute of Anthropology, Xiamen University, Xiamen, China, [4]State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen, China, [5]State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen, China

Keywords: genetic ancestry, genetic markers, asia, asian populations, genetic admixture

Editorial on the Research Topic

A Genetic Perspective on Asian Populations

Asia is the world's largest and most geographically and ethnolinguistically diverse continent. Those various ethnolinguistic groups inhabit different geography and climatic extremes. Such linguistic and geographic diversifications could be correlated with the genetic variation of human populations. So far, several previous studies have reported on the genetic variation of Asian populations. Since there is an enormous diversity of modern human populations in Asia, abundant present-day populations and numerous human remains have been left for investigation. Therefore, we established this Research Topic and we initially intended to call for research articles on populations from all regions across Asia. However, we were not able to recruit research papers on populations from some parts of the continent. Nevertheless, the papers in our Research Topic still cover diverse ethnolinguistic groups of vast regions of Asia. Finally, we collected one review article and a total of 13 research papers that examined genetic variations from diverse present-day Asian populations and ancient DNA data using different kinds of genetic markers and methods.

We summarized the main results of each study according to geographic location; we began with Northern East Asia/North of China. Wei investigated the association between bi-allelic polymorphisms of two genes: *ACE* (I/D) and *ACTN3* (R/X), and some anthropometric, physical and strength traits of rowing athletes who are ethnically Han Chinese from the north of China. The alleles of these genes were genotyped by conventional PCR-based method, and allelic frequencies of the studied groups were compared between genders and between populations, specifically the Han Chinese and East Asian groups. The results showed an association between *ACE* I allele and XX genotype and male endurance traits, while *ACE* D allele is associated with female strength traits. The author also compared and discovered some differences in the frequency of *ACTN3* R/X alleles between rowers and other Chinese populations, suggesting the application of using these polymorphisms as biomarkers of genetic traits in Chinese rowing athletes. Dong et al. reported allelic frequency of 23 single nucleotide polymorphisms (SNPs) on eight skeletal muscle strength-related genes in five northern Chinese ethnic groups from Heilongjiang Province: Hezhen, Manchu, Daur, Ewenki, and Mongolian. The latter three groups showed close genetic affinity. There are statistically genetic differences between rs1815739 (*ACTN3* gene) and rs7975232 (*VDR* gene) among the five ethnic groups, and the authors suggested that natural selection could explain these results. Another association study was conducted by Cho et al.; the authors investigated the correlation between genetic signals and nine anthropometric traits, e.g. height, weight and body mass index. They analysed genome-wide data generated by KoreanChip from a huge number of Korean participants. They successfully identified nine novel genetic variants: eight

Edited by:
Marc Via,
University of Barcelona, Spain

Reviewed by:
Georgios Athanasiadis,
Sankt Hans Hospital, Denmark

***Correspondence:**
Wibhu Kutanan
wibhu@kku.ac.th
Piya Changmai
piya.changmai@osu.cz
Chuan-Chao Wang
wang@xmu.edu.cn

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

Received: *25 February 2022*
Accepted: *24 May 2022*
Published: *08 June 2022*

Citation:
Kutanan W, Changmai P and Wang C-C (2022) Editorial: A Genetic Perspective on Asian Populations. Front. Genet. 13:883843. doi: 10.3389/fgene.2022.883843

in chromosome 6 at the gene *RP11-513I15.6* and one in chromosome 12 near the gene *RP11-977G19.10* that are associated with height in Korean females. The authors also found six Asian-specific genetic variants that could be developed for further investigation on forensic anthropology.

Another genome-wide study using Illumina Array of Mongolians from Baotou city of Inner Mongolia Autonomous Region was done by (Yang et al.). Merged data from present-day East Asian populations and ancient samples across Asia were intensively analysed with multiple tests to explore Mongolian genetic architecture. In consistent with prehistorical and historical evidence that indicated complex demographic movements in Eastern Eurasia, the Mongolian showed three genetic components: one enriched in present-day Sino-Tibetan speaking groups/ancient Neolithic millet farmers of Yellow River Basin, one from ancient Neolithic hunter-gatherers of North Asia and another from Western Eurasian-related ancestral component related to the Western Steppe herders and Iranian farmers. The authors also revealed signals of recent positive selection in the *MHC* region that related to the human immune response. Zhou et al. developed methods for multiplex amplification and genotyping of 52 ancestry informative deletion/insertion polymorphisms (AIDIPs) and tested its efficiency with the Han population from Eastern China. The authors compared a subset of these AIDIPs with populations from African, European, American and South Asian and found genetic distinction of East Asian from other groups. The Eastern Han exhibited close genetic relatedness to East Asians and obtained East Asian ancestral components with high proportions. Because this AIDIPs panel had high cumulative discrimination power in Eastern Han, the authors suggested that this panel could be beneficial for additional forensic investigation apart from the commonly used method. Another study by Zhang et al. genotyped 38 X-chromosome InDel polymorphisms (X-InDels) of Han Chinese from Henan Province. This forensic investigation exhibited the effectiveness of this 38 X-InDels panel as a complementary tool for forensic applications, e.g. individual identification and parentage testing of trios, though this panel can only differentiate inter-continental populations but not intra-continental groups. The authors suggested further study could improve the power of this panel with other compound makers.

In Northwestern China, Wang et al. sequenced mitochondrial (mt) DNA in the control region using a massively parallel sequencing platform in the Altaic-speaking Kirgiz group from Northwest China. The authors compared sequence data of Kirgiz with reference populations worldwide. The Kirgiz had more maternal relatedness to East Asian populations; the mtDNA haplogroups of the Kirgiz group were ~70% of East Asian prevalent haplogroups and ~30% of haplogroups from West Asia.

The Maoming Han from Guangdong Province of Southeastern China were regarded as the descendants of Gaoliang aborigines. Maoming Han speak Cantonese, which is another branch of the Sino-Tibetan language family. Fan et al. studied the genetic structure of Maoming Han from Guangdong using forensic markers. Genotypes of 27 Y-STRs from 431 Maoming Han were done with AmpFLSTR® Yfiler® Plus PCR Amplification Kit. The authors reported allelic frequency and forensic parameters that support the effectiveness of using this marker set for forensic applications in Maoming Han. The Maoming Han exhibited closest paternal relatedness to Hakka, another Han Chinese group from Guangdong Province.

Southwestern East Asia/southwestern China is the region that harbors geographic, cultural and ethnolinguistic diversities. The Tai-Kadai and Hmong-Mien speaking populations were groups distributed in this region. Chen et al. employed the battery of Infinium Global Screening Array to generate genome-wide data of Tai-Kadai speaking-Maonan from Guizhou Province. The authors included previously published present-day and ancient East Asian and Southeast Asian populations for genetic comparison. The fine-scale genetic structure of Maonan was explored; the Guizhou Maonan had the closest genetic relationship with Guangxi Maonan. Both of them showed genetic affinity to other geographically Tai-Kadai speaking populations. Genetic admixtures with neighboring Guizhou populations were also observed. The ancestries of Tai-Kadai speaking populations were mainly related to ancient southern East Asians with an additional minor proportion from northern East Asians. Bin et al. study the fine-scale genetic structure of Tai-Kadai speaking Sui from Guangxi and Guizhou using genome-wide SNPs; there are substructures of Guizhou Sui and Guangxi Sui. The former showed relationships with other Tai-Kadai speaking groups in the vicinity, while the latter had additional genetic components from Hmong-Mien-speaking populations and Northern East Asians. In general, the Sui populations showed close genetic relatedness to various southern Chinese and Southeast Asian populations, particularly Tai-Kadai and Hmong-Mien speaking groups. The Sui also showed a genetic affinity with ancient individuals from southern China, supporting southern China as their original place.

Using the same platform as Chen et al., Liu et al. analysed genome-wide data of the Sichuan Miao group together with other Hmong-Mien speaking groups from China and Southeast Asia and also with other present-day and ancient data across East Asia. There is a new ancestral lineage that existed in the Hmong-Mien speaking groups suggesting their common origin. However, some interactions in different areas also influenced the genetic structure of some Hmong-Mien speaking groups, e. g. Dao from Vietnam, Iu Mien from Thailand and Miao and She from Chongqin. With analyses of sharing patterns of the new ancestral lineage specific to the Hmong-Mien groups and estimated admixture times, the authors suggested that Southwest China was the original place of Hmong-Mien speaking groups then recently migrated southward to Mainland Southeast Asia. Apart from the Tai-Kadai, Hmong-Mien, and Sino-Tibetan speaking groups which scattered in Southwestern China, the Mongolic-speaking Mongolians and Tungusic-speaking Manchus were also inhabited in this area. The Mongolian and Manchus moved from the North to Guizhou not older than 800 years ago. Chen et al. compared Guizhou Mongolian and Manchus with multiple present-day and ancient East Asian populations. The southwestern Mongolic and Tungusic speaking groups had mixed ancestries: one related to northern ancestor and one related to southern

indigenous East Asian. The admixture dating results are consistent with historical evidence that indicated admixture events during the Mongolians Empire expansion during the formation of the Yuan dynasty.

Southwestern China shares a border with the Tibetan Plateau to the west. With extreme environment for human occupation, e.g. high altitude, perennial low temperature, extreme aridity and severe hypoxia, geneticists have paid much attention to this area to study genetic adaptation. He et al. recruited genome-wide data generated from different platforms from previous studies and investigated the fine genetic structure of the Tibetan groups from geographically diverse regions: Ü-Tsang Tibetan, Ando Tibetan and Kham Tibetan groups compared with present-day and ancient samples. The authors found different genetic structures of three Tibetan groups which were possibly influenced by cultural and geographic factors. The Ü-Tsang Tibetans possessed a stronger Onge/Hoabinhian related affinity, Ando Tibetans harbored greater Western Eurasian related ancestry, and Kham Tibetans had much more Neolithic Southeast Asian ancestry.

There is only one article on Southeast Asia that submitted to this Research Topic. Hoh et al. reviewed the genetic studies of native groups in the Malay Peninsular and Borneo over the past century. The authors presented the contents according to chronology and types of biological and genetic markers that were generated data from primary to advance technologies: anthropological and physical traits, blood groups, protein variations, mitochondrial and autosomal DNA polymorphisms and whole-genome sequences.

In sum, albeit we were not able to recruit papers on populations from some parts of Asia, the papers in our Research Topic still cover populations of vast regions of Asia, especially East and Southeast Asia. The collection of fourteen articles in this Research Topic showed new population genetic results that were produced by heterogeneous technologies, reflecting the richness of research approaches in East/Southeast Asian genetics. Although nowadays many developments of advanced sequencing and microarray technologies and together with the innovative tools of bioinformatics have enabled researchers to deeply reconstruct population history based on modern and ancient DNA samples. We are still encouraging researchers who focused on STRs and InDels to provide a basic view of population structure, with overlapping advantages in forensic investigation. Further explorations with more present-day populations and ancient samples, particularly in Southeast Asia, will be provided new insights into the population structure, demographic history and natural selection of functional genes that could be useful for biomedical studies and forensic investigation in Asian populations.

AUTHOR CONTRIBUTIONS

WK, PC, and C-CW co-wrote this editorial based on this Research Topic's contributions. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We thank all authors who submitted their works for this Research Topic and thank other editors and reviewers for their valuable contributions to this Research Topic.

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 13 May 2021
doi: 10.3389/fgene.2021.669215

A Genome-Wide Association Study of Novel Genetic Variants Associated With Anthropometric Traits in Koreans

Hye-Won Cho[1†], Hyun-Seok Jin[2†] and Yong-Bin Eom[1,3]*

[1] Department of Medical Sciences, Graduate School, Soonchunhyang University, Asan-si, South Korea, [2] Department of Biomedical Laboratory Science, College of Life and Health Sciences, Hoseo University, Asan-si, South Korea, [3] Department of Biomedical Laboratory Science, College of Medical Sciences, Soonchunhyang University, Asan-si, South Korea

Most previous genome-wide association studies (GWAS) have identified genetic variants associated with anthropometric traits. However, most of the evidence were reported in European populations. Anthropometric traits such as height and body fat distribution are significantly affected by gender and genetic factors. Here we performed GWAS involving 64,193 Koreans to identify the genetic factors associated with anthropometric phenotypes including height, weight, body mass index, waist circumference, hip circumference, and waist-to-hip ratio. We found nine novel single-nucleotide polymorphisms (SNPs) and 59 independent genetic signals in genomic regions that were reported previously. Of the 19 SNPs reported previously, eight genetic variants at *RP11-513I15.6* and one genetic variant at the *RP11-977G19.10* region and six Asian-specific genetic variants were newly found. We compared our findings with those of previous studies in other populations. Five overlapping genetic regions (*PAN2*, *ANKRD52*, *RNF41*, *HGMA1*, and *C6orf106*) had been reported previously but none of the SNPs were independently identified in the current study. Seven of the nine newly found novel loci associated with height in women revealed a statistically significant skeletal expression of quantitative trait loci. Our study provides additional insight into the genetic effects of anthropometric phenotypes in East Asians.

Keywords: genome-wide association study, Korean, East Asian, *CUX2*, rs7133285, eQTL, anthropometric traits

OPEN ACCESS

Edited by:
Wibhu Kutanan,
Khon Kaen University, Thailand

Reviewed by:
Gyaneshwer Chaubey,
Banaras Hindu University, India
Chuan-Chao Wang,
Xiamen University, China

***Correspondence:**
Yong-Bin Eom
omnibin@sch.ac.kr

[†] These authors have contributed equally to this work

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

Received: *18 February 2021*
Accepted: *31 March 2021*
Published: *13 May 2021*

Citation:
Cho H-W, Jin H-S and Eom Y-B (2021) A Genome-Wide Association Study of Novel Genetic Variants Associated With Anthropometric Traits in Koreans. Front. Genet. 12:669215. doi: 10.3389/fgene.2021.669215

INTRODUCTION

Human anthropometric traits, including height, body mass index (BMI), and fat distribution, differ substantially according to gender and genetic factors. Particularly, height, which has been associated with multiple diseases, is highly representative of the heritable phenotypic trait (Akiyama et al., 2019). Anthropometric traits related to obesity such as body size and composition are highly associated with metabolic syndrome (Wen et al., 2016). Furthermore, anthropometry is occasionally considered the traditional and basic tool of biological anthropology. However, it also plays an essential role in forensic science (Krishan, 2006; Thamizhselvi and Geetha, 2019).

A high degree of ethnic differences in adult anthropometric traits has been reported (Marigorta and Navarro, 2013). Many previous large-scale genome-wide association studies (GWAS) conducted in European populations revealed genetic variants correlated with various anthropometric traits. However, there is limited evidence supporting the variations associated with anthropometric traits in East Asian populations (Tachmazidou et al., 2017; Rask-Andersen et al., 2019). Therefore, given that various populations carry specific genetic variants, identification of genetic variants in large East Asian population samples is important in understanding the genetic determinants associated with anthropometric traits.

To date, more than 106 loci have been associated with anthropometric traits such as height and fat distribution (Tachmazidou et al., 2017). Tachmazidou et al. (2017) performed a GWAS of 12 anthropometric traits correlated with height and body mass in European populations and discovered six novel loci related to height and hip circumference (*CCDC36*, *HCG18*, *ZNF143*, *RP11-63E9.1*, *DDX51*, and *RP11-788M5.4*) and 28 independent genetic variants in previously reported genes. Randall et al. (2013) identified seven signals that were significant in women (located near *GRB14/COBLL1*, *LYPLAL1/SLC30A20*, *VEGFA*, *ADAMTS9*, *MAP3K1*, *HSD17B4*, and *PPARG*). More recently, Rask-Andersen et al. (2019) confirmed independent genetic signals related to the adiposity phenotype in the UK Biobank and identified clear differences between males and females, especially regarding fat distribution in the legs and trunk. Akiyama et al. (2019) performed a large-scale genetic association study and characterized 22 rare and 42 low-frequency height-associated single-nucleotide polymorphisms (SNPs) in a Japanese population.

Therefore, to explore the specific genetic signals of anthropometric traits in a Korean population, we performed a study using KoreanChip (KCHIP, Seoul, South Korea). The present study investigated the genetic factors related to the anthropometric phenotypes of 64,193 participants from two independent Korean cohorts. Here we performed GWAS for each gender to identify novel genetic variants reaching a genome-wide significance threshold (p-value $< 1 \times 10^{-8}$). We found not only nine novel SNPs that had not been reported previously, which were associated with height in females, but also 59 independent genetic signals in genomic regions that had been reported previously.

MATERIALS AND METHODS

Study Design and Participants

The present study included two independent cohorts at the discovery and replication stages. The participants in the discovery stage (phase 1) were recruited from the Ansan/Ansung cohorts of the Korean Genome and Epidemiology Study (KoGES) between 2001 and 2002 (Kim et al., 2017), known as the Korea Association REsource (De Vries et al., 2019) project. The study involved 10,038 participants and included the genetic data of 5,493 participants (2,616 men and 2,877 women; age, 40–69 years). The participants in the replication stage (phase 2) were selected from the Health Examinee (HEXA) study cohort of the KoGES, which included a total of 173,357 participants recruited between 2004 and 2013 (Kim et al., 2017). This study included participants from urban (Seoul, Incheon, Daejeon, Daegu, Ulsan, Busan, and Gwangju) and rural (Gyeonggi, Sejong, Gangwon, Chungcheongbuk, Chungcheongnam, Gyeongsangbuk, Gyeongsangnam, Jeollabuk, Jeollanam, and Jeju) areas, and all participants were between 40 and 79 years of age. Among a total of 173,357 participants with baseline data in the HEXA study, only 58,700 participants were selected for the replication analysis. Height (cm), weight (kg), waist circumference (cm), and hip circumference (cm) were examined, and the BMI [weight (kg)/height (m^2)] and waist-to-hip ratio (WHR = waist/hip) were computed. To reflect body fat distribution independent of overall adiposity, waist circumference, hip circumference, and the WHR were also analyzed and adjusted for BMI (waistBMIadj, hipBMIadj, and whrBMIadj). Since the anthropometric traits differed by gender in various aspects, males and females were analyzed separately (Randall et al., 2013). Participants with values greater than three standard deviations (SD) (depending upon cohort, sex, and trait) were also excluded from the study.

Genotyping and Quality Control

Genotype data were provided by the Center for Genome Science, Korea National Institute of Health. DNA samples were separated and extracted from the peripheral blood of the participants. DNA genotyping of both the discovery and replication GWAS populations was performed using the Korea Biobank Array, which was designed by the Center for Genome Science, Korea National Institute of Health, South Korea, and referred to as the KoreanChip (KCHIP; Seoul, South Korea). The KCHIP array included a total of 833,535 single nucleotide variants for autosomal chromosomes (Han et al., 2021). The location of the genes was assigned according to the National Center for Biotechnology Information Human Genome Build 37 (hg19). The detailed KoreanChip analysis was reported previously (Moon et al., 2019; Jin et al., 2020). We excluded samples matching one of the following criteria: (i) genotyping accuracy less than 96–99% (Heid et al., 2010), (ii) excessive heterozygosity, and (iii) sex inconsistencies. SNPs were removed with (i) a missing call rate >5% (Heid et al., 2010), (ii) a minor allele frequency <1%, and (iii) a p-value in the Hardy–Weinberg equilibrium test $<10^{-4}$. A total of 465 K variants were included after the quality control. After quality control and imputation, a total of 8,056,211 SNPs were used for this GWAS.

Statistical Analysis

Most statistical analyses were performed using PLINK, version 1.90 beta[1] (Purcell et al., 2007). Imputation of the genotype data was executed using IMPUTE v2 with data from the 1,000 genome phase 3 haplotypes serving as the reference panel (Chung et al., 2020; Oh et al., 2020). Only SNPs with an r^2 value $\leq$ 95% with no linkage disequilibrium to each other were included in our study. GWAS were performed to identify SNPs associated

[1]https://www.cog-genomics.org/plink2

with anthropometric traits *via* linear regression analysis with an additive model. Age and area were fitted as fixed covariates, and BMI was added to the adjustment as described above. The cutoff *p*-value suggesting the genome-wide significance level was $P < 10^{-5}$ in the discovery stage (phase 1) and $P < 10^{-8}$ in the replication and combined (discovery + replication) stages. GTEx Portal databases[2] were used for expression quantitative trait loci (eQTL) analysis (GTEx Consortium, 2015), Haploview[3] was used for Manhattan plots, and a regional plot was generated using LocusZoom[4]. Functional annotations such as protein motifs were analyzed using HaploReg[5], and functional variants were identified by RegulomeDB[6].

Ethical Review

This study was approved by the Institutional Review Board of the Korea National Institute of Health (KBN-2021-003) and Soonchunhyang University (202012-BR-086-01). Written informed consent was obtained from all participants.

RESULTS

Identification of Loci Related to Anthropometric Traits in the Discovery Stage

The participants' characteristics in the discovery and replication stages are listed in **Table 1**. We performed GWAS of 5,493 participants (2,616 men and 2,877 women) in the discovery stage and selected SNPs reaching the signal cutoffs for association at $P < 10^{-5}$ and $P < 10^{-8}$ in the discovery and replication stages, respectively. Manhattan plot showed genome-wide association between height and women in the discovery phase (**Supplementary Figure 1**). Nine novel SNPs, which were associated with height in women, located on autosomal chromosomes 6 and 12 at the genes *RP11-513I15.6* and *RP11-977G19.10* were identified, and eight genetic variants were found in *RP11-513I15.6* (**Table 2**). In addition, 59 significant independent signals at previously reported regions and 19 signals identified in previous studies for different traits were found. Among the nine anthropometric traits, two male-related traits, including weight and WHR, and one female-related trait (height) reached a *p*-value $< 10^{-5}$. The *HGMA1*, *C6orf106*, and *GRM4* genes reached the significance level for weight in men, and the *CUX2* gene reached significance for WHR in men (**Supplementary Table 1**). The genetic locations of these nine novel SNPs and their recombination rates in the discovery stage are plotted by their position in **Figure 1**. Moreover, the *LINC02456*, *GRM4*, *HMGA1*, *PAN2*, *SMIM29*, *C6orf106*, *ANKRD52*, *RNF41*, and *SLC39A5* loci reached the significance level for height in women. *HMGA1*, *C6orf106*, and *GRM4* were associated with anthropometric traits in both sexes. In particular, five genetic loci including rs1187115 ($P_{men} = 1.73 \times 10^{-6}$, $P_{women} = 5.32 \times 10^{-7}$), rs10807137 ($P_{men} = 5.44 \times 10^{-6}$, $P_{women} = 6.88 \times 10^{-7}$), rs370788671 ($P_{men} = 6.85 \times 10^{-6}$, $P_{women} = 2.17 \times 10^{-6}$), and rs9469745 ($P_{men} = 9.74 \times 10^{-6}$, $P_{women} = 8.09 \times 10^{-6}$) at *HMGA1* and rs6457765 ($P_{men} = 5.53 \times 10^{-6}$, $P_{women} = 1.22 \times 10^{-6}$) at *C6orf106* were identified as common anthropometric-related genetic variants in men and women.

Genetic Variants Showing Association Signals in the Replication and Combination Stages

Single-nucleotide polymorphisms that reached a *p*-value $< 10^{-5}$ in the discovery set were selected and re-evaluated in other stages. We performed a replication analysis of genetic variants found in the discovery stage, and finally, a total of 94 SNPs were analyzed in the combination stage. A total of 58,700 participants (20,293 men and 38,407 women) were included in the replication stage, and 64,193 participants (22,909 men and 41,284 women) were included in the combination stage. The genetic variants associated with waist circumference and BMI in the discovery stage did not meet the cutoff for signal association at $P < 10^{-5}$ and $P < 10^{-8}$ in both the discovery and

[2]https://gtexportal.org
[3]https://www.broadinstitute.org/
[4]http://locuszoom.org/
[5]https://pubs.broadinstitute.org
[6]https://regulomedb.org/

TABLE 1 | Characteristics of the study participants.

Characteristics	Men			Women		
	Discovery	Replication	Combination	Discovery	Replication	Combination
No.	2,616	20,293	22,909	2,877	38,407	41,284
Age (M years ± SD)	51.08 ± 8.33	55.18 ± 8.42	54.71 ± 8.51	51.98 ± 8.64	53.07 ± 7.70	53.00 ± 7.77
Height (M cm ± SD)	68.03 ± 9.46	168.69 ± 5.61	168.50 ± 5.65	153.98 ± 5.43	156.53 ± 5.16	156.36 ± 5.20
Weight (M kg ± SD)	167.00 ± 5.68	69.37 ± 8.72	69.22 ± 8.82	58.78 ± 7.94	57.62 ± 7.25	57.70 ± 7.31
BMI (M kg/m^2 ± SD)	24.36 ± 2.84	24.37 ± 2.60	24.37 ± 2.63	24.78 ± 3.03	23.53 ± 2.78	23.62 ± 2.81
Hip circumference (M cm ± SD)	93.90 ± 5.59	95.71 ± 5.35	95.51 ± 7.28	93.38 ± 5.66	93.21 ± 5.36	93.22 ± 5.38
Waist circumference (M cm ± SD)	83.68 ± 7.45	85.62 ± 7.23	85.40 ± 7.28	81.40 ± 9.26	78.08 ± 7.87	78.31 ± 8.02
WHR	0.89 ± 0.06	0.89 ± 0.05	0.89 ± 0.05	0.87 ± 0.09	0.84 ± 0.06	0.84 ± 0.06

BMI, body mass index; WHR, waist–hip ratio; M, mean value; SD, standard deviation.

TABLE 2 | Identified novel variants with genome-wide significant associations.

SNP	Nearest gene	Trait	Chromosome position	Minor allele	MAF	Discovery (*n* = 2,869)		Replication (*n* = 37,321)		Combination (*n* = 40,169)	
						β ± SE	*P* value	β ± SE	*P* value	β ± SE	*P* value
rs1307273	*RP11-513I15.6*	Height–women	6:34240886	G	0.205	0.82 ± 0.17	9.27×10^{-7}	0.33 ± 0.04	5.19×10^{-14}	0.37 ± 0.04	3.37×10^{-17}
rs9469757	*RP11-513I15.6*	Height–women	6:34232864	C	0.170	0.87 ± 0.18	1.16×10^{-6}	0.37 ± 0.05	1.41×10^{-14}	0.40 ± 0.05	3.10×10^{-17}
rs1759637	*RP11-513I15.6*	Height–women	6:34237130	C	0.193	0.82 ± 0.17	1.62×10^{-6}	0.32 ± 0.05	4.49×10^{-12}	0.35 ± 0.04	8.54×10^{-15}
rs7133285	***RP11-977G19.10***	**Height–women**	**12:56699429**	**A**	**0.232**	**−0.75 ± 0.16**	$\mathbf{2.25 \times 10^{-6}}$	**−0.67 ± 0.04**	$\mathbf{2.27 \times 10^{-54}}$	**−0.67 ± 0.04**	$\mathbf{4.19 \times 10^{-58}}$
rs9469761	*RP11-513I15.6*	Height–women	6:34236429	A	0.169	0.85 ± 0.18	2.53×10^{-6}	0.37 ± 0.05	1.38×10^{-14}	0.40 ± 0.05	3.25×10^{-17}
rs1776890	*RP11-513I15.6*	Height–women	6:34237747	G	0.194	0.79 ± 0.17	3.62×10^{-6}	0.32 ± 0.05	9.86×10^{-13}	0.35 ± 0.04	1.51×10^{-15}
rs13207853	*RP11-513I15.6*	Height–women	6:34238946	C	0.181	0.81 ± 0.18	4.64×10^{-6}	0.33 ± 0.05	1.57×10^{-12}	0.37 ± 0.05	1.59×10^{-15}
rs2797961	*RP11-513I15.6*	Height–women	6:34237797	G	0.205	0.77 ± 0.17	4.67×10^{-6}	0.32 ± 0.04	6.31×10^{-13}	0.35 ± 0.04	9.07×10^{-16}
rs200808496	*RP11-513I15.6*	Height–women	6:34237376	T	0.203	0.75 ± 0.17	7.40×10^{-6}	0.31 ± 0.04	9.98×10^{-12}	0.33 ± 0.04	5.34×10^{-14}

Age and residential area were included as covariants in all genetic models. Variants stated in the manuscript are indicated in bold. MAF, minor allele frequency; β, regression coefficient; SE, standard error; N, the number of participants included in each stage.

replication stages. Moreover, a higher number of genetic signals were related to anthropometric traits in women than in men, and the statistical significance between SNPs and the traits was also higher in women. Among the nine novel SNPs related to height in women, rs7133285 in the *RP11-977G19.10* region had an even higher significance in the replication and combination stages compared with the discovery stage ($P_{rep} = 2.27 \times 10^{-54}$, $P_{com} = 4.19 \times 10^{-58}$) (**Table 2**). A similar trend was detected with the genetic loci rs146426492 ($P_{rep} = 2.70 \times 10^{-53}$, $P_{com} = 4.07 \times 10^{-57}$), rs72648137 ($P_{rep} = 9.28 \times 10^{-48}$, $P_{com} = 7.81 \times 10^{-52}$), rs76280383 ($P_{rep} = 1.24 \times 10^{-29}$, $P_{com} = 2.00 \times 10^{-33}$), and rs76459740 ($P_{rep} = 2.23 \times 10^{-29}$, $P_{com} = 2.64 \times 10^{-33}$), which were located on chromosome 12 and included in the 59 significant independent signals at previously reported regions and were definitely associated with height in women in the replication and combination stages.

Functional Annotations of Associated Variants

Based on the GTEx databases, we analyzed the eQTL of the novel SNPs and new SNPs in previously reported regions. Among the nine newly found loci associated with height in women, statistically significant skeletal muscle eQTL were found for seven genetic variants (rs1307273, rs1759637, rs9469761, rs1776890, rs13207853, rs2797961, and rs200808496) (**Figure 2A**). In particular, rs1776890 and rs200808496, belonging to the skeleton eQTL, were associated with the expression of not only *NUDT3* but also *C6orf106* and *RPS10*. Additionally, we identified these seven height-related genetic variants as eQTL in subcutaneous adipose tissue that affects body fat distribution. Among the independent loci belonging to previously reported regions (*C6orf106*), rs6457765, rs147736074, and rs13201774 were eQTL expressed in the skeletal muscle ($P = 6.80 \times 10^{-6}$–9.50×10^{-5}, eQTL effect size = −0.24 – −0.11) (**Figure 2B**, **Supplementary Table 2**, and **Supplementary Figure 2**).

Geographic Distribution of Genomic Variants

We analyzed the frequency of rs7133285, rs146426492, rs72648137, rs76280383, and rs76459740 that showed a higher significance in the replication and combination stages than in the discovery stage. The GGV browser identifies the frequencies of the genetic variants in diverse populations based on 1,000 genomes (hg19) (**Figure 3**). The minor allele frequency of the rs7133285, which was distributed throughout the world, was higher in Africa and East Asia than in Europe or America (**Figure 3A**). Four variants, including rs146426492, rs72648137, rs76280383, and rs76459740, showed almost similar patterns, representing Asian genetic variants specific to the Asian population (**Figures 3B–F**). The minor allele frequencies in the cohort used in the present study are shown in **Table 2** and **Supplementary Table 1**.

DISCUSSION

In the present study, we performed GWAS of anthropometric traits in a non-European population using KCHIP optimized for the Korean population (Moon et al., 2019) and presented the GWAS findings of anthropometric traits. Many studies related to anthropometric traits have been performed, and recent studies investigated Asian populations. However, few GWAS of anthropometric traits involving Korean population have been performed. Cho et al. (2009) performed a large-scale

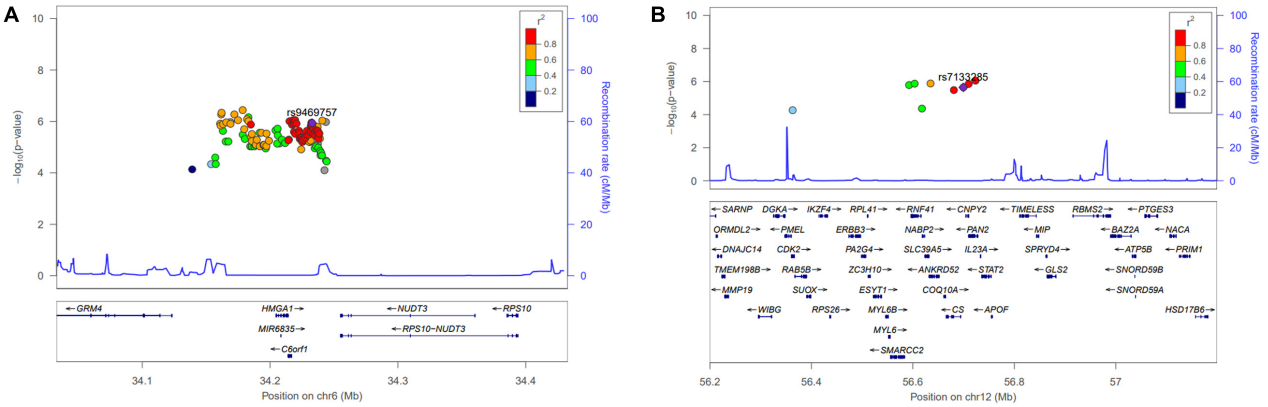

FIGURE 1 | Regional association plot for the newly identified single-nucleotide polymorphisms (SNPs). Height-related signals in the discovery genome-wide association studies are plotted as $-\log_{10}P$ values. The color of each SNP plot shows its linkage disequilibrium (using r^2 values) with the novel SNP (purple diamond) within the association locus. The *y*-axis on the right shows the recombination rate from the HapMap database. The image above was constructed using the LocusZoom program (http://locuszoom.org/). **(A)** Regional plot of rs9469757 ($P = 1.16 \times 10^{-6}$) around the *RP11-513I15.6* region of chromosome 6 with height in women. **(B)** Regional plot of rs7133285 ($P = 2.25 \times 10^{-6}$) around the *RP11-977G19.10* region of chromosome 12.

genetic association analysis of Koreans and found genetic factors related to eight quantitative traits including height, BMI, WHR, blood pressure, pulse rate, and bone density. The analysis was performed with Affymetrix5.0, which is a commercial SNP array designed for European or multiethnic populations (Cho et al., 2009; Kim et al., 2011). However, limitations in the capture of functional signals from next-generation sequencing were recently detected, including monomorphic variants in the Korean population. Therefore, we performed GWAS with KCHIP and identified nine novel genetic variants, 59 independent genetic signals in genomic regions that were reported previously, and 19 previously reported signals.

Overall, our results indicated similarities in the genetic signals associated with height in East Asians and Europeans. Among the 59 independent genetic signals in genomic regions that were reported previously (**Supplementary Table 1**), *C6orf106* (also known as *ILRUN*) (Weedon et al., 2008; Berndt et al., 2013; Tachmazidou et al., 2017; Kichaev et al., 2019), *GRM4* (Kichaev et al., 2019), *PAN2* (Allen et al., 2010), *SMIM29* (N'Diaye et al., 2011; Wojcik et al., 2019), and *ANKRD52* (Kichaev et al., 2019) were related to height in European populations. Besides this, *C6orf106*, which was related to weight in men in our study, also showed an association with lipid levels [high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, and apolipoprotein A1 and B], metabolic syndrome, WHR-adjusted BMI, waist circumference, and BMI in diverse populations (De Vries et al., 2019; Richardson et al., 2020; Zhu et al., 2020). Except for the newly identified loci, 19 SNPs reached the cutoff for signal associations at $P < 10^{-5}$ and $P < 10^{-8}$ in the discovery and replication stages, respectively, that were previously reported for different or similar traits in other populations (**Supplementary Table 3**). In particular, rs671 belonging to *ALDH2* is a well-known variant related to metabolic syndrome and body mass in Asians (Wen et al., 2014; Zhu et al., 2017), and recently, Akiyama et al. (2019) reported that rs671 was associated with height in a Japanese population.

Analysis of a non-European population such as Korean for specific height-related signals including eight variants at the *RP11-513I15.6* region and one variant at the *RP11-977G19.10* region, which were not previously reported, may have an advantage. The comparison of a previous study of genetic variants influencing anthropometric traits in Koreans, using Affymetrix5.0, with the present study (Cho et al., 2009; Kim et al., 2010) showed that only the *HMGA1* region was associated with height in both studies. SNPs with higher statistical significance were found in the replication and combination stages compared with the discovery stage. Above all, the *PAN2* gene was found in an East Asian meta-analysis of GWAS, which interestingly revealed the highest statistical significance for adult height (He et al., 2015). Indeed, in a European GWA study, the *ANKRD52* gene, which was found to be associated with height and BMI (Kichaev et al., 2019; Zhu et al., 2020), showed significance in the replication and combined stages of the present study (**Supplementary Table 1**), and rs72648137 belonging to the *ANKRD52* gene in the present study was an Asian-specific genetic signal (**Figure 3C**). The genetic variant (rs76280383) belonging to the *RNF41* gene, which is known to be associated with congenital heart diseases in the Chinese Mongolian population (Zhang et al., 2016), was also an Asian-specific mutation (**Figure 3D**). Cho et al. (2009) stated that the genetic signals of anthropometric traits, especially BMI and height, overlapped with findings previously reported in a European population. Accordingly, the results of our study offer insights into the similarities and differences based on genetic factors associated with height and underscore the need for analysis of various populations to broaden our understanding of the genetic basis of anthropometric traits.

There were 16 independent signals related to weight or WHR of men at previously reported regions, which were located in four regions including *HGMA1*, *C6orf106*, *GRM4*, and *CUX2* (**Supplementary Table 1**). One study reported that the genetic variants in the *HGMA1* region were associated with an increased risk of type 2 diabetes (Bianco et al., 2015). Furthermore, *HGMA1*

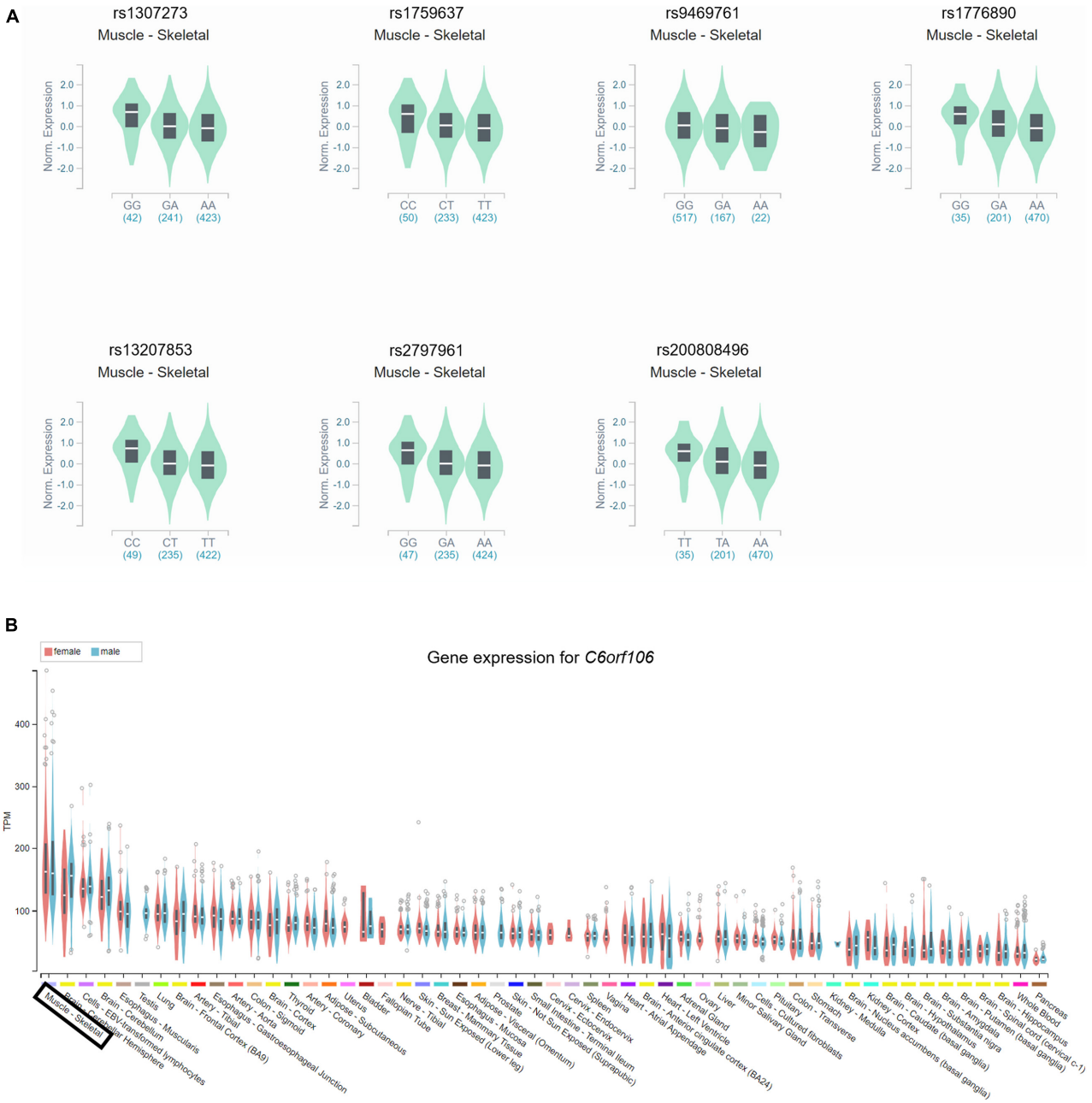

FIGURE 2 | Identification of the gene expression of the *C6orf106* gene and seven novel SNPs in eQTL. The gene expression of each genotype in the skeletal muscle was presented using GTEx Portal and showed statistical significance. **(A)** Expression of each genotype of rs1307273, rs1759637, rs9469761, rs1776890, rs13207853, rs2797961, and rs200808496 in the RP11-513I15.6 region. All of the seven variants showed statistical significance reaching $P < 10^{-4}$. **(B)** Gene expression of the *C6orf106* gene in the skeletal muscle.

and *C6orf106* have been reported to show a genetic association with body fat ratios in Caucasians (Rask-Andersen et al., 2019), and both *GRM4* and *CUX2* were associated with the γ-GT-catalyzed reaction in excessive alcohol consumption (Chen et al., 2020). Additionally, the *CUX2* gene has been investigated in genome association studies related to serum uric acid levels, coronary artery disease, and hypertension (Cho et al., 2020; German et al., 2020; Matsunaga et al., 2020). However, to the best of our knowledge, the association with anthropometric trait (especially the WHR) GWAS signals has never been reported previously. Our GWAS demonstrated that the *CUX2* gene was associated with the WHR in Asian men ($P_{dis} = 1.44 \times 10^{-6}$, $P_{rep} = 2.49 \times 10^{-15}$, $P_{com} = 4.63 \times 10^{-19}$) and revealed the minor allele of rs76892715 in the *CUX2* gene in the East Asian population, but not in European, American, and African populations (**Figure 3F**).

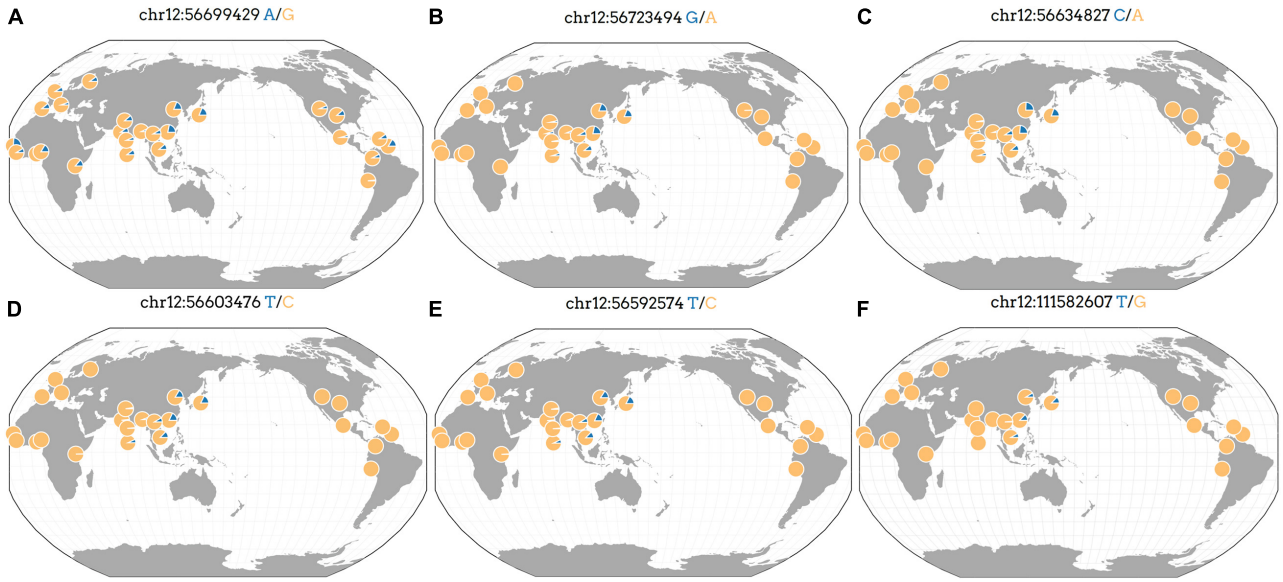

FIGURE 3 | Geographic distribution of Asian-specific variants identified as independent signals in previously reported genes. The position of the genetic variant is described at the top, and the blue part indicates an allele frequency based on 1,000 genomes (hg19). **(A)** rs7133285, **(B)** rs146426492, **(C)** rs72648137, **(D)** rs76280383, **(E)** rs76459740, and **(F)** rs76892715.

In this study, there were significant genetic signals associated with weight and the WHR in males, but no new SNPs related to female fat distribution traits were found. However, previous studies, which analyzed the distribution of body fat and reported contradictory results, indicated that the genetic influence affecting fat distribution was more powerful in females than in males (Rask-Andersen et al., 2019). Other studies also revealed sexual dimorphism in the genetic effects related to fat distribution-related traits (Heid et al., 2010; Randall et al., 2013; Winkler et al., 2015). As shown in our results, rs76892715 in the *CUX2* region was associated with the WHR, a trait representing abdominal obesity in men (**Supplementary Table 1**). The *CUX2* gene is expressed higher in men than in women in general (**Supplementary Figure 3**), and males may be influenced more by the direct genetic association with *CUX2* than females. Further studies are required to elucidate these genetic signals between fat distribution and gender. Another limitation of our study was the lack of cohorts comprised of diverse populations and genotyping with the related assay chips, suggesting the need for further studies in the future.

In conclusion, we analyzed nine anthropometric traits and found nine novel genetic signals that had not been previously reported and 59 genetic independent variants in genomic regions that had been reported previously. Our study discovered novel loci in two regions including *RP11-513I15.6* and *RP11-977G19.10* associated with height in Korean women. Of the genetic loci previously associated with quantitative traits in non-Asian populations, 19 similar genetic variants that reached the cutoff for signal association were presented. Six Asian-specific genetic variants were also found, suggesting that both Asian and European populations show not only overlapping genetic signals but also characteristic anthropometric traits. Thus, anthropometric trait GWAS may enrich our perspective of anthropometric traits in East Asians, and optimization of ethnicity-specific genetic variants to distinguish nationality may contribute to the foundation of forensic anthropology.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board of the Korea National Institute of Health (KNIH and KBN-2021-003) and Soonchunhyang University (202012-BR-086-01). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Y-BE and H-SJ participated in the design of the study, contributed to data reduction/analysis, and interpretation of the results. H-WC contributed to data analysis and interpretation of the results. All authors contributed to manuscript writing, reviewed and approved the final version of the manuscript, and agreed with the order of presentation of the authors.

FUNDING

This study was supported by the Soonchunhyang University Research Fund and a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2020R1F1A1071977).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.669215/full#supplementary-material

Supplementary Figure 1 | Manhattan plot showing a genome-wide association between height and women in the discovery stage. The red line suggests that the genome-wide significance level was of p-value $< 10^{-5}$ in the discovery stage.

Supplementary Figure 2 | Expression of each genotype of **(A)** rs6457765, **(B)** rs147736074, and **(C)** rs13201774 in the *C6orf106* region of the skeletal muscle.

Supplementary Figure 3 | Gender-based gene expression of *CUX2*.

Supplementary Table 1 | Identified independent signals at a previously reported gene with genome-wide significant associations.

Supplementary Table 2 | Significant single tissue eQTL results for nine novel genetic variants and signals belonging to C6orf106 using the GTEx portal.

Supplementary Table 3 | Replication of 19 previously discovered loci in our study populations.

REFERENCES

Akiyama, M., Ishigaki, K., Sakaue, S., Momozawa, Y., Horikoshi, M., Hirata, M., et al. (2019). Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nat. Commun.* 10:4393. doi: 10.1038/s41467-020-15202-2

Allen, H. L., Estrada, K., Lettre, G., Berndt, S., Weedon, M. N., and Rivadeneira, F. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832–838. doi: 10.1038/nature09410

Berndt, S. I., Gustafsson, S., Mägi, R., Ganna, A., Wheeler, E., Feitosa, M. F., et al. (2013). Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat. Genet.* 45, 501–512. doi: 10.1038/ng.2606

Bianco, A., Chiefari, E., Nobile, C. G., Foti, D., Pavia, M., and Brunetti, A. (2015). The association between HMGA1 rs146052672 variant and type 2 diabetes: a transethnic meta-analysis. *PLoS One* 10:e0136077. doi: 10.1371/journal.pone.0136077

Chen, I. C., Kuo, P. H., Yang, A. C., Tsai, S. J., Liu, T. H., Liu, H. J., et al. (2020). CUX2, BRAP and ALDH2 are associated with metabolic traits in people with excessive alcohol consumption. *Sci. Rep.* 10:18118. doi: 10.1038/s41598-020-75199-y

Cho, S. K., Kim, B., Myung, W., Chang, Y., Ryu, S., Kim, H. N., et al. (2020). Polygenic analysis of the effect of common and low-frequency genetic variants on serum uric acid levels in Korean individuals. *Sci. Rep.* 10:9179. doi: 10.1038/s41598-020-66064-z

Cho, Y. S., Go, M. J., Kim, Y. J., Heo, J. Y., Oh, J. H., Ban, H.-J., et al. (2009). A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat. Genet.* 41, 527–534. doi: 10.1038/ng.357

Chung, G. E., Shin, E., Kwak, M.-S., Yang, J. I., Lee, J.-E., Choe, E. K., et al. (2020). The association of genetic polymorphisms with nonalcoholic fatty liver disease in a longitudinal study. *BMC Gastroenterol.* 20:344. doi: 10.1186/s12876-020-01469-8

De Vries, P. S., Brown, M. R., Bentley, A. R., Sung, Y. J., Winkler, T. W., Ntalla, I., et al. (2019). Multiancestry genome-wide association study of lipid levels incorporating gene-alcohol interactions. *Am. J. Epidemiol.* 188, 1033–1054. doi: 10.1093/aje/kwz005

German, C. A., Sinsheimer, J. S., Klimentidis, Y. C., Zhou, H., and Zhou, J. J. (2020). Ordered multinomial regression for genetic association analysis of ordinal phenotypes at Biobank scale. *Genet. Epidemiol.* 44, 248–260. doi: 10.1002/gepi.22276

GTEx Consortium. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110

Han, N., Oh, J. M., and Kim, I.-W. (2021). Combination of genome-wide polymorphisms and copy number variations of pharmacogenes in Koreans. *J. Pers. Med.* 11:33. doi: 10.3390/jpm11010033

He, M., Xu, M., Zhang, B., Liang, J., Chen, P., Lee, J.-Y., et al. (2015). Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. *Hum. Mol. Genet.* 24, 1791–1800. doi: 10.1093/hmg/ddu583

Heid, I. M., Jackson, A. U., Randall, J. C., Winkler, T. W., Qi, L., Steinthorsdottir, V., et al. (2010). Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat. Genet.* 42, 949–960. doi: 10.1038/ng.685

Jin, E. H., Park, B., Kim, Y. S., Choe, E. K., Choi, S. H., Kim, J. S., et al. (2020). A novel susceptibility locus near GRIK2 associated with erosive esophagitis in a Korean Cohort. *Clin. Transl. Gastroenterol.* 11:e00145. doi: 10.14309/ctg.0000000000000145

Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M. K., et al. (2019). Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* 104, 65–75. doi: 10.1016/j.ajhg.2018.11.008

Kim, J. J., Lee, H. I., Park, T., Kim, K., Lee, J. E., Cho, N. H., et al. (2010). Identification of 15 loci influencing height in a Korean population. *J. Hum. Genet.* 55, 27–31. doi: 10.1038/jhg.2009.116

Kim, Y., Han, B. G., and Group, K. (2017). Cohort profile: the Korean genome and epidemiology study (KoGES) consortium. *Int. J. Epidemiol.* 46:e20. doi: 10.1093/ije/dyx105

Kim, Y. J., Go, M. J., Hu, C., Hong, C. B., Kim, Y. K., Lee, J. Y., et al. (2011). Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. *Nat. Genet.* 43:990. doi: 10.1038/ng.939

Krishan, K. (2006). Anthropometry in forensic medicine and forensic science-'Forensic Anthropometry'. *Int. J. Forensic. Sci.* 2, 1–3.

Marigorta, U. M., and Navarro, A. (2013). High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS Genet.* 9:e1003566. doi: 10.1371/journal.pgen.1003566

Matsunaga, H., Ito, K., Akiyama, M., Takahashi, A., Koyama, S., Nomura, S., et al. (2020). Transethnic meta-analysis of genome-wide association studies identifies three new loci and characterizes population-specific differences for Coronary artery disease. *Circ. Genom. Precis. Med.* 13:e002670. doi: 10.1161/CIRCGEN.119.002670

Moon, S., Kim, Y. J., Han, S., Hwang, M. Y., Shin, D. M., Park, M. Y., et al. (2019). The Korea biobank array: design and identification of coding variants associated with blood biochemical traits. *Sci. Rep.* 9:1382. doi: 10.1038/s41598-018-37832-9

N'Diaye, A., Chen, G. K., Palmer, C. D., Ge, B., Tayo, B., Mathias, R. A., et al. (2011). Identification, replication, and fine-mapping of Loci associated with adult height in individuals of african ancestry. *PLoS Genet.* 7:e1002298. doi: 10.1371/journal.pgen.1002298

Oh, S. W., Lee, J. E., Shin, E., Kwon, H., Choe, E. K., Choi, S. Y., et al. (2020). Genome-wide association study of metabolic syndrome in Korean populations. *PLoS One* 15:e0227357. doi: 10.1371/journal.pone.0227357

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

Randall, J. C., Winkler, T. W., Kutalik, Z., Berndt, S. I., Jackson, A. U., Monda, K. L., et al. (2013). Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genet.* 9:e1003500. doi: 10.1371/journal.pgen.1003500

Rask-Andersen, M., Karlsson, T., Ek, W. E., and Johansson, Å (2019). Genome-wide association study of body fat distribution identifies adiposity loci and

sex-specific genetic effects. *Nat. Commun.* 10:339. doi: 10.1038/s41467-018-08000-4

Richardson, T. G., Sanderson, E., Palmer, T. M., Ala-Korpela, M., Ference, B. A., Davey Smith, G., et al. (2020). Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: a multivariable Mendelian randomisation analysis. *PLoS Med.* 17:e1003062. doi: 10.1371/journal.pmed.1003062

Tachmazidou, I., Süveges, D., Min, J. L., Ritchie, G. R., Steinberg, J., Walter, K., et al. (2017). Whole-genome sequencing coupled to imputation discovers genetic signals for anthropometric traits. *Am. J. Hum. Genet.* 100, 865–884. doi: 10.1016/j.ajhg.2017.04.014

Thamizhselvi, E., and Geetha, V. (2019). "A comparative study of anthropometric measures and its significance on diverse applications," in *Proceedings of the 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, (New York, NY: IEEE), doi: 10.1109/ICSCAN.2019.8878748

Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., et al. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Genet.* 40:575. doi: 10.1038/ng.121

Wen, W., Kato, N., Hwang, J.-Y., Guo, X., Tabara, Y., Li, H., et al. (2016). Genome-wide association studies in East Asians identify new loci for waist-hip ratio and waist circumference. *Sci. Rep.* 6:17958. doi: 10.1038/srep17958

Wen, W., Zheng, W., Okada, Y., Takeuchi, F., Tabara, Y., Hwang, J.-Y., et al. (2014). Meta-analysis of genome-wide association studies in East Asian-ancestry populations identifies four new loci for body mass index. *Hum. Mol. Genet.* 23, 5492–5504. doi: 10.1093/hmg/ddu248

Winkler, T. W., Justice, A. E., Graff, M., Barata, L., Feitosa, M. F., Chu, S., et al. (2015). The influence of age and sex on genetic associations with adult body size and shape: a large-scale genome-wide interaction study. *PLoS Genet.* 11:e1005378. doi: 10.1371/journal.pgen.1005378

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature* 570, 514–518. doi: 10.1038/s41586-019-1310-4

Zhang, Y., Jin, S., Li, W., Gao, G., Zhang, K., and Huang, J. (2016). Association between RNF41 gene c.-206 T> a genetic polymorphism and risk of congenital heart diseases in the Chinese Mongolian population. *Genet. Mol. Res.* 15, 1–7. doi: 10.4238/gmr.15028089

Zhu, Y., Zhang, D., Zhou, D., Li, Z., Li, Z., Fang, L., et al. (2017). Susceptibility loci for metabolic syndrome and metabolic components identified in Han Chinese: a multi-stage genome-wide association study. *J. Cell. Mol. Med.* 21, 1106–1116. doi: 10.1111/jcmm.13042

Zhu, Z., Guo, Y., Shi, H., Liu, C.-L., Panganiban, R. A., Chung, W., et al. (2020). Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. *J. Allergy Clin. Immunol.* 145, 537–549. doi: 10.1016/j.jaci.2019.09.03

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 18 June 2021
doi: 10.3389/fgene.2021.690504

Insights Into Forensic Features and Genetic Structures of Guangdong Maoming Han Based on 27 Y-STRs

Haoliang Fan[1,2,3†], Qiqian Xie[1†], Yanning Li[1,4], Lingxiang Wang[2], Shao-Qing Wen[2*] and Pingming Qiu[1*]*

[1] School of Forensic Medicine, Southern Medical University, Guangzhou, China, [2] Institute of Archaeological Science, Fudan University, Shanghai, China, [3] School of Basic Medicine and Life Science, Hainan Medical University, Haikou, China, [4] School of Basic Medicine, Gannan Medical University, Ganzhou, China

OPEN ACCESS

Edited by:
Chuan-Chao Wang,
Xiamen University, China

Reviewed by:
Cemal Gurkan,
Turkish Cypriot DNA Laboratory, Cyprus
Pankaj Shrivastava,
State Forensic Science Laboratory, Sagar, India

***Correspondence:**
Shao-Qing Wen
wenshaoqing@fudan.edu.cn
orcid.org/0000-0003-1223-4720
Pingming Qiu
qiupmfy@126.com
orcid.org/0000-0002-5579-1124
Haoliang Fan
fanhaoliang198931@163.com
orcid.org/0000-0002-3214-0177

[†] These authors have contributed equally to this work

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

Received: 03 April 2021
Accepted: 14 May 2021
Published: 18 June 2021

Citation:
Fan H, Xie Q, Li Y, Wang L, Wen S-Q and Qiu P (2021) Insights Into Forensic Features and Genetic Structures of Guangdong Maoming Han Based on 27 Y-STRs. Front. Genet. 12:690504. doi: 10.3389/fgene.2021.690504

Maoming is located in the southwest region of Guangdong Province and is the cradle of Gaoliang culture, which is the representative branch of Lingnan cultures. Historical records showed that the amalgamations between Gaoliang aborigines and distinct ethnic minorities had some influences on the shaping of Gaoliang culture, especially for the local Tai-kadai language-speaking Baiyue and Han Chinese from Central China. However, there is still no exact genetic evidence for the influences on the genetic pool of Maoming Han, and the genetic relationships between Maoming Han and other Chinese populations are still unclear. Hence, in order to get a better understanding of the paternal genetic structures and characterize the forensic features of 27 Y-chromosomal short tandem repeats (Y-STRs) in Han Chinese from Guangdong Maoming, we firstly applied the AmpFLSTR® Yfiler® Plus PCR Amplification Kit (Thermo Fisher Scientific, Waltham, MA, United States) to genotype the haplotypes in 431 Han males residing in Maoming. A total of 263 different alleles were determined across all 27 Y-STRs with the corresponding allelic frequencies from 0.0004 to 0.7401, and the range of genetic diversity (GD) was 0.4027 (DYS391) to 0.9596 (DYS385a/b). In the first batch of 27 Yfiler data in Maoming Han, 417 distinct haplotypes were discovered, and nine off-ladder alleles were identified at six Y-STRs; in addition, no copy number variant or null allele was detected. The overall haplotype diversity (HD) and discrimination capacity (DC) of 27 Yfiler were 0.9997 and 0.9675, respectively, which demonstrated that the 6-dye and 27-plex system has sufficient system effectiveness for forensic applications in Maoming Han. What is more, the phylogenetic analyses indicated that Maoming Han, which is a Southern Han Chinese population, has a close relationship with Meizhou Kejia, which uncovered that the role of the gene flows from surrounding Han populations in shaping the genetic pool of Maoming Han cannot be ignored. From the perspectives of genetics, linguistics, and geographies, the genetic structures of Han populations correspond to the patterns of the geographical-scale spatial distributions and the relationships of language families. Nevertheless, no exact genetic evidence supports the intimate relationships between Maoming Han and Tai-Kadai language-speaking populations and Han populations of Central Plains in the present study.

Keywords: Maoming Han, Gaoliang culture, Y-STR, forensic features, genetic structures

INTRODUCTION

Maoming, a city located in the southwest of Guangdong Province (**Figure 1**), is the cradle of Gaoliang culture (Zhou, 2019). Gaoliang culture, one of the representative Lingnan cultures, could be dated back to the Han Dynasty (111 B.C.) in Chinese history (He, 2012). The aborigines living in Gaoliang mountainous areas and the basins between Jian River and Moyang River are the inheritors of Gaoliang culture, which are represented by the customs of *Nianli* (a special celebration for New Year) and *Piaose* (a form of dramatic plastic arts on moving stages) (Chen, 2013). Since the Southern and Northern Dynasties (420–589 A.D.), the intermarriages accelerated national amalgamations between Gaoliang aborigines and other ethnic minorities in ancient Gaoliang District (He, 2012; Chen, 2013)12). Therefore, Gaoliang culture was influenced by the convergences between Gaoliang aborigines and different ethnic groups (Gao, 2007). Moreover, some archeological records also hinted that the population structures of Gaoliang aborigines might be affected by the local Baiyue (a Tai-kadai language-speaking population in ancient China) and Han Chinese from Central China with the increasingly social activities of mixed marriages, population migrations, and trade contacts in the long course of history (Gao, 2007). Maoming Hans, the descendants of Gaoliang aborigines, speak Cantonese (*Gaoyang Pian*), which is one branch of Sino-Tibetan language family (Ding, 2010). From the perspective of languages, the language of Maoming Han (Cantonese) did not seem to be impacted by Tai-Kadai groups (Baiyue). Hence, there is still no exact genetic evidence for the influences on the genetic pool of Maoming Han, and the genetic relationships between Maoming Han and other surrounding populations are still unclear.

Y-chromosomal short tandem repeats (Y-STRs) have been regarded as a valuable tool in forensic genetics (Kayser, 2017), genealogy (Kayser et al., 2007), human evolution (Underhill and Kivisild, 2007), archeology (Calafell and Larmuseau, 2017), population history (Kayser et al., 2000; Jobling and Tyler-Smith, 2017), and male medical genetics (Hughes and Page, 2016; Jobling and Tyler-Smith, 2017). In addition, Y chromosomal variant analysis for determining the patterns of present and past flows of genes between populations is very helpful (Oppenheimer, 2012). The use of Y-STRs also allows the simultaneous analysis of closely related and distantly related populations (Ballantyne and Kayser, 2012). The 6-dye and 27-plex AmpFLSTR® Yfiler® Plus PCR Amplification Kit (Thermo Fisher Scientific, Waltham, MA, United States) includes 17 Yfiler loci (DYS19, DYS385a/b, DYS389I/II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, and Y GATA H4) plus three highly polymorphic Y-STR loci (DYS460, DYS481, and DYS533) and seven rapidly mutating Y-STR loci (DYF387S1a/b, DYS449, DYS518, DYS570, DYS576, and DYS627) in an effort to improve discrimination of related individuals (Gopinath et al., 2016). Y-chromosome STR haplotype reference database (YHRD)[1] is an internet-accessible worldwide reference database of Y chromosome profiles, which contributed to provide a worldwide and high-quality Y-STR haplotype data from distinct human populations for forensic purposes and population genetics (Willuweit and Roewer, 2015). As yet, little is known about the genetic backgrounds of the aforementioned 27 Y-STRs in Maoming Han, and the forensic-related Y chromosome variation data in Guangdong Maoming still remains blank in YHRD.

Hence, in order to get a better understanding of the paternal genetic structure and characterize the forensic resolution of 27 Y-STRs in Han Chinese from Guangdong Maoming, we used the 6-dye and 27-plex Y-STR system to genotype the haplotypes in 431 Han males residing in Maoming city. Furthermore, we explored the genetic relationships between Maoming Han and Chinese populations of Southern and Northern China from the perspectives of geographies, linguistics, and genetics.

MATERIALS AND METHODS

Sample Preparation

In this study, a total of 431 unrelated Han Chinese males were recruited from Maoming city, Guangdong Province, China (**Figure 1**). The inclusion criteria were as follows: (1) healthy individuals without any underlying diseases (including but not limited to cardiovascular diseases, metabolic diseases, chronic wasting diseases, immunologic diseases, etc.); (2) unrelated males and any two individuals who have no blood relationship for up to three generations; (3) the volunteers' parents and grandparents are aboriginals and have non-consanguineous marriages of the same ethnic group for at least three generations, which was confirmed by the volunteers' self-declared statements; and (4) the language Cantonese is the mother tongue of Maoming volunteers, and any self-declared Maoming Han who could not speak Cantonese would be excluded from our cohort. Blood samples of all Maoming volunteers were collected using FTA cards (Whatman^{TM}, GE Healthcare, Chicago, IL, United States) with written informed consents from participants. All the experimental procedures were performed following the standards of the Declaration of Helsinki. This study was approved by the Medical Ethics Committee of Hainan Medical University (no. HYLL-2020-012).

DNA Extraction, Amplification, and Genotyping

Genomic DNA was extracted using the TIANamp Blood Spot DNA Kit (TIANGEN BIOTECH, Beijing, China) according to the manufacturer's protocol. The quantity of the DNA templates was determined using Qubit^{TM} dsDNA HS Assay Kit (Thermo Fisher Scientific, Waltham, MA, United States) on the Qubit 4.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, United States) according to the manufacturer's instructions. Based on the quantitative results, DNA samples were normalized to 2.0 ng/μl and stored at −20°C until amplification.

The amplification of the 6-dye multiplex PCR-CE-based AmpFLSTR® Yfiler® Plus PCR Amplification Kit (Thermo Fisher Scientific, Waltham, MA, United States) was performed in a single multiplex PCR reaction (25 μl in total, containing

[1]https://yhrd.org/

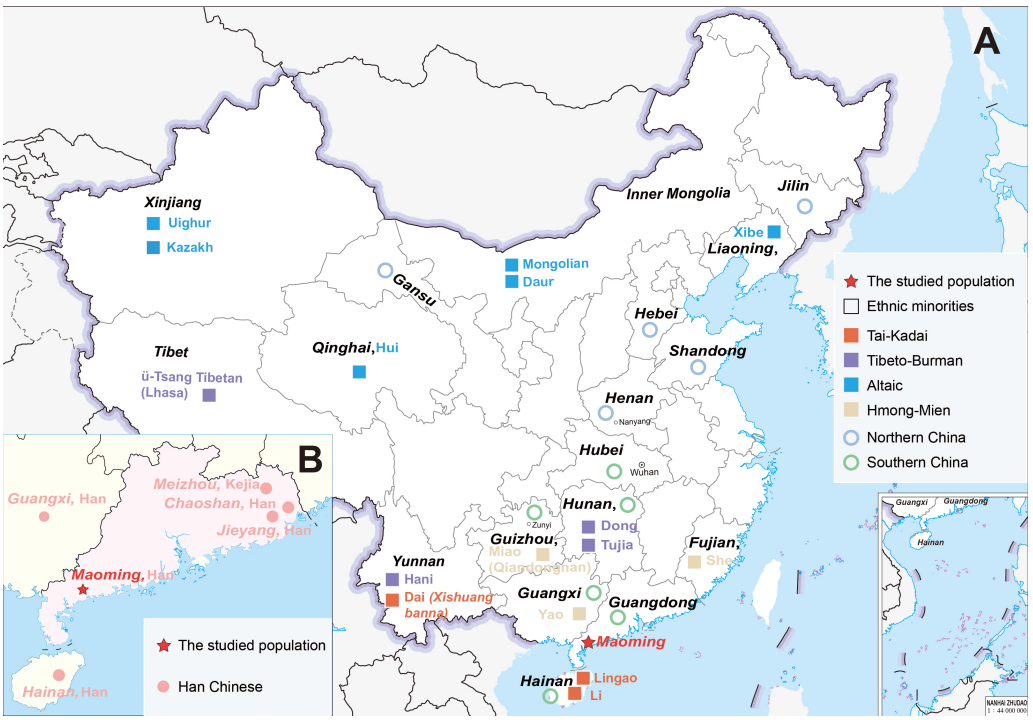

FIGURE 1 | Geographical locations of population distributions and sampling information. **(A)** The geographical distribution of Maoming Han and other Chinese populations (Han Chinese and ethnic minorities), which was analyzed in the present study. **(B)** The geographical distribution of Maoming Han and other five Han Chinese populations in South China (our studied group is marked as a red pentagram).

10 μl master mix, 5 μl primer mix, and 10 μl genomic DNA) on a Veriti® 96-Well Thermal Cycler System (Thermo Fisher Scientific, Waltham, MA, United States) following the manufacturer's instructions. Amplified products were separated by capillary electrophoresis (CE) on a 3500xL Genetic Analyzer (Thermo Fisher Scientific, Waltham, MA, United States). The separation of CE-based amplified products was conducted according to our previous studies (Fan et al., 2019a; Liu et al., 2020).

Statistic and Population Genetic Analyses

Allele and haplotype frequencies as well as forensic parameters were calculated using direct counting. The relevant forensic parameters contained genetic diversity (GD), haplotype diversity (HD), discrimination capacity (DC), and random match probability (RMP). GD was calculated according to the following formula:

$$\mathrm{GD} = \frac{n}{n-1} \times (1 - {p_i}^2)$$

where n is the total sample size, and p_i indicates the frequency of i-th allele. HD was computed in the same formula as GD, except that p_i refers to the frequency of i-th haplotype. DC is equal to the ratio of different haplotypes to the total sample size. Computed with the formula $\mathrm{RMP} = \sum {p_i}^2$, RMP is the probability that a particular DNA profile would appear in a population and that a "match" would occur by coincidence. In forensic statistics, a lower RMP value indicates higher strength of evidence provided by genetic analysis.

Population pairwise genetic distance (R_{st}) is commonly used for estimating the population differences and computing the genetic relationships among different populations (Fan et al., 2019b; Li et al., 2020). By using the "AMOVA&MDS tool" on YHRD, pairwise R_{st} and corresponding p values based on 17 Yfiler between Maoming Han and reference populations were estimated by analysis of molecular variance (AMOVA) and visualized in multidimensional scaling (MDS) plot, which were used show the reduced dimensionality spatial representation of the populations. Additionally, phylogenetic relationships among Han Chinese populations from Southern and Northern mainland China as well as those between 6 Han Chinese and 16 ethnic minorities were depicted in the Molecular Evolutionary Genetics Analysis-X (MEGA-X) software (Kumar et al., 2018) by a neighbor-joining (N-J) phylogenetic tree (Saitou and Nei, 1987) based upon the R_{st} genetic distance matrix, respectively.

Quality Control

The recommendations of the DNA Commission of the Chinese National Standards, the Scientific Working Group on DNA Analysis Methods (SWGDAM) (SWGDAM, 2010), and the DNA Commission of the International Society of Forensic Genetics (ISFG) (Gusmao et al., 2006; Carracedo et al., 2013; Roewer et al., 2020) for analysis of Y-STRs were strictly followed. Control DNA 007 was employed as a positive control, while ddH_2O

was used as a negative control for each batch of amplification and genotyping. Additionally, the laboratory has passed the proficiency testing for Y-STR typing organized by YHRD and has been accredited in accordance with ISO/IEC 17025:2005 and the China National Accreditation Service for Conformity Assessment (CNAS). The haplotype data of 431 unrelated male individuals from Guangdong Maoming Han population in the present study have been submitted to YHRD database and received the accession number YA004720 (Maoming Han, n = 431). The Y-STR profiles with off-ladders were re-amplified and re-genotyped by Goldeneye DNA^{TM} ID 27YB system (Goldeneye® Technology Ltd., Beijing, China).

RESULTS AND DISCUSSION

In the present study, a total of 431 unrelated male individuals from Han Chinese in Guangdong Maoming were genotyped including 27 Y-STR loci using the AmpFLSTR® Yfiler® Plus PCR Amplification Kit (Thermo Fisher Scientific, Waltham, MA, United States). In order to evaluate the forensic features of Maoming Han population, we set up two datasets, Yfiler set and Yfiler Plus set, including 17 and 27 Y-STRs, respectively. In addition, a series of comprehensive population genetic analyses were conducted between Maoming Han and other southern and northern Chinese populations. In short, the aims of this study were to feature the forensic characteristics of 27 Y-STRs in Maoming Han, clarify the paternal genetic structures of Maoming Han, and get a better understanding of the genetic relationships between Maoming Han and other Chinese populations from the perspectives of geographics, linguistics, and genetics.

Forensic Characteristics

Forensic Features of Yfiler Set (17 Y-STRs)

As illustrated in **Supplementary Table 1**, a total of 147 distinct alleles were identified across all 17 Y-STRs in Maoming Han with the corresponding allelic frequencies from 0.0023 to 0.7401 (DYS391). Overall, 17 Yfiler loci were relatively highly polymorphic in Maoming Han. The range of allele numbers was 4 (DYS391, DYS437, and DYS438) ∼ 55 (DYS385a/b), and the lowest and highest estimates of GD corresponded to loci DYS391 (0.4027) and DYS385a/b (0.9596). Except for DYS391 (0.4027) and DYS438 (0.4049), the GD values for other 17 Yfiler loci were greater than 0.5. The haplotypes and haplotype frequencies of Yfiler in Maoming Han are shown in **Supplementary Table 2**. There were 371 different haplotypes observed in 431 Maoming Han individuals, of which 328 (88.41%) were unique, 33 occurred twice (S011–S043), 6 (S005–S010) were observed thrice, 1 (H004) was shared by 4 individuals, and 3 (S001–S003) were shared by 5 individuals. We observed four confirmed microvariants [18.2 (twice) and 19.2 at DYS448 and 18.2 at DYS458]. The overall HD was 0.9994 with a DC of 0.8608.

Forensic Features of Yfiler Plus Set (27 Y-STRs)

Allele frequency distributions and haplotype frequencies of Yfiler Plus for Maoming Han are presented in **Supplementary Tables 3**, **4**. A total of 263 different alleles were observed, and the number of distinct alleles ranged from 4 for DYS391, DYS437, and DYS438 to 55 for DYS385a/b. Allele frequencies varied from 0.0004 to 0.7401. All 10 newly added loci got GD values higher than 0.5, especially for the added multi-copy DYF387S1a/b (0.9682). DYS385a/b (0.9596) on the one hand while DYS391 (0.4027) and DYS438 (0.4049) on the other marked the extremes of the GD distribution (with GD values less than 0.5). Genotyping with the 27 Y-STRs determined 417 distinct haplotypes in the population of Maoming Han, of which 405 (97.12%) were unique, 10 different haplotypes were identified twice (H003–H012), and 2 (H001-H002) appeared thrice. In addition to 18.2 and 19.2 at DYS448 and 18.2 at DYS458, intermediate alleles were also observed at the DYS449 (34.2), DYS518 (37.2), DYF387S1 (37.2), and DYS627 (17.2 and 18.2) loci. The overall HD and DC were calculated to be 0.9997 and 0.9675, respectively.

In this study, duplicated or triplicated alleles and null alleles were not detected in both Yfiler set and Yfiler Plus set. The analysis of genotype data revealed that DYS385a/b and DYF387S1a/b showed higher GD in Maoming Han, which were the same as other Chinese populations (Fan et al., 2018a). Forensic parameters based on different sets of Y-STR loci were calculated and listed in **Table 1**, indicating that as the number of Y-STR loci increased, more distinct haplotypes were identified, and HD and DC were also increased in the present study.

Genetic Differences Between Maoming Han and Han Chinese Populations From Southern and Northern Mainland China

From the intercontinental perspective, a MDS was performed between Maoming Han and 21 worldwide populations (Kim et al., 2001; Miranda et al., 2001; Roewer et al., 2005; Mizuno et al., 2008; Alam et al., 2010; Laouina et al., 2011; Wolfgramm Ede et al., 2011; Ramos-Luis et al., 2014; Xu et al., 2015; Han et al., 2016; Rapone et al., 2016; Wang et al., 2017; Alonso Morales et al., 2018; Fan et al., 2018c; Singh et al., 2018; Henry et al., 2019; Lang et al., 2019; Li et al., 2019; Salvador et al., 2019; Reid and Heathfield, 2020). As shown in **Supplementary Figure 1**,

TABLE 1 | Forensic features of Yfiler and Yfiler Plus loci in 431 Maoming Han.

Number of observed haplotypes	Y-filer (17Y-STR)	Y-filer Plus (27Y-STR)
1 (unique)	328	405
2	33	10
3	6	2
4	1	–
5	3	–
Sample size	431	431
Number of unique haplotypes	371	417
Proportion of unique haplotypes	0.8841	0.9712
HD	0.9994	0.9997
DC	0.8608	0.9675
RMP	0.0033	0.0027

HD, haplotype diversity; DC, discrimination capacity; RMP, random match probability.

Maoming Han clustered with Han Chinese populations, while other populations got together in accordance with geographical patterns relatively. In addition, to further explore the genetic affinity among Han Chinese populations in mainland China, the degree of differentiation between Maoming Han and Han populations from different administrative divisions of China was assessed by AMOVA and visualized in an MDS plot. Pairwise R_{st} and corresponding p values based on the 17 Yfiler among Maoming Han and 11 other Han Chinese populations from North to South across the mainland China (Yang et al., 2014; Shu et al., 2015; Han et al., 2016; Wang et al., 2016, 2017; Xu et al., 2016; Yao et al., 2016; Jiang et al., 2017; Chen et al., 2018; Fan et al., 2018c; Lang et al., 2019; Yin et al., 2020) are listed in **Supplementary Table 5**. Significant genetic differences were observed between Maoming Han and all other Han populations ($p = 0.05$) except for Guangdong Han. However, after Bonferroni's correction (p values = $0.05/431 \approx 0.0001$, $n = 431$), there were differences between Maoming Han and Northern Han Chinese from Gansu ($R_{st} = 0.0156$, $p < 0.0001$) and Shandong ($R_{st} = 0.0123$, $p < 0.0001$), while other Han Chinese populations had no significances with Maoming Han, especially for Guangdong Han ($R_{st} = 0.0008$, $p = 0.1509$), indicating the genetic affinity between Southern Han and Northern Han in mainland China. The MDS plot (**Figure 2A**) based on R_{st} values clearly demonstrated that the Northern Han populations, Gansu, Jilin, Hebei, Shandong, and Henan (Nanyang), were grouped in the bottom left side, while Hainan, Hubei (Wuhan), Hunan, Guizhou (Zunyi), and Guangdong isolated from the northern cluster and gathered into a Southern Han cluster in the upper right side with Maoming Han.

Subsequently, to make further confirmation about the genetic relationships between Maoming Han and other Han Chinese populations, an N-J phylogenetic tree based on R_{st} values was constructed (**Figure 2B**). We found that two main branches could be clearly identified in the N-J phylogenetic tree. The upper branch was Southern Han cluster, which was composed of Maoming, Guangdong, Guangxi, Hainan, Guizhou (Zunyi), Hunan, and Hubei (Wuhan), while Henan (Nanyang), Hebei, Shandong, Gansu, and Jilin got together in the bottom clade as Northern Han cluster. From the perspective of genetics, the analyses above indicated that Maoming Han is a Southern Han population and has relatively close relationships with Guangdong Han, followed by Guangxi Han ($R_{st} = 0.0031$, $p = 0.0155$) and Hainan Han ($R_{st} = 0.0033$, $p = 0.0183$). The phylogenetic structures of Han Chinese populations from Southern and Northern mainland China in the phylogenetic dendrogram were in line with the results of the MDS. From the geographical scale, Guangdong, Guangxi, Hainan, Guizhou, and Hunan belong to southern administrative divisions of mainland China, while Jilin, Gansu, Shandong, Hebei, and Henan are subordinated to the northern administrative divisions of mainland China (**Figure 1**). Maoming, lying in the southwest of Guangdong Province, has close geographical distances with Guangxi Zhuang Autonomous Region and Hainan Province. Even though different Han Chinese populations from distinct administrative divisions of mainland China have genetic and linguistic homogeneousness, the genetic distances and population structures of Han Chinese are in accordance with the geographical-scale pattern to a certain extent in mainland China.

Genetic Affinities and Differentiations Among Maoming Han, Other Han Populations, and Ethnic Minorities From China

According to the history records, the population structures of Maoming Han were mainly affected by the intermarriages with local Tai-kadai language-speaking Baiyue population and the south migrations of Han Chinese from Central China (He, 2012; Chen, 2013)12), while the above population analyses between our studied population and other Southern and Northern Han populations did not hint the relatively intimate relationships between Maoming Han and Han populations of Central Plains. To reveal the genetic structures among Maoming Han, surrounding Han populations and other 16 Chinese ethnic groups (Zhu et al., 2005; Shi et al., 2011; Shan et al., 2014; Zeng et al., 2014; Gao et al., 2015; Guo et al., 2015; Ou et al., 2015; Shu et al., 2015; Bian et al., 2016; Fu et al., 2016; Hu et al., 2017; Wang et al., 2017, 2019; Zhang et al., 2017; Zhao et al., 2017; Chen et al., 2018; Fan et al., 2018a,b,c; Du et al., 2019; Lang et al., 2019; Song et al., 2019; Xie et al., 2019; Ding et al., 2020; Feng et al., 2020; Guan et al., 2020), pairwise R_{st} and corresponding p values were calculated based on 17 Yfiler. As presented in **Supplementary Table 6**, no difference was observed between Maoming Han and Meizhou Kejia (R_{st} 0.0007, $p = 0.2899$), while significant genetic differences were observed between Maoming Han and all other Han Chinese or ethnic groups ($p < 0.05$). However, after Bonferroni's correction (p values ≈ 0.0001), there were no differences between Maoming Han and surrounding Han Chinese populations (**Supplementary Table 7**). Furthermore, we found that Inner Mongolia Daur ($R_{st} = 0.1614$), Xinjiang Uighur ($R_{st} = 0.1407$), and Lhasa U-Tsang Tibetan ($R_{st} = 0.1294$) in Northern China had the longest genetic distances with Maoming Han, while the closest genetic distance was seen in Meizhou Kejia ($R_{st} = 0.0007$), followed by Guangxi Han ($R_{st} = 0.0031$), Hainan Han ($R_{st} = 0.0033$), then by Jieyang Han ($R_{st} = 0.0039$), and Chaoshan Han ($R_{st} = 0.0083$).

On the basis of R_{st} values of 22 Chinese populations, an MDS plot (**Figure 3**) and an N-J phylogenetic tree (**Figure 4**) were performed to depict the forensic genetic landscape of Chinese Han and ethnic groups. As shown in **Figure 3A**, the Han Chinese populations were closely related to each other and therefore formed a Southern Han cluster, while other 16 minorities were relatively isolated from the Southern Han cluster and dispersed into four main clusters, which were in accord with the distributions of language families in some degree. The Hmong-Mien language-speaking groups, Miao, Yao, and She, clustered together at the upper left, and Dai, Lingao, and Li gathered in the bottom left as the Tai-Kadai language-speaking cluster, while the Tibeto-Burman-language speaking and Altaic-language speaking groups located together at the bottom right with relative separated positions. In addition, **Figure 3B** indicates the genetic relationships between Maoming Han and surrounding

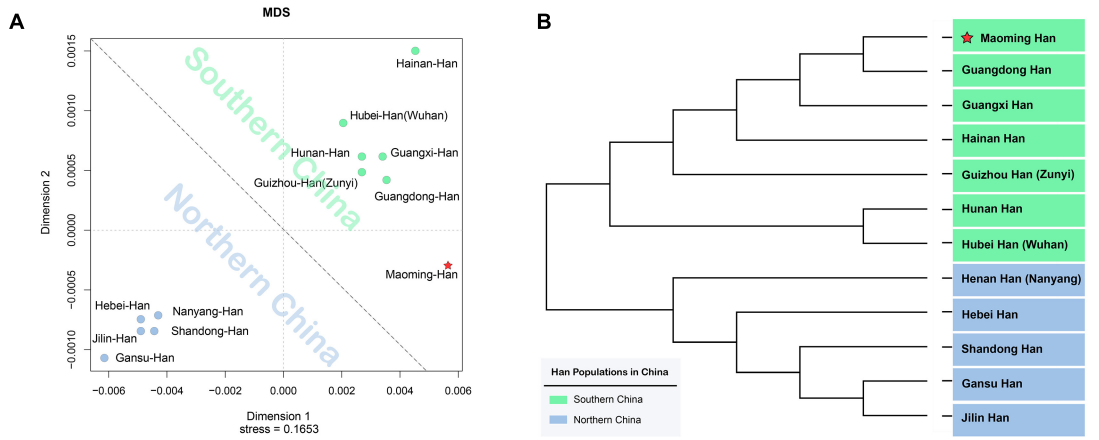

FIGURE 2 | Genetic similarities and differences between Maoming Han and other 11 Han Chinese populations from Northern and Southern China based on R_{st} values. **(A)** Multidimensional scaling (MDS) plot between Maoming Han and other 11 Han Chinese populations; **(B)** neighbor-joining (N-J) phylogenetic tree among 12 Han Chinese populations from Southern and Northern mainland China.

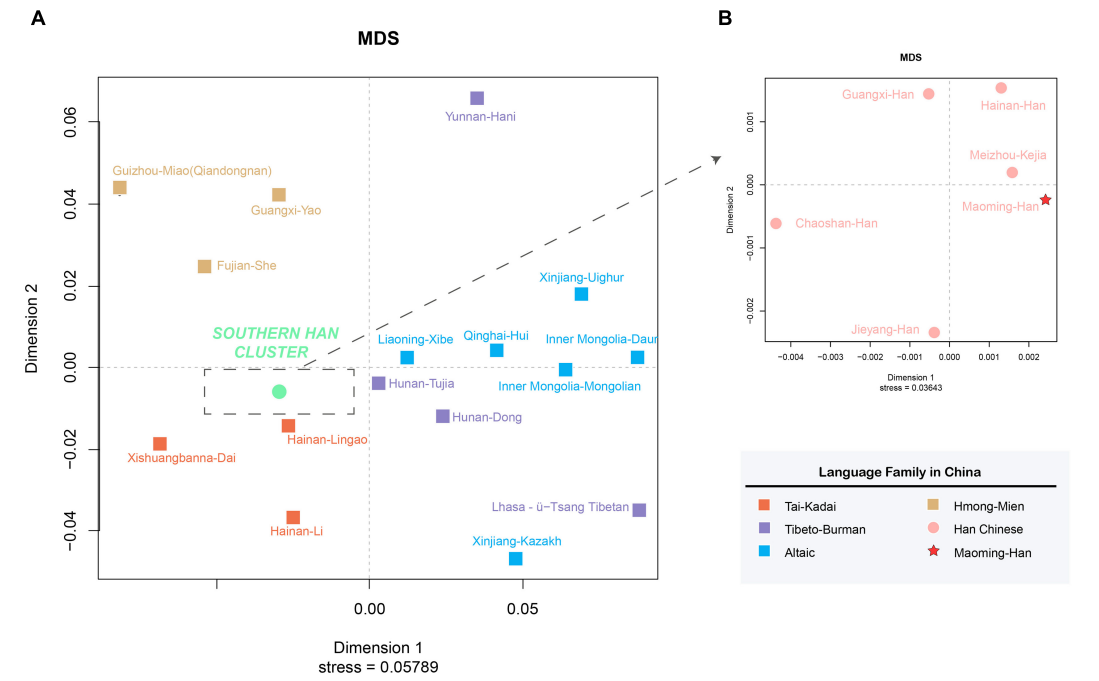

FIGURE 3 | Multidimensional scaling plots among Maoming Han, surrounding Han populations, and Chinese ethnic minorities based on pairwise R_{st} values. **(A)** MDS plot among Maoming Han, five other Chinese Han populations, as well as 16 ethnic minorities from China; **(B)** MDS plot between Maoming Han and surrounding Southern Han Chinese populations.

Han populations, which indicated that Maoming Han had a close relationship with Meizhou Kejia. Kejia, also known as Hakka, is a branch of Han Chinese that has a wide distribution in Guangdong Province. The genetic pool of Maoming Han was influenced by the surrounding Han populations, while no direct genetic evidence verified that the Tai-Kadai language-speaking populations contributed to the Maoming Han genetic pool. Furthermore, Meizhou Kejia was first clustered with the Maoming Han, followed by Guangxi Han and Hainan Han, then by Jieyang and Chaoshan in the phylogenetic tree (**Figure 4**). The tree also revealed that different populations were gathered into two cluster according to their geographical distributions and separated into two main branches: one represented the Altaic language-speaking populations; the other one stood for the Sino-Tibetan language-speaking populations (Han Chinese, Tibeto-Burman, Hmong-Mien, and Tai-Kadai), which was roughly congruent with the results of corresponding MDS (**Figure 3**).

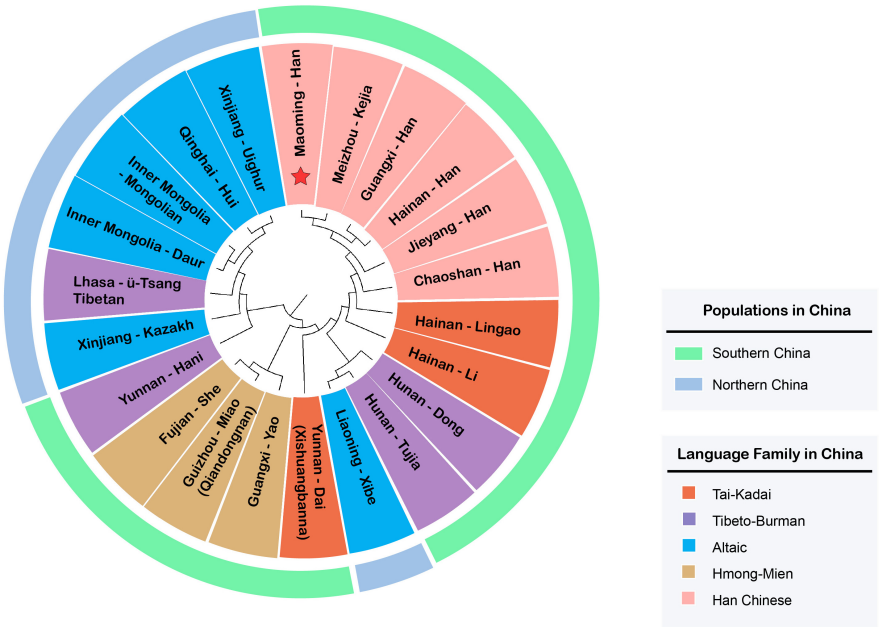

FIGURE 4 | Phylogenetic analysis among the Maoming people, five other Chinese Han populations, as well as 16 reference minorities from China based on pairwise R_{st} values (the language families and geographic distributions of different populations are displayed in the inner circle and the outer layer, respectively).

From the perspective of linguistics, geographies, and genetics, the phylogenetic analyses (both the MDS plots and N-J phylogenetic tree) demonstrated that Maoming Han was isolated from Chinese ethnic minority groups relatively and had a relatively close genetic relationships with Southern Han populations, especially for those with the same dialects and intimate geographical distances (Meizhou Kejia, Guangxi Han, and Hainan Han), which indicated that there might be gene flows between Maoming Han and the surrounding Han populations. In addition, the genetic structures of Han populations correspond to the patterns of the geographical-scale spatial distributions and the relationships of language families. In total, the results of above population genetic analyses indicated that Maoming Han, which is a Southern Han Chinese population, has a relatively close relationship with Meizhou Kejia; therefore, the role of the gene flows from surrounding Han populations in shaping the genetic pool of Maoming Han cannot be ignored.

CONCLUSION

In the present study, a total of 431 unrelated Guangdong Maoming Han were investigated using the AmpFLSTR® Yfiler® Plus PCR Amplification Kit (Thermo Fisher Scientific, Waltham, MA, United States). The high-quality 27 Y-STR haplotype data of Maoming Han were obtained and submitted to YHRD with the accession number YA004720. Overall, 263 different alleles were identified across all 27 Y-STRs with the number of distinct alleles from 4 to 55. Allele frequencies varied from 0.0004 to 0.7401, and the lowest and highest estimates of GD corresponded to loci DYS391 (0.4027) and DYS385a/b (0.9596), respectively. Genotyping with the 27 Y-STRs determined 417 distinct haplotypes in the population of Maoming Han, of which 405 (97.12%) were unique. In the first batch of 27 Yfiler data for Maoming Han, nine intermediate alleles were detected at six Y-STR loci; in addition, duplicated or triplicated alleles and null alleles were not observed. Based on the comparisons of forensic parameters for different sets of Y-STRs (17 Yfiler set and 27 Yfiler Plus set), it demonstrated that the improvements of HD and DC are accompanied by the increasing numbers of Y-STRs. The overall HD and DC of 27 Yfiler in Maoming Han were calculated to be 0.9997 and 0.9675, respectively.

From the perspectives of genetics, linguistics, and geographies, different Han Chinese populations from distinct administrative divisions of mainland China have genetic and linguistic homogeneousness, and the genetic distances and population structures of Han Chinese are in accordance with the geographical-scale pattern to a certain extent in mainland China. Maoming Han, a Southern Han population, has a relatively close genetic relationship with Meizhou Kejia, which has the same language family and has intimate geographical distances with Maoming Han, while no exact genetic evidence supports that there are intimate relationships between Maoming Han and Tai-Kadai language-speaking populations and Han populations of Central Plains. At the same time, we found that the genetic structures of Han populations correspond to the patterns of the geographical-scale spatial distributions and the relationships of language families. As a whole, the sufficient systematic efficiencies of AmpFLSTR® Yfiler® Plus PCR Amplification Kit in Maoming Han demonstrated that it can be widely applied in the population of Guangdong Maoming Han for forensic purposes, and Maoming Han, which is a Southern

Han Chinese population, and has a relatively close relationship with Meizhou Kejia; therefore, the role of the gene flows from surrounding Han populations in shaping the genetic pool of Maoming Han cannot be ignored.

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 were reviewed and approved by Medical Ethics Committee of the Hainan Medical University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

HF made significant contributions in the conceptualization, resources, software, formal analysis, and project administration. YL and QX made significant contributions in the investigation. LW performed the validation. QX performed the data curation. HF and QX performed the visualization, wrote and prepared the original draft, and reviewed and edited the manuscript. PQ and S-QW made significant contributions in the supervision of the study. PQ acquired funding for the study. All authors reviewed the manuscript.

FUNDING

This study was supported by grants from the National Natural Science Foundation of China (NSFC, No. 81971786).

ACKNOWLEDGMENTS

We would like to thank the donors who contributed samples for this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.690504/full#supplementary-material

REFERENCES

Alam, S., Ali, M. E., Ferdous, A., Hossain, T., Hasan, M. M., and Akhteruzzaman, S. (2010). Haplotype diversity of 17 Y-chromosomal STR loci in the Bangladeshi population. *Forensic Sci. Int. Genet.* 4, e59–e60. doi: 10.1016/j.fsigen.2009.05.005

Alonso Morales, L. A., Casas-Vargas, A., Rojas Castro, M., Resque, R., Ribeiro-Dos-Santos, ÂK., Santos, S., et al. (2018). Paternal portrait of populations of the middle Magdalena river region (Tolima and Huila, Colombia): new insights on the peopling of central America and northernmost South America. *PLoS One* 13:e0207130. doi: 10.1371/journal.pone.0207130

Ballantyne, K. N., and Kayser, M. (2012). Additional Y-STRs in forensics: why, which, and when. *Forensic Sci. Rev.* 24, 63–78.

Bian, Y., Zhang, S., Zhou, W., Zhao, Q., Siqintuya, Zhu, R., et al. (2016). Analysis of genetic admixture in uyghur using the 26 Y-STR loci system. *Sci. Rep.* 6:19998. doi: 10.1038/srep19998

Calafell, F., and Larmuseau, M. H. D. (2017). The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research. *Hum. Genet.* 136, 559–573. doi: 10.1007/s00439-016-1740-0

Carracedo, A., Butler, J. M., Gusmao, L., Linacre, A., Parson, W., Roewer, L., et al. (2013). New guidelines for the publication of genetic population data. *Forensic Sci. Int. Genet.* 7, 217–220. doi: 10.1016/j.fsigen.2013.01.001

Chen, P., He, G., Zou, X., Zhang, X., Li, J., Wang, Z., et al. (2018). Genetic diversities and phylogenetic analyses of three Chinese main ethnic groups in southwest China: a Y-chromosomal STR study. *Sci. Rep.* 8:15339. doi: 10.1038/s41598-018-33751-x

Chen, Y. (2013). *Cultural Business Card in Local Maoming. Nan Fang Lun Kan*(09). 87–89.

Ding, J., Fan, H., Zhou, Y., Wang, Z., Wang, X., Song, X., et al. (2020). Genetic polymorphisms and phylogenetic analyses of the Ü-Tsang Tibetan from Lhasa based on 30 slowly and moderately mutated Y-STR loci. *Forensic Sci. Res.* doi: 10.1080/20961790.2020.1810882 [Epub ahead of print].

Ding, S. (2010). The dialects and cultures in maoming. *Forward Position* 14, 164–166.

Du, W., Wu, W., Wu, Z., Guo, L., Wang, B., and Chen, L. (2019). Genetic polymorphisms of 32 Y-STR loci in Meizhou Hakka population. *Int. J. Legal Med.* 133, 465–466. doi: 10.1007/s00414-018-1845-1

Fan, H., Wang, X., Chen, H., Li, W., Wang, W., and Deng, J. (2019a). The Ong Be language-speaking population in Hainan island: genetic diversity, phylogenetic characteristics and reflections on ethnicity. *Mol. Biol. Rep.* 46, 4095–4103. doi: 10.1007/s11033-019-04859-8

Fan, H., Wang, X., Chen, H., Long, R., Liang, A., Li, W., et al. (2018a). The evaluation of forensic characteristics and the phylogenetic analysis of the Ong Be language-speaking population based on Y-STR. *Forensic Sci. Int. Genet.* 37, e6–e11. doi: 10.1016/j.fsigen.2018.09.008

Fan, H., Wang, X., Chen, H., Zhang, X., Huang, P., Long, R., et al. (2018b). Population analysis of 27 Y-chromosomal STRs in the Li ethnic minority from Hainan province, southernmost China. *Forensic Sci. Int. Genet.* 34, e20–e22. doi: 10.1016/j.fsigen.2018.01.007

Fan, H., Wang, X., Ren, Z., He, G., Long, R., Liang, A., et al. (2019b). Population data of 19 autosomal STR loci in the Li population from Hainan province in southernmost China. *Int. J. Legal Med.* 133, 429–431. doi: 10.1007/s00414-018-1828-2

Fan, H., Zhang, X., Wang, X., Ren, Z., Li, W., Long, R., et al. (2018c). Genetic analysis of 27 Y-STR loci in Han population from Hainan province, southernmost China. *Forensic Sci. Int. Genet.* 33, e9–e10. doi: 10.1016/j.fsigen.2017.12.009

Feng, R., Zhao, Y., Chen, S., Li, Q., Fu, Y., Zhao, L., et al. (2020). Genetic analysis of 50 Y-STR loci in Dong, Miao, Tujia, and Yao populations from Hunan. *Int. J. Legal Med.* 134, 981–983. doi: 10.1007/s00414-019-02115-z

Fu, X., Fu, Y., Liu, Y., Guo, J., Liu, Y., Guo, Y., et al. (2016). Genetic polymorphisms of 26 Y-STR loci in the Mongolian minority from Horqin district. China. *Int. J. Legal Med.* 130, 941–946. doi: 10.1007/s00414-016-1387-3

Gao, H. (2007). A Brief Analysis of the Integration of the Han and Li Nationalities in Ancient Gaoliang. Journal of Guangzhou Institute of Socialism(04). 38–41.

Gao, T., Yun, L., Gu, Y., He, W., Wang, Z., and Hou, Y. (2015). Phylogenetic analysis and forensic characteristics of 12 populations using 23 Y-STR loci. *Forensic Sci. Int. Genet.* 19, 130–133. doi: 10.1016/j.fsigen.2015.07.006

Gopinath, S., Zhong, C., Nguyen, V., Ge, J., Lagace, R. E., Short, M. L., et al. (2016). Developmental validation of the Yfiler((R)) plus PCR amplification kit: an enhanced Y-STR multiplex for casework and database applications. *Forensic Sci. Int. Genet.* 24, 164–175. doi: 10.1016/j.fsigen.2016.07.006

Guan, T., Song, X., Xiao, C., Sun, H., Yang, X., Liu, C., et al. (2020). Analysis of 23 Y-STR loci in Chinese Jieyang Han population. *Int. J. Legal Med.* 134, 505–507. doi: 10.1007/s00414-019-02019-y

Guo, F., Zhang, L., and Jiang, X. (2015). Population genetics of 17 Y-STR loci in Xibe ethnic minority from Liaoning province, Northeast China. *Forensic Sci. Int. Genet.* 16, 86–87. doi: 10.1016/j.fsigen.2014.12.007

Gusmao, L., Butler, J. M., Carracedo, A., Gill, P., Kayser, M., Mayr, W. R., et al. (2006). DNA commission of the international society of forensic genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. *Int. J. Legal Med.* 120, 191–200. doi: 10.1007/s00414-005-0026-1

Han, Y., Li, L., Liu, X., Chen, W., Yang, S., Wei, L., et al. (2016). Genetic analysis of 17 Y-STR loci in Han and Korean populations from Jilin province, Northeast China. *Forensic Sci. Int. Genet.* 22, 8–10. doi: 10.1016/j.fsigen.2016.01.003

He, H. (2012). *Expanding the Research Field and Creating a New Cultural Field: A Review of Feng Guixiong and Wu Gang's Madam Xian's Family and Yangjiang in Sui and Tang dynasties. Nan Fang Lun Kan(03)*. 91–93.

Henry, J., Dao, H., Scandrett, L., and Taylor, D. (2019). Population genetic analysis of Yfiler(®) Plus haplotype data for three South Australian populations. *Forensic Sci. Int. Genet.* 41, e23–e25. doi: 10.1016/j.fsigen.2019.03.021

Hu, L., Gu, T., Fan, X., Yuan, X., Rao, M., Pang, J. B., et al. (2017). Genetic polymorphisms of 24 Y-STR loci in Hani ethnic minority from Yunnan province, Southwest China. *Int. J. Legal Med.* 131, 1235–1237. doi: 10.1007/s00414-017-1543-4

Hughes, J. F., and Page, D. C. (2016). The history of the Y chromosome in man. *Nat. Genet.* 48, 588–589. doi: 10.1038/ng.3580

Jiang, W., Gong, Z., Rong, H., Guan, H., Zhang, T., Zhao, Y., et al. (2017). Population genetics of 26 Y-STR loci for the Han ethnic in Hunan province, China. *Int. J. Legal Med.* 131, 115–117. doi: 10.1007/s00414-016-1411-7

Jobling, M. A., and Tyler-Smith, C. (2017). Human Y-chromosome variation in the genome-sequencing era. *Nat. Rev. Genet.* 18, 485–497. doi: 10.1038/nrg.2017.36

Kayser, M. (2017). Forensic use of Y-chromosome DNA: a general overview. *Hum. Genet.* 136, 621–635. doi: 10.1007/s00439-017-1776-9

Kayser, M., Brauer, S., Weiss, G., Underhill, P. A., Roewer, L., Schiefenhovel, W., et al. (2000). Melanesian origin of polynesian Y chromosomes. *Curr. Biol.* 10, 1237–1246. doi: 10.1016/s0960-9822(00)00734-x

Kayser, M., Vermeulen, M., Knoblauch, H., Schuster, H., Krawczak, M., and Roewer, L. (2007). Relating two deep-rooted pedigrees from central Germany by high-resolution Y-STR haplotyping. *Forensic Sci. Int. Genet.* 1, 125–128. doi: 10.1016/j.fsigen.2007.02.004

Kim, Y. J., Shin, D. J., Kim, J. M., Jin, H. J., Kwak, K. D., Han, M. S., et al. (2001). Y-chromosome STR haplotype profiling in the Korean population. *Forensic Sci. Int.* 115, 231–237. doi: 10.1016/s0379-0738(00)00332-7

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Lang, M., Liu, H., Song, F., Qiao, X., Ye, Y., Ren, H., et al. (2019). Forensic characteristics and genetic analysis of both 27 Y-STRs and 143 Y-SNPs in eastern Han Chinese population. *Forensic Sci. Int. Genet.* 42, e13–e20. doi: 10.1016/j.fsigen.2019.07.011

Laouina, A., El Houate, B., Yahia, H., Azeddoug, H., Boulouiz, R., and Chbel, F. (2011). Allele frequencies and population data for 17 Y-STR loci (The AmpFlSTR® Y-filer™) in Casablanca resident population. *Forensic Sci. Int. Genet.* 5, e1–e3. doi: 10.1016/j.fsigen.2010.10.016

Li, L., Xu, Y., Luis, J. R., Alfonso-Sanchez, M. A., Zeng, Z., Garcia-Bertrand, R., et al. (2019). Cebú, Thailand and Taiwanese aboriginal populations according to Y-STR loci. *Gene X* 1:100001. doi: 10.1016/j.gene.2018.100001

Li, W., Wang, X., Wang, X., Wang, F., Du, Z., Fu, F., et al. (2020). Forensic characteristics and phylogenetic analyses of one branch of Tai-Kadai language-speaking Hainan Hlai (Ha Hlai) via 23 autosomal STRs included in the Huaxia() platinum system. *Mol. Genet. Genomic Med.* 8:e1462. doi: 10.1002/mgg3.1462

Liu, J., Wang, R., Shi, J., Cheng, X., Hao, T., Guo, J., et al. (2020). The construction and application of a new 17-plex Y-STR system using universal fluorescent PCR. *Int. J. Legal Med.* 134, 2015–2027. doi: 10.1007/s00414-020-02291-3

Miranda, J. J., Benecke, M., Hidding, M., and Schmitt, C. (2001). Y-chromosomal short tandem repeat haplotypes at the loci DYS393, DYS19, DYS392, and DYS385-I/II, DYS390, DYS389-I/II, and DYS391 in a Filipino population sample. *J. Forensic Sci.* 46, 1250–1253.

Mizuno, N., Nakahara, H., Sekiguchi, K., Yoshida, K., Nakano, M., and Kasai, K. (2008). 16 Y chromosomal STR haplotypes in Japanese. *Forensic Sci Int* 174, 71–76. doi: 10.1016/j.forsciint.2007.01.032

Oppenheimer, S. (2012). Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 770–784. doi: 10.1098/rstb.2011.0306

Ou, X., Wang, Y., Liu, C., Yang, D., Zhang, C., Deng, S., et al. (2015). Haplotype analysis of the polymorphic 40 Y-STR markers in Chinese populations. *Forensic Sci. Int. Genet.* 19, 255–262. doi: 10.1016/j.fsigen.2015.08.007

Ramos-Luis, E., Blanco-Verea, A., Brión, M., Van Huffel, V., Sánchez-Diz, P., and Carracedo, A. (2014). Y-chromosomal DNA analysis in French male lineages. *Forensic Sci. Int. Genet.* 9, 162–168. doi: 10.1016/j.fsigen.2013.12.008

Rapone, C., D'Atanasio, E., Agostino, A., Mariano, M., Papaluca, M. T., Cruciani, F., et al. (2016). Forensic genetic value of a 27 Y-STR loci multiplex (Yfiler(§) Plus kit) in an Italian population sample. *Forensic Sci. Int. Genet.* 21, e1–e5. doi: 10.1016/j.fsigen.2015.11.006

Reid, K. M., and Heathfield, L. J. (2020). Allele frequency data for 23 Y-chromosome short tandem repeats (STRs) for the South African population. *Forensic Sci. Int. Genet.* 46:102270. doi: 10.1016/j.fsigen.2020.102270

Roewer, L., Andersen, M. M., Ballantyne, J., Butler, J. M., Caliebe, A., Corach, D., et al. (2020). DNA commission of the international society of forensic genetics (ISFG): recommendations on the interpretation of Y-STR results in forensic analysis. *Forensic Sci. Int. Genet.* 48:102308. doi: 10.1016/j.fsigen.2020.102308

Roewer, L., Croucher, P. J., Willuweit, S., Lu, T. T., Kayser, M., Lessig, R., et al. (2005). Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. *Hum. Genet.* 116, 279–291. doi: 10.1007/s00439-004-1201-z

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425. doi: 10.1093/oxfordjournals.molbev.a040454

Salvador, J., Rodriguez, J., Carandang, L., Agmata, A., Honrado, M., Delfin, F., et al. (2019). Filipino DNA variation at 36 Y-chromosomal short tandem repeat (STR) marker units. *Philipp. J. Sci.* 148, 43–52.

Shan, W., Ablimit, A., Zhou, W., Zhang, F., Ma, Z., and Zheng, X. (2014). Genetic polymorphism of 17 Y chromosomal STRs in Kazakh and Uighur populations from Xinjiang. China. *Int. J. Legal Med.* 128, 743–744. doi: 10.1007/s00414-013-0948-y

Shi, M., Bai, R., Bai, L., and Yu, X. (2011). Population genetics for Y-chromosomal STRs haplotypes of Chinese Xibe ethnic group. *Forensic Sci. Int. Genet.* 5, e119–e121. doi: 10.1016/j.fsigen.2010.08.004

Shu, L., Li, L., Yu, G., Yu, B., Liu, Y., Li, S., et al. (2015). Genetic analysis of 17 Y-STR loci in Han, Dong, Miao and Tujia populations from Hunan province, central-southern China. *Forensic Sci. Int. Genet.* 19, 250–251. doi: 10.1016/j.fsigen.2015.07.007

Singh, M., Sarkar, A., and Nandineni, M. R. (2018). A comprehensive portrait of Y-STR diversity of Indian populations and comparison with 129 worldwide populations. *Sci. Rep.* 8:15421. doi: 10.1038/s41598-018-33714-2

Song, M., Wang, Z., Zhang, Y., Zhao, C., Lang, M., Xie, M., et al. (2019). Forensic characteristics and phylogenetic analysis of both Y-STR and Y-SNP in the Li and Han ethnic groups from Hainan island of China. *Forensic Sci. Int. Genet.* 39, e14–e20. doi: 10.1016/j.fsigen.2018.11.016

SWGDAM (2010). *Scientific Working Group on DNA Analysis Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories*.

Underhill, P. A., and Kivisild, T. (2007). Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu. Rev. Genet.* 41, 539–564. doi: 10.1146/annurev.genet.41.110306.130407

Wang, C. Z., Su, M. J., Li, Y., Chen, L., Jin, X., Wen, S. Q., et al. (2019). Genetic polymorphisms of 27 Yfiler(®) plus loci in the daur and Mongolian ethnic minorities from Hulunbuir of inner Mongolia autonomous region, China. *Forensic Sci. Int. Genet.* 40, e252–e255. doi: 10.1016/j.fsigen.2019.02.003

Wang, M., Wang, Z., Zhang, Y., He, G., Liu, J., and Hou, Y. (2017). Forensic characteristics and phylogenetic analysis of two Han populations from the southern coastal regions of China using 27 Y-STR loci. *Forensic Sci. Int. Genet.* 31, e17–e23. doi: 10.1016/j.fsigen.2017.10.009

Wang, Y., Zhang, Y. J., Zhang, C. C., Li, R., Yang, Y., Ou, X. L., et al. (2016). Genetic polymorphisms and mutation rates of 27 Y-chromosomal STRs in a Han population from Guangdong province, Southern China. *Forensic Sci. Int. Genet.* 21, 5–9. doi: 10.1016/j.fsigen.2015.09.013

Willuweit, S., and Roewer, L. (2015). The new Y chromosome haplotype reference database. *Forensic Sci. Int. Genet.* 15, 43–48. doi: 10.1016/j.fsigen.2014.11.024

Wolfgramm Ede, V., Silva, B. C., Aguiar, V. R., Malta, F. S., de Castro, A. M., Ferreira, A. C., et al. (2011). Genetic analysis of 15 autosomal and 12 Y-STR loci in the Espirito Santo state population, Brazil. *Forensic Sci. Int. Genet.* 5, e41–e43. doi: 10.1016/j.fsigen.2010.05.001

Xie, M., Song, F., Li, J., Lang, M., Luo, H., Wang, Z., et al. (2019). Genetic substructure and forensic characteristics of Chinese Hui populations using 157 Y-SNPs and 27 Y-STRs. *Forensic Sci. Int. Genet.* 41, 11–18. doi: 10.1016/j.fsigen.2019.03.022

Xu, H., Wang, C. C., Shrestha, R., Wang, L. X., Zhang, M., He, Y., et al. (2015). Inferring population structure and demographic history using Y-STR data from worldwide populations. *Mol. Genet. Genomics* 290, 141–150. doi: 10.1007/s00438-014-0903-8

Xu, J., Li, L., Wei, L., Nie, Z., Yang, S., Xia, M., et al. (2016). Genetic analysis of 17 Y-STR loci in Han population from Shandong province in East China. *Forensic Sci Int Genet* 22, e15–e17. doi: 10.1016/j.fsigen.2016.01.016

Yang, Y., Yuan, W., Guo, F., and Jiang, X. (2014). Population data of 17 Y-STR loci in Nanyang Han population from Henan province, central China. *Forensic Sci. Int. Genet.* 13, 145–146. doi: 10.1016/j.fsigen.2014.07.013

Yao, H. B., Wang, C. C., Tao, X., Shang, L., Wen, S. Q., Zhu, B., et al. (2016). Genetic evidence for an East Asian origin of Chinese Muslim populations Dongxiang and Hui. *Sci. Rep.* 6:38656. doi: 10.1038/srep38656

Yin, C., Su, K., He, Z., Zhai, D., Guo, K., Chen, X., et al. (2020). Genetic reconstruction and forensic analysis of Chinese Shandong and Yunnan Han populations by Co-analyzing Y chromosomal STRs and SNPs. *Genes (Basel)* 11:743. doi: 10.3390/genes11070743

Zeng, Z., Rowold, D. J., Garcia-Bertrand, R., Calderon, S., Regueiro, M., Li, L., et al. (2014). Taiwanese aborigines: genetic heterogeneity and paternal contribution to oceania. *Gene* 542, 240–247. doi: 10.1016/j.gene.2014.03.005

Zhang, J., Wang, J., Liu, Y., Shi, M., Bai, R., and Ma, S. (2017). Haplotype data for 27 Y-chromosomal STR loci in the Chaoshan Han population, South China. *Forensic Sci. Int. Genet.* 31, e54–e56. doi: 10.1016/j.fsigen.2017.08.003

Zhao, Q., Bian, Y., Zhang, S., Zhu, R., Zhou, W., Gao, Y., et al. (2017). Population genetics study using 26 Y-chromosomal STR loci in the Hui ethnic group in China. *Forensic Sci. Int. Genet.* 28, e26–e27. doi: 10.1016/j.fsigen.2017.01.018

Zhou, J. (2019). *A Preface to "Gaoliang Culture List" in Gaozhou City*. Available online at: http://blog.sina.com.cn/s/blog_15029bea60102zec2.html (accessed December 07, 2019).

Zhu, B., Li, X., Wang, Z., Wu, H., He, Y., Zhao, J., et al. (2005). Y-STRs haplotypes of Chinese Mongol ethnic group using Y-PLEX 12. *Forensic Sci. Int.* 153, 260–263. doi: 10.1016/j.forsciint.2004.11.002

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.

The handling editor declared a past co-authorship with the authors LW and S-QW.

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 June 2021
doi: 10.3389/fgene.2021.685285

Genomic Insights Into the Admixture History of Mongolic- and Tungusic-Speaking Populations From Southwestern East Asia

Jing Chen[1†], Guanglin He[2†], Zheng Ren[1], Qiyan Wang[1], Yubo Liu[1], Hongling Zhang[1], Meiqing Yang[1], Han Zhang[1], Jingyan Ji[1], Jing Zhao[2], Jianxin Guo[2], Kongyang Zhu[2], Xiaomin Yang[2], Rui Wang[2], Hao Ma[2], Chuan-Chao Wang[2,3] and Jiang Huang[1]**

[1] Department of Forensic Medicine, Guizhou Medical University, Guiyang, China, [2] State Key Laboratory of Cellular Stress Biology, State Key Laboratory of Marine Environmental Science, Department of Anthropology and Ethnology, Institute of Anthropology, National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China, [3] School of Basic Medical Sciences, Zhejiang University School of Medicine, Hangzhou, China

OPEN ACCESS

Edited by:
Horolma Pamjav,
Ministry of Interior, Hungary

Reviewed by:
Balazs Egyed,
Eötvös Loránd University, Hungary
Jatupol Kampuansai,
Chiang Mai University, Thailand

***Correspondence:**
Chuan-Chao Wang
wang@xmu.edu.cn
Jiang Huang
mmm_hj@126.com

[†] These authors have contributed equally to this work

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

Received: 24 March 2021
Accepted: 05 May 2021
Published: 22 June 2021

Citation:
Chen J, He G, Ren Z, Wang Q, Liu Y, Zhang H, Yang M, Zhang H, Ji J, Zhao J, Guo J, Zhu K, Yang X, Wang R, Ma H, Wang C-C and Huang J (2021) 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

As a major part of the modern *Trans*-Eurasian or Altaic language family, most of the Mongolic and Tungusic languages were mainly spoken in northern China, Mongolia, and southern Siberia, but some were also found in southern China. Previous genetic surveys only focused on the dissection of genetic structure of northern Altaic-speaking populations; however, the ancestral origin and genomic diversification of Mongolic and Tungusic–speaking populations from southwestern East Asia remain poorly understood because of the paucity of high-density sampling and genome-wide data. Here, we generated genome-wide data at nearly 700,000 single-nucleotide polymorphisms (SNPs) in 26 Mongolians and 55 Manchus collected from Guizhou province in southwestern China. We applied principal component analysis (PCA), ADMIXTURE, *f* statistics, *qpWave/qpAdm* analysis, *qpGraph*, TreeMix, Fst, and ALDER to infer the fine-scale population genetic structure and admixture history. We found significant genetic differentiation between northern and southern Mongolic and Tungusic speakers, as one specific genetic cline of Manchu and Mongolian was identified in Guizhou province. Further results from ADMIXTURE and *f* statistics showed that the studied Guizhou Mongolians and Manchus had a strong genetic affinity with southern East Asians, especially for inland southern East Asians. The *qpAdm*-based estimates of ancestry admixture proportion demonstrated that Guizhou Mongolians and Manchus people could be modeled as the admixtures of one northern ancestry related to northern Tungusic/Mongolic speakers or Yellow River farmers and one southern ancestry associated with Austronesian, Tai-Kadai, and Austroasiatic speakers. The *qpGraph*-based phylogeny and neighbor-joining tree further confirmed that Guizhou Manchus and Mongolians derived approximately half of the ancestry from their northern ancestors and the other half from southern Indigenous East Asians. The estimated admixture time ranged from 600 to 1,000 years ago, which further confirmed the admixture events were mediated *via* the Mongolians Empire expansion during the formation of the Yuan dynasty.

Keywords: population history, genetic structure, genetic admixture, East Asia, population genetics

INTRODUCTION

The East Asian continent has abundant ethnolinguistic diversity and profound history of the populations. The Altaic languages, including Mongolic, Tungusic, and Turkic, are widely distributed in northern East Asia, Siberia, and part region of Central Asia. Previous studies from a genetic perspective have mainly demonstrated the northern East Asian affinity of Mongolic and Tungusic–speaking populations based on the genome-wide single-nucleotide polymorphism (SNP) data or sharing IBD fragments (Yunusbayev et al., 2015; Pugach et al., 2016; Jeong et al., 2020; Kilinc et al., 2021). Based on the large-scale sampling of the ancient and present-day populations from Mongolia, Lake Baikal, to Amur River Basin, it is observed that the Mongolians and Tungusic-speaking groups have a higher proportion of genetic component related to the Devil's Gate people who were early Neolithic hunter–gatherers in northeastern East Asia dating to more than 7.7 thousand years ago (Siska et al., 2017), as well as Mongolians Neolithic people (Jeong et al., 2020; Wang C. C. et al., 2021). The massive migration of Neolithic people between the eastern Mongolians plateau and the Amur River basin had shaped the culture and genetic structure of Bronze Age and Iron Age and even historic pastoralist empires (Xiongnu, Xianbei, Rouran, Khitan, and Uyghur) (Jeong et al., 2020). This identified ancestry component was referred to as the ancient northeast Asian ancestry compared with the ancient components from Ancient Northern Eurasians and also played an important genetic contribution to modern Mongolic and Tungusic speakers. The genetic similarity of Mongolic and Tungusic populations is also shown in a similar pattern of the paternal Y chromosomes (Wei et al., 2017a,b, 2018a; Zhang et al., 2018). The Y-haplogroup C2*, C2a, and C2b have been identified as the founder paternal lineages of the Tungusic population through whole Y-chromosome sequencing (Wei et al., 2018b). Especially, haplogroup C2a-F5484 has contributed largely to both modern Mongolians and Tungusic populations (Liu et al., 2020). Because of the vast geographic distribution, the present-day Mongolian populations in northern East Asia were suggested to have a distinct genetic substructure due to substantial gene flows between northern Eurasian populations in the past as revealed by whole-genome sequencing (Bai et al., 2018; Zhao et al., 2020). Previous genetic surveys mainly focused on the northern Altaic-speaking populations; however, the ancestral origin and genomic diversification of Mongolic and Tungusic–speaking populations from southwestern East Asia remain poorly understood because of the paucity of high-density sampling and genome-wide data.

Guizhou province, located at the eastern end of the Yunnan-Guizhou Plateau, harbors a diverse array of ethnic groups and linguistic backgrounds including the Mongolic and Tungusic languages (Wang Q. et al., 2021). According to local chronicles and folklore, during the Yuan Dynasty, the Mongolian people were recruited to various regions including Guizhou for their southward or westward expeditions[1], while the settlement of the Tungusic-speaking Manchus in Guizhou was related to the implementation of military plans by the Qing Dynasty. However, the genetic profile of the Manchus and Mongolian speakers in southern China is still very much in its infancy. Here, we generated genome-wide data at nearly 700,000 SNPs in 26 Mongolian and 55 Manchu individuals collected from three populations in Guizhou province and compared with available data of both modern and ancient East Asian individuals to explore their fine-scale population genetic structure.

MATERIALS AND METHODS

Sampling and Genotyping

We collected saliva samples from 26 Mongolians and 55 Manchus in Guizhou province, southwestern China (**Supplementary Figure 1**). These samples were collected randomly from unrelated participants whose parents and grandparents are Indigenous people and have a non-consanguineous marriage of the same ethnical group for at least three generations. The ethnicities of all participates were used as their self-declaration based on their family migration history and corresponding family records. Our study and sample collection were reviewed and approved by the Medical Ethics Committee of Guizhou Medical University and followed the recommendations provided by the revised Helsinki Declaration of 2000. The participants provided their written informed consent before they were invited to have participated in this study. We used PureLink Genomic DNA Mini Kit (Thermo Fisher Scientific) to extract DNA and measure the concentration *via* the Nanodrop-2000. Infinium® Global Screening Array (GSA, Shenzhen, China) was used to genotype approximately 700,000 SNPs, which covered SNPs from the autosome, Y-chromosome, and merohedral DNA. Raw data in the binary form (bed, bim, and fam) were initial filtered using PLINK 1.9 (Chang et al., 2015) based on our predefined threshold of the genotyping success rate, missing site rates, minor allele frequency, and Hardy–Weinberg equilibrium (–maf 0.01,–hwe 1e-6, mind: 0.01, and geno: 0.01). A final dataset with 6,992,479 SNPs was used to perform the following population genetic analysis.

Data Merging

We merged our population data of 81 newly genotyped samples with previously published modern and ancient populations from Human Origins (HO) dataset (Patterson et al., 2012) and the 1240K dataset from the David Reich laboratory[2], and other recently published ancient East Asians populations (Ning et al., 2020; Yang et al., 2020; Wang C. C. et al., 2021). The 1240K dataset harbored higher-density SNP data from ancient populations, especially for the genome-wide ancient data *via* the capture sequence or whole-genome sequence; however, HO dataset not only has all these ancient DNA data but only has more modern population reference data genotyped *via* the Affymetrix HO array, which can provide more representative source population to construct the modern population genetic background. The detailed information of our used reference population data was

[1]https://en.wikipedia.org/wiki/Yuan_dynasty

[2]https://reich.hms.harvard.edu/downloadablegenotypes-present-day-and-ancient-dna-data-compiled-published-papers

listed in **Supplementary Table 1**. We finally generated two combined datasets used in subsequent analysis covering 72,532 in the merged HO dataset and 193,846 SNPs in the merged 1240K dataset, respectively.

Principal Component Analysis

We carried out the principal component analysis (PCA) using the smartpca package built-in EIGENSOFT (Patterson et al., 2012). We performed PCA based on present-day East Asian populations and then projected the ancient samples onto the basal axis based on the top two components using the lsqproject: YES option, which accounts for samples with substantial missing data. We did not perform any outlier removal iterations (numoutlieriter: 0). We set all other options to the default and assessed the statistical significance with a Tracy–Widom test using the twstats program of EIGENSOFT.

ADMIXTURE Analysis

To further explore the ancestry composition and genetic similarity of our studied groups with geographically close ancient and present-day populations, we carried out model-based clustering analysis using ADMIXTURE 1.23 (Alexander et al., 2009) by combining the present-day and ancient worldwide populations samples with our 81 individuals. We performed model-based ADMIXTURE analysis based on the unlinked SNP data (–indep-pairwise 200 25 0.4). We ran ADMIXTURE with default fivefold cross-validation (-CV = 5), varying the number of ancestral populations between $K = 2$ and $K = 20$ in 100 bootstraps with different random seeds. We used the unsupervised ADMIXTURE approach, in which allele frequencies for unadmixed ancestral populations are unknown and are computed during the analysis. We used point estimation and terminated the block relaxation algorithm when the objective function delta < 0.0001. We chose the best run according to the highest log-likelihood. We used cross-validation to identify an "optimal" number of clusters. We observed the lowest CV error at $K = 11$.

Admixture and Outgroup f_3 Statistics

We used the *qp3pop* in ADMIXTOOLS (Patterson et al., 2012) to perform the outgroup f_3 (Reference1, Reference2; Mbuti) to assess the shared genetic drift between reference populations 2 and reference populations 2 since their separation from an African outgroup population of Mbuti using the default parameters. Then, we used the *qp3pop* to perform the admixture-f_3 (Reference1, Reference2; Target populations) to explore the admixture signatures in our studied Guizhou Manchus and Mongolian samples with different Eurasian ancestral source candidates, where a significant negative-f_3 value with |Z-score| larger than three denoted that the targeted population was an admixture between two parental populations.

f_4 Statistics

We computed f_4 statistics of the form $f_4(X, Y;$ Test, Outgroup) using the *qpDstat* program in ADMIXTOOLS with default parameters and estimated standard errors using the block jackknife (Patterson et al., 2012). The statistics can show if the population test is symmetrically related to X and Y or shares an excess of alleles with either of the two.

qpAdm Estimation

We investigated the admixture source numbers, plausible admixture sources, and the corresponding admixture proportions based on *qpWave* and *qpAdm* programs in ADMIXTOOLS (Patterson et al., 2012) using the following outgroups: Mbuti, Papuan, Australian, Mixe, Russia_MA1_HG, Onge, Atayal, Ust_Ishim, Russia_Kostenki14, and China_Tianyuan. Parameter of "allsnps: YES" was used here. We used the spatiotemporally different Yellow River basin farmers as the northern sources and Fujian or Taiwan modern and ancient as the southern sources to perform the two population qpAdm model. To further dissect the admixture proportions from inland or coastal southern East Asians, we additionally included ancient populations from Southeast Asia as the third source to conduct three-way admixture models.

TreeMix and *qpGraph*

Phylogenetic relationship with migration events among modern East Asians was performed using TreeMix and *qpGraph* to explore admixture models with population splits and gene flow in Manchus and Mongolians. We followed the basic model to reconstruct the deep population genomic history of our targeted populations (Wang C. C. et al., 2021).

ALDER-Based Admixture Times

Admixture dates from the possible admixture sources for Manchus and Mongolians were estimated using ALDER (Loh et al., 2013). We used geographically different northern and southern East Asians as candidate sources to estimate the admixture time. We used Plink 1.9 (Chang et al., 2015) and our in-house script to calculate the pairwise Fst indexes (Weir and Cockerham, 1984).

Y-Chromosomal and mtDNA Haplogroup Assignment

There were 26,341 paternal lineages informative SNPs and 4,198 maternal-informative SNPs genotyped *via* the Infinium® GSA. Ancestral or derived statuses of these SNPs were used to identify the terminal haplogroup. We used in-house tools (unpublished software) to assign the Y-chromosomal paternal lineage following the basic regulations reaccommodated *via* the International Society of Genetic Genealogy[3]. We classified the maternal mitochondrial haplogroups used HaploGrep 2 (Weissensteiner, 2016).

RESULTS

We successfully genotyped approximately 700,000 genome-wide SNPs in 26 Mongolians and 55 Manchus in the Guizhou province, China. We then merged our data with worldwide modern and

[3]https://isogg.org/

ancient published populations from the HO dataset and 1240K dataset, which included modern populations from Altaic, Sino-Tibetan, Austronesian, Austroasiatic, Hmong-Mien, and Tai-Kadai speakers in East Asia (Wang C. C. et al., 2021), as well as ancient DNA data from Nepal (Jeong et al., 2016), Mongolia (Jeong et al., 2020), Siberia (Lazaridis et al., 2014; Raghavan et al., 2014a,b, 2015; Rasmussen et al., 2014; Mathieson et al., 2015; Damgaard et al., 2018; de Barros Damgaard et al., 2018; Sikora et al., 2019), North and South China (Yang et al., 2017, 2020; Ning et al., 2020; Wang C. C. et al., 2021), and Southeast Asia (Lipson et al., 2018; McColl et al., 2018). To understand the general patterns of relatedness between Guizhou Manchus, Mongolians, and published populations, we first performed PCA to provide a overview pattern of the population structure across East Asia (**Figure 1**). We observed the following five genetic clusters correlating well with geographic and linguistic categories within East Asia: (I) a northern Altaic cluster consisting of Tungusic and Mongolic–speaking groups in North China, Mongolia, and Siberia; (II) a southern China/Southeast Asia cluster with Austroasiatic, Tai-Kadai, and Austronesian speaking groups; (III) a western Tibetan Plateau cluster being made up of Tibeto-Burman–speaking populations; (IV) a southern inland East Asian

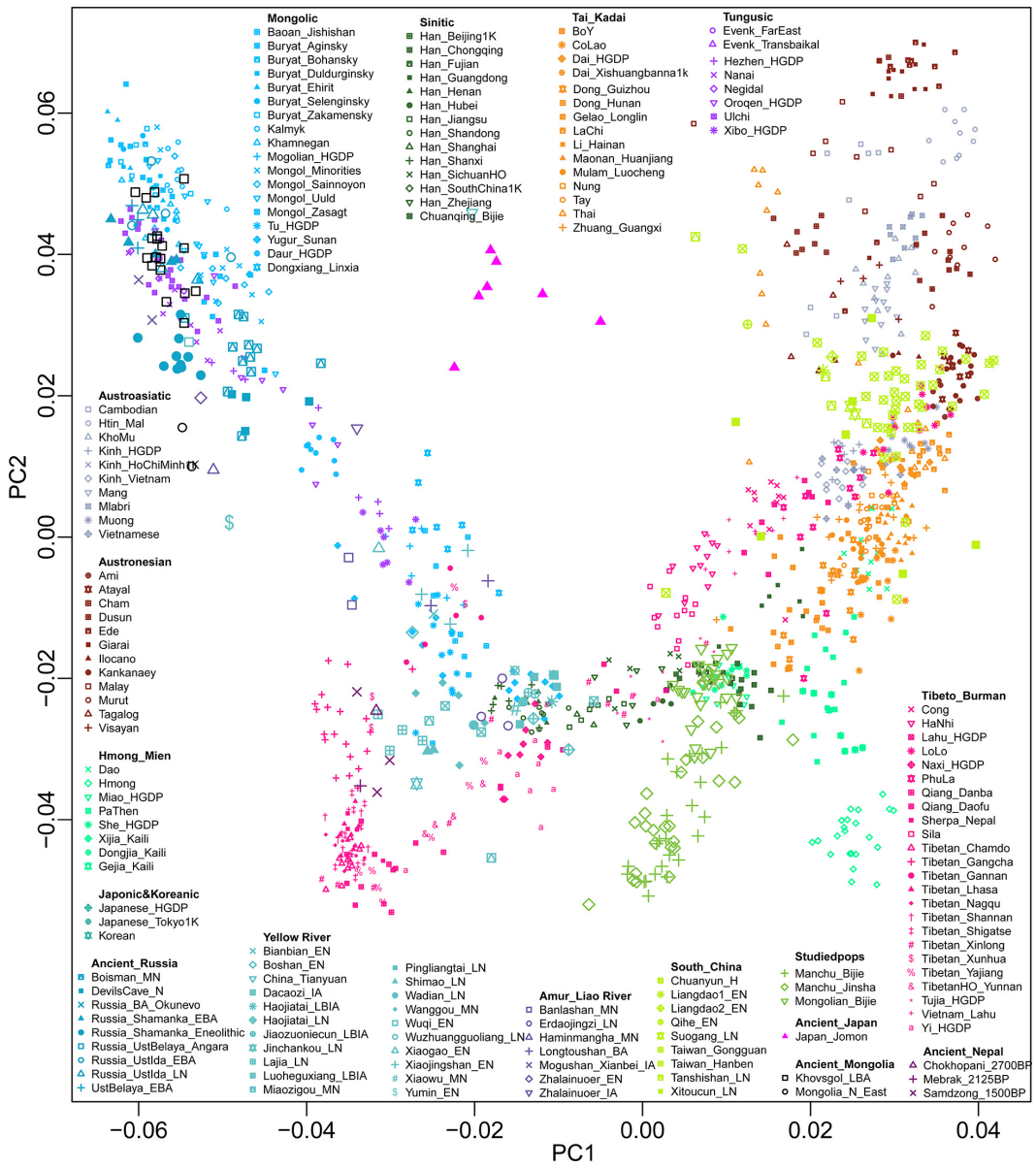

FIGURE 1 | Patterns of genetic relationship among East Asian populations inferred from principal component analysis. Genetic background was constructed based on the genetic variations from modern populations and their top two components. Modern populations were color-coded on the basis of their language family categories. All ancient populations were projected onto it.

Hmong-Mien cluster comprising Hmong, Dao, Gejia, Dongjia, and Xijia; and (VI) a new identified southern Chinese Altaic cluster consisting of Tungusic and Mongolic–speaking groups. Our studied Tungusic and Mongolic–speaking populations from Guizhou province formed a unique genetic cline, which was located at an intermediate position between the western Tibetan Plateau cluster and Hmong-Mien cluster and partially overlapped with previously published Sinitic and Hmong-Mien speaking populations.

In the model-based ADMIXTURE clustering analysis, we used cross-validation to identify an "optimal" number of clusters. We observed the lowest CV error at $K = 11$. At $K = 11$, we observed three ancestral components in our studied Guizhou Manchus and Mongolian samples (**Figure 2**). One of these components is enriched in the ancient Nepalese and also found at the highest proportions in Tibetans, with the second component with maximum representation in the Tai-Kadai- and Austroasiatic-speaking populations. The remaining ancestry component in our studied populations was maximized in Austronesian speakers and also enriched in ancient samples from southeast China including Fujian and Taiwan. In general, we found our Manchus and Mongolians are genetically similar to the Hmong-Mien–speaking populations and Han Chinese in South China.

To formally test the genetic affinity observed in PCA and ADMIXTURE and find the potential ancestral sources for Guizhou Manchus and Mongolians, we measured allele sharing and admixture signals *via* outgroup f_3 and admixture-f_3 statistics. Specifically, in the outgroup f_3 statistics of the form $f_3(X$, Guizhou Manchus/Mongolians; Mbuti), Guizhou Manchus shared more alleles with Han Chinese, She, Ami, and Miao. When X represented ancient individuals, Guizhou Manchus was found to share more alleles with Neolithic-Iron Age Yellow River farming populations including Haojiatai, followed by Jiaozuoniecun and Luoheguxiang ancients. Guizhou Mongolians shared more alleles with Han Chinese, Ami, ancient Gongguan samples from Taiwan, She, and Miao (**Figure 3** and **Supplementary Table 2A**). Besides, we used admixture-f_3 statistics of the form $f_3(X, Y$; Guizhou Manchus/Mongolians) to model possible admixtures, where X and Y were East Asian populations that might be the source candidates for modeling the admixture in Guizhou Manchus or Mongolians when getting negative Z scores. However, we observed only one significant signal of admixture ($Z < -3$) in the Mongolian_Bijie when using Tibetan as the northern East Asian source and Austronesian-speaking Igorot people as the southern East Asian source (**Supplementary Tables 2B–D**). This suggests that the allele frequencies of Mongolian_Bijie are intermediate between those of a northern group related to Tibetans and a southern group related to the Austronesian-speaking people. We also calculated pairwise Fst genetic distances among these populations (**Supplementary Table 3**), and the patterns observed here were consistent with the f_3-based results.

We then performed f_4 statistics to explore genetic substructure between studied groups and other modern/ancient East Asians in the form f_4 (*study group 1*, *study group 2*; *East Asians*, *Mbuti*). We observed significant negative f_4 values in f_4 (*Manchu_Jinsha*, *Mongolian_Bijie*; *East Asians*, *Mbuti*) (**Supplementary Table 4A**) when we used ancient Hanben samples from Taiwan, Atayal, and early Neolithic Liangdao1 people in the position of "East Asians," showing that Bijie Mongolians shared the most derived alleles with ancient or modern southern East Asians compared with Jinsha Manchus. We have not observed significant f_4 values in f_4 (*Manchu_Jinsha*, *Manchu_Bijie*; *East Asians*, *Mbuti*) (**Supplementary Table 4B**), suggesting Manchus from Jinsha and Bijie form a clade with a closer genetic relationship compared with other East Asians. We observed suggestive evidence that Bijie Mongolians may obtain additional gene flow from southern East Asians compared with Bijie Manchus by finding of marginal negative Z scores of f_4 (*Manchu_Bijie*, *Mongolian_Bijie*; *East Asians*, *Mbuti*) (**Supplementary Table 4C**).

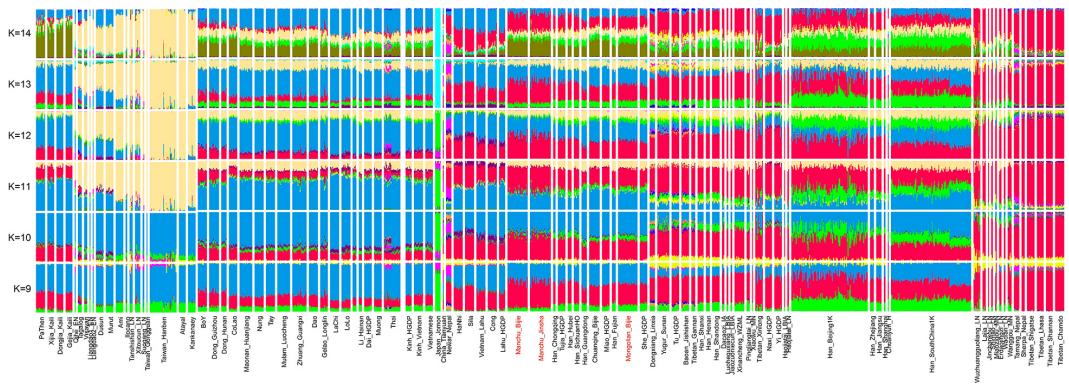

FIGURE 2 | Results of model-based ADMIXTURE clustering analysis. Clustering patterns were visualized with the predefined ancestral sources ranging from 9 to 14 among East Asians (K: 9–14). Here, we can identify late Neolithic to Iron Age Taiwan Hanben dominant ancestry widely distributed in Austronesian speakers, LoChi or Lolo-dominant ancestry maximized in Tai-Kadai–speaking populations, Tibetan-dominant ancestry widely distributed in Tibeto-Burman–speaking populations, and others, all of these ancestries were color-coded by different colors.

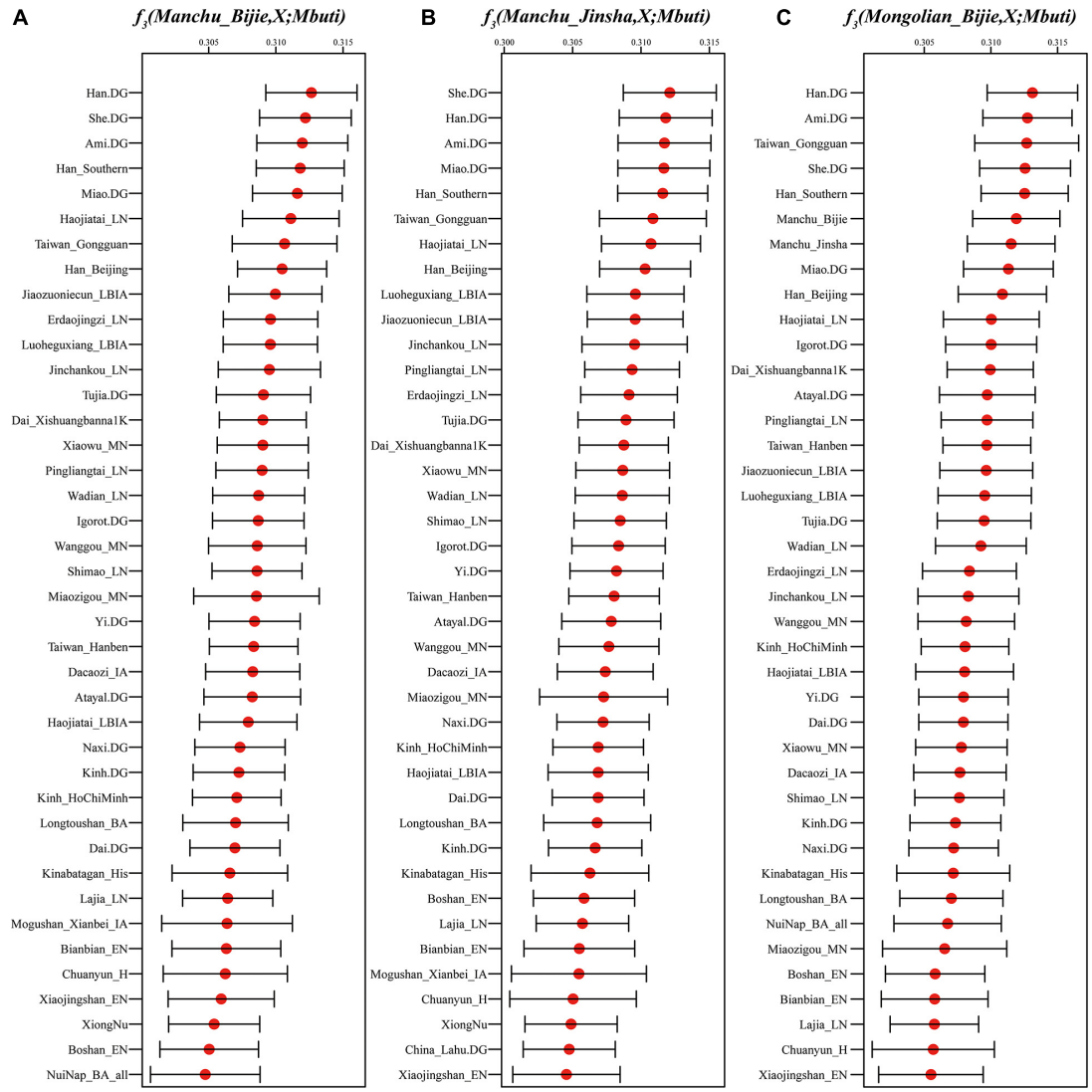

FIGURE 3 | Shared genetic drift estimated *via* admixture-f_3 statistics in different testing forms based on the merged 1240K dataset focused on Bijie Manchus **(A)**, Jinsha Manchus **(B)**, and Bijie Mongolians **(C)**. Bar denotes the three times of error bar.

We found that Guizhou Manchu and Mongolian people harbored more northern Mongolic and Tungusic–related ancestry compared with Guizhou indigenous populations by the observation of significant positive values in f_4 (*Guizhou Manchus/Mongolians*, *Guizhou indigenous populations*; *northern Mongolians/Tungusic populations*, *Mbuti*) (**Supplementary Table 5**). Further evidence demonstrated that studied populations harbored more southern East Asian–related ancestry compared to ancient northern East Asians *via* the significant negative f_4 statistics in the form f_4 (*ancient Yellow River millet farmer*, *Guizhou Manchus/Mongolians*; *southern East Asians*, *Mbuti*) (**Supplementary Table 6**). Compared with ancient populations in southeast China including Fujian and Taiwan, we observed that Guizhou Manchus and Mongolians shared more alleles with northern East Asians *via* significant negative f_4 values in from of f_4 (*Taiwan_Hanben/Xitoucun_LN/Tanshishan_LN*, *Guizhou Manchus/Mongolians*; *northern East Asians*, *Mbuti*) (**Supplementary Table 7**). Similarly, when compared with present-day southern Sinitic, Austronesian, Tai-Kadai, Hmong-Mien–speaking populations from southern China and the Islands of Southeast Asia, Guizhou Manchus and Mongolians have excess allele sharing with northern East Asians (**Supplementary Table 8**).

Considering the observed excess allele sharing and possible sources for our studied Manchus and Mongolians people, we applied *qpWave* and *qpAdm* methods to model their ancestry. We used all available ancient northern populations (Bianbian, Boshan, Xiaogao, Xiaowu, Luoheguxiang, Dacaozi_IA, Longtoushan_BA, Shimao_LN, Miaozigou_MN, and Yumin_EN) as the northern sources and Iron Age Hanben samples from Taiwan as the southern sources to estimate the admixture proportions. The Southern East Asian Hanben-like

ancestry proportion spanned from 16.5 to 35.7% when using Yellow River farmers as the northern source, whereas the proportion reached 56.7% when using Yumin_EN (hunter–gatherers in Inner Mongolia) (**Supplementary Table 9A**). To explore if there was any genetic influence from inland southern East Asians related to Austroasiatic speakers, we conducted three-way admixture models by adding ancient Southeast Asians as a third source. The best-fitted three-way admixture proximal models for Manchus and Mongolians are as deriving ancestry from northern ancient Yellow River farming populations, Austronesian-related ancient Southern East Asians (Taiwan_Hanben/Gongguan, Xitoucun), and Austroasiatic-related ancient Southeast Asians (GuaCha_LN, MaiDaDieu_LN, ManBac_LN, NamTun_LN, PhaFaen_Hoabinhian, and TamHang_BA) (**Supplementary Table 9B**).

In the TreeMix analysis (**Figure 4**), we observed Mongolian-speaking groups in southern Siberia and Tungusic-speaking groups in the Amur River basin cluster together as the northern branch, while the Austronesian, Austroasiatic, Hmong-Mien, and Tai-Kadai speakers from southern China cluster together forming the southern branch. Our studied Mongolians and Manchus groups, Tibeto-Burman and Sinitic populations were located at an intermediate position between the northern and southern branches. Specifically, the two Guizhou Manchus groups in this study clustered together first and then clustered with the Guizhou Mongolians group at an intermediate position between the Sinitic and Hmong-Mien–speaking populations. The clustering pattern was consistent with the patterns observed in the aforementioned PCA, ADMIXTURE, and *f* statistics–based analysis that Guizhou Manchus and Mongolians had experienced genetic influence from surrounding southern Indigenous populations since their separation from northern ancestors and migrated to Guizhou.

We further used *qpGraph* to reconstruct the deep evolutionary history of the Mongolians group in Guizhou. We used two ancient Neolithic samples from the Mongolians Plateau as the northern source and used the samples from the middle Neolithic Xiaowu site as a representative of the ancient Yellow River millet farmers. We used Iron Age Hanben samples from Taiwan as

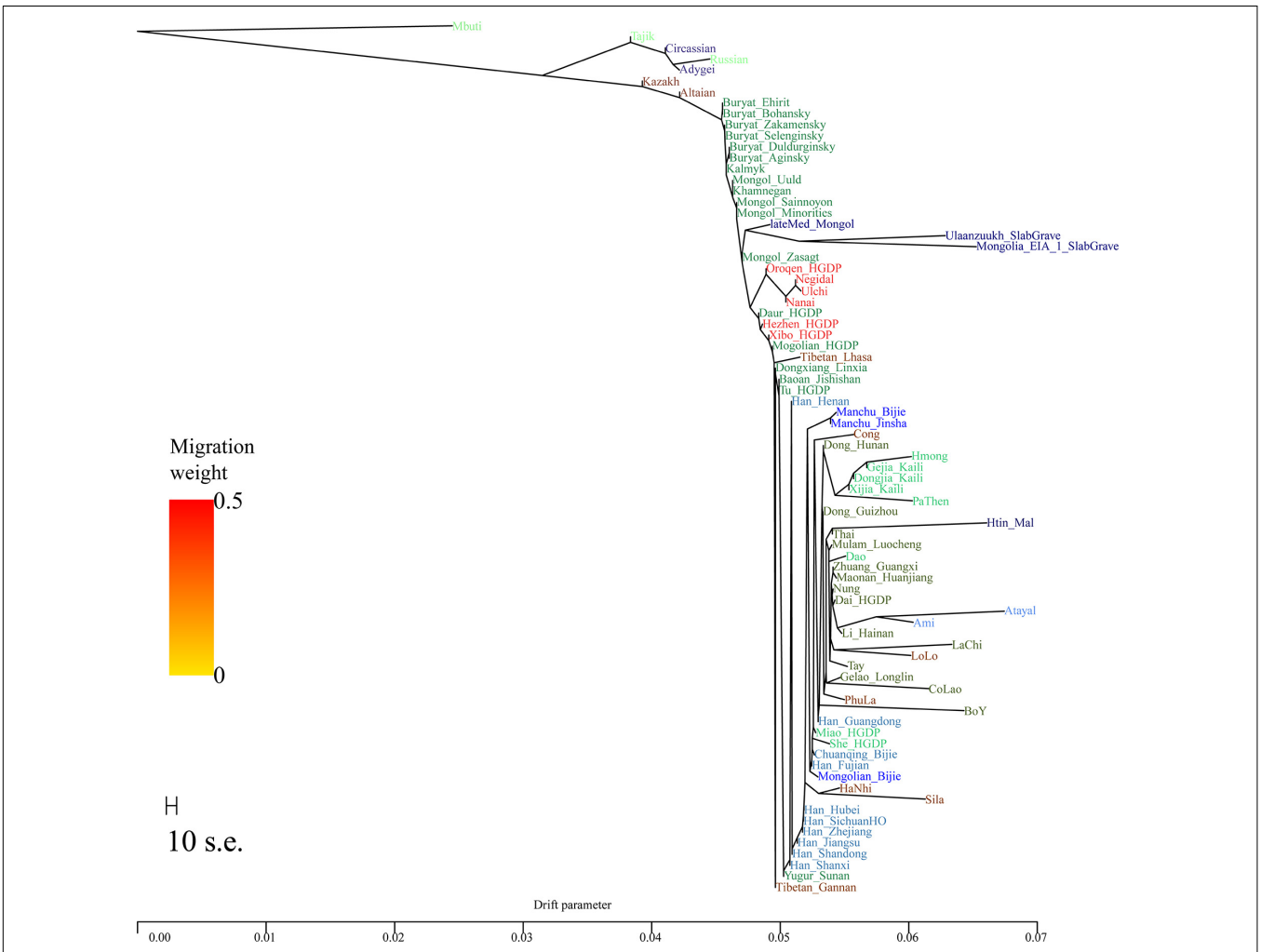

FIGURE 4 | Phylogenetic relationship among northern Altaic, central Sino-Tibetan and southern Austronesian, Austroasiatic, Hmong-Mien, and Tai-Kadai speakers. Mbuti population from central Africa were set as the root. Different populations were marked according to their linguistic affinity.

the southern source. The reconstructed phylogeny showed that the genetic contribution of the ancient northern East Asians to the Bijie Mongolians is 44%, whereas the proportion from the southern East Asians is approximately 56% (**Figure 5**).

We next used ALDER software to estimate when the admixture occurred. We tried different modern populations from the north and south of East Asia as possible ancestral groups. We observed that most of the average time that admixture

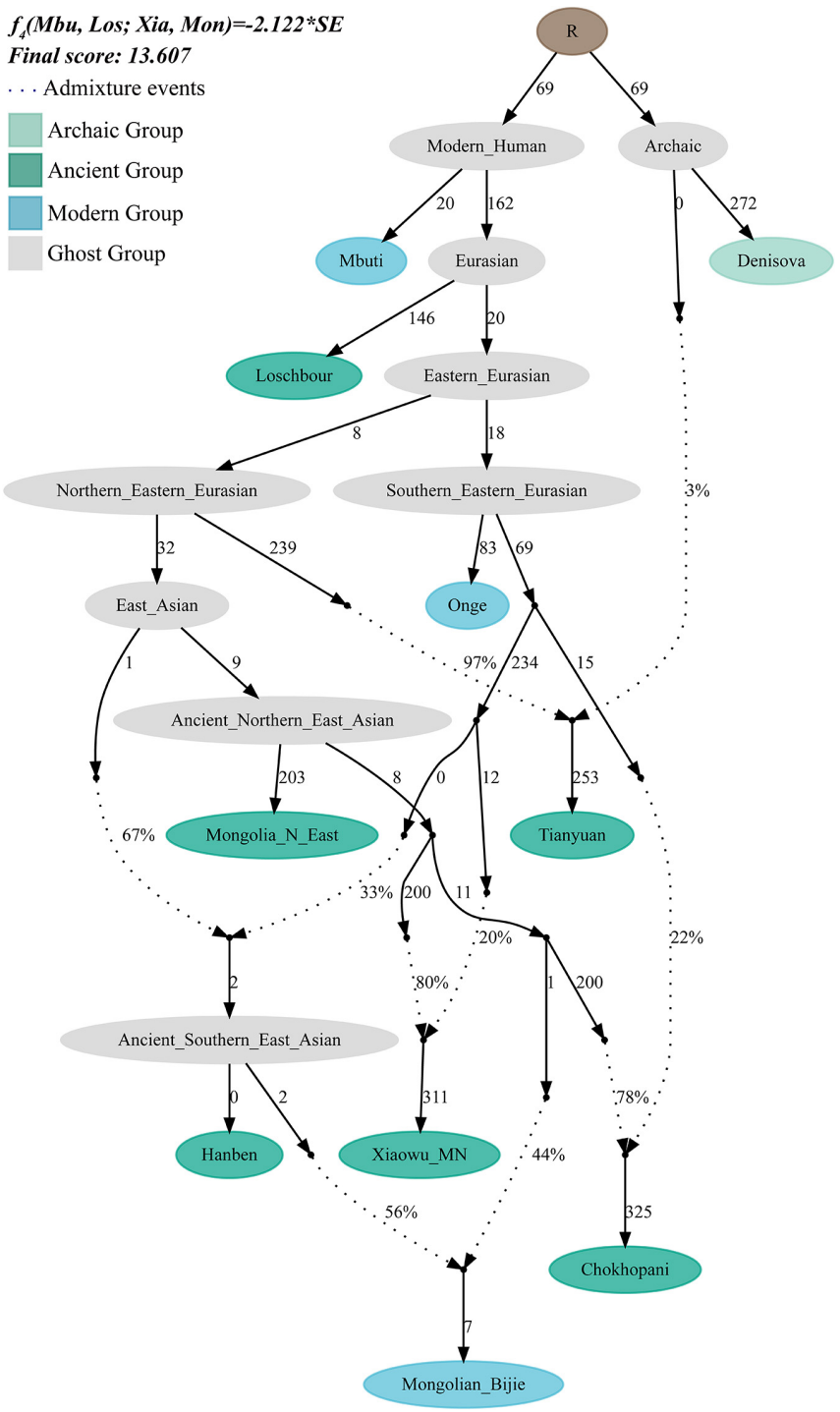

FIGURE 5 | The suggested admixture model of southern Mongolian people *via qpGraph*. The merged 1240K dataset was used. Dotted line denotes the admixture events, and their corresponding admixture proportions also marked. One hundred times of genetic drift (f_2 values) were denoted. Ancient populations, modern targeted, and ghost populations were color-coded.

occurred is around 1,000 AD, which is concordant with the historically documented expansion of the Mongol Empire and the establishment of the Yuan Dynasty (**Figure 6**).

We successfully obtained 62 uniparental Y-chromosome lineages and 81 mtDNA lineages. Among 55 studied Manchus samples, we identified 37 maternal lineages with terminal lineage frequencies ranging from 0.0182 to 0.0727 (B4g:4, F1a:4, F1a1:4), and B4, B5a, F, D4, and M7 were the dominant maternal lineages. We obtained 14 terminal paternal lineages among 43 males with frequencies ranging from 0.0233 to 0.3953 (O1b1a2a1-F1759/F2064/CTS5847/CTS8414/Z24393/F3314/F3323/CTS11890/F3478′: 17). We also identified some Manchus samples with paternal haplogroup C2c1b7∼-Z45293′. For the studied Mongolians, we identified 23 different maternal lineages with frequencies ranging from 0.0384 to 0.0769 (M7b1a1e1: 3), with A5b1, B5a1c1, and M7b1a1 identified at least twice among the Mongolian samples. The high-frequency paternal lineages of our Mongolian samples are O1b1a1a1a1a2a1-Z24050′ (11) and O1a1a2a1-Z23266 (6) (**Supplementary Table 10**). We also made population compassion among paternal and maternal lineages from ethnically and geographically Guizhou populations; population clustering patterns showed that Mongolic and Tungusic–speaking populations had a close relationship with geographically close populations, suggesting extensive population admixture occurred among them (**Supplementary Figures 2–3**).

DISCUSSION

Strong associations between population genetic structure and linguistic similarity were subsequently evidenced among Afroasiatic, Nilo-Saharan, Niger-Congo, and Khoisan language families in Africa (Martin et al., 2018; Patin and Quintana-Murci, 2018; Gurdasani et al., 2019), as well as language families in Asia (Chen et al., 2019; He et al., 2020a,b,c). Recent genome-wide modern and ancient DNA data have demonstrated that obvious population stratifications existed in East Asia with four regional dominant ancestries. The 7,000-year-old eastern Mongolians Neolithic people–related ancestry was widely distributed in modern Tungusic and Mongolic speakers in northern and northeastern China, Mongolia, and southern

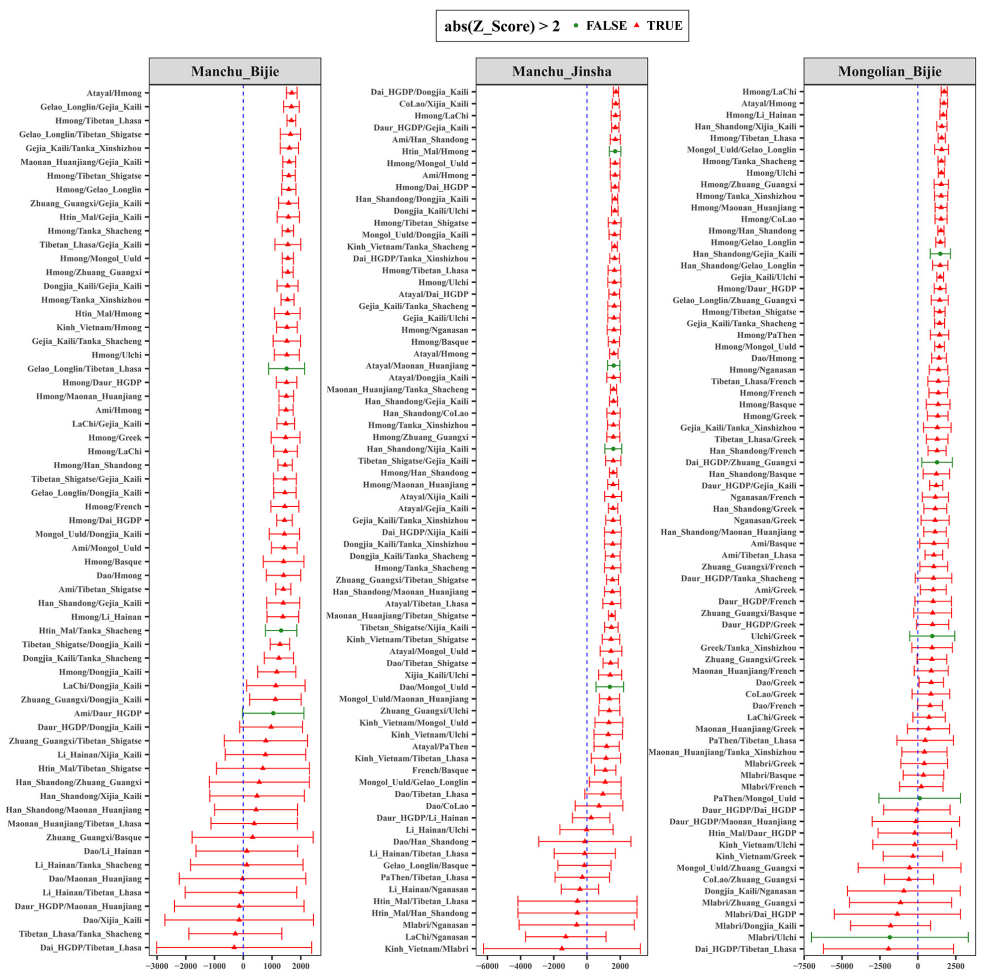

FIGURE 6 | ALDER-based admixture time with different ancestral sources. The length of 29 years of one generation was used.

Siberia (Ning et al., 2020; Wang C. C. et al., 2021). The Tibetan-related ancestry, which was represented by Neolithic Upper and Middle Yellow River farmers, was widely distributed in modern Tibetan-Burman–speaking populations and also a dominant component in Sinitic speakers (Jeong et al., 2016; Massilani et al., 2020; Zhang and Fu, 2020; Wang C. C. et al., 2021). For southern China and Southeast Asia, one ancestry component was widely distributed in Hmong-Mien–speaking populations mainly collected from Guizhou province and Vietnam (Lipson et al., 2018; McColl et al., 2018; Yang et al., 2020; Wang C. C. et al., 2021). The other southern ancestry was dominated in Austronesian-speaking populations (Lipson et al., 2018; McColl et al., 2018; Yang et al., 2020; Wang C. C. et al., 2021), also dominant in Tai-Kadai–speaking Li in Hainan island (He et al., 2020b). However, some exceptions were also identified in China, which may be caused by large-scale population movements and genetic admixture events in the recent and prehistoric time, for example, the East–West admixture along the Silk Road (Yao et al., 2021), and some western Eurasian ancestry was also identified in Iron Age Xinjiang people (Ning et al., 2019). Ancient genome data in East Asia also have illuminated three main Neolithic population expansions that have participated in the formation of modern observed distributed patterns of genetic structure and language families (Wang C. C. et al., 2021). Holocene population movements from the Amur River basin and eastern Mongolia Plateau were associated with the formation of the genetic structure of Mongolic and Tungusic–speaking populations. Similarly, population expansion from the Yellow River basin and the Yangtze River basin, respectively, contributed to the formation of Sino-Tibetan speakers (Wang et al., 2020) and other southern East Asians, as well as the Southeast Asians (Larena et al., 2021; Wang C. C. et al., 2021).

Here, we presented the fine-scale genetic structure of Mongolic and Tungusic–speaking populations (Mongolians and Manchus) in Guizhou and reconstructed their demographic history. We observed significant genetic differences between southern Mongolic and Tungusic speakers from Guizhou and their counterparts from northern East Asia (North China, Mongolia, and southern Siberia). We observed two different genetic clines among all Mongolic and Tungusic–speaking populations in the PCA plots, and Guizhou populations have deviated to the southern East Asian clusters comprising Austronesian, Austroasiatic, and Tai-Kadai populations, as well as close to Hmong-Mien clines. However, northern Mongolic and Tungusic speakers formed another genetic cluster that was located far away from the southern ones. We identified different ancestry components in northern and southern populations in the model-based ADMIXTURE results with the studied Guizhou populations sharing similar genetic profiles with southern East Asians. We observed suggestive evidence in f_3 statistics that Guizhou Manchus and Mongolians derived ancestry from both northern and southern East Asia. But for the northern Mongolic and Tungusic–speaking populations, we can find significant admixture signatures with one source from East Asians and the other from western Eurasians or northern Siberians. The genetic distance-related indexes (Fst and outgroup f_3 statistics) consistently supported the studied Guizhou populations having a strong southern East Asian affinity, but northern Mongolic and Tungusic speakers showing a clear northern East Asian affinity. We observed the Y-chromosome and mtDNA haplogroups in Guizhou Manchus and Mongolians are the lineages that are frequent in southern China, showing a different genetic profile from that in northern Mongolic and Tungusic speakers. Recent genetic studies focused on northern Mongolian and Manchu populations found that paternal lineages of C2a and C2b were widely distributed in these populations, which is rarely found in our focused Guizhou Manchus and Mongolians.

Furthermore, we also identified the genetic differences between studied Manchus and Mongolians with southern East Asians. Our studied Manchus and Mongolians did not group together with geographically close Guizhou populations, such as Guizhou Han, Chuanqing, Gejia, Gongjia, and Xijia. Compared with southern East Asians, Guizhou Manchus and Mongolians shared excess alleles with northern Mongolic/Tungusic–speaking populations, as shown in significant negative f_4 values in f_4 (southern East Asians, studied Guizhou populations; northern East Asians, Mbuti). The *qpGraph*-based phylogeny with admixture events further showed a large proportion of the ancestry of Guizhou Mongolians derived from Yellow River farmers, who were genetically close to Mongolians Neolithic populations. The ALDER-based estimates of admixture times ranged from 500 to 1,500 years ago, which was consistent with the time of Mongolians Empire expansion and the formation of the Yuan dynasty. The excess affinity of Guizhou Manchus and Mongolians with northern populations, when compared with Guizhou Indigenous groups, highlights the role of the southern expansion of northern Mongolians.

Previous genetic, linguistic, and archeological documents from Guizhou and other southwestern China showed that Southwestern East Asia had the highest diversity in genetics, language, and culture. Thus, these complex mixture natures promote the admixture process between southward migrated Manchus and Mongolians and local populations. These strong genetic affinities also supported *via* genome-wide data or traditional genetic markers from southwestern populations (Chen et al., 2018a,b; He et al., 2019, 2020b,c, 2021). However, both of our ALDER-based admixture dates and historically documented migration history of Mongolians in the Yuan Dynasty and Manchus in the Qing Dynasty showed the plausible admixture events that occurred recently. Cultural anthropologies also showed these migrated populations had their specific lifestyles, language, and other customs. Besides, the relatively isolated resediment environments further confirmed some extent genetic isolation between Mongolians, Manchus, and other geographically close populations. It is interesting to identify the genetic affinity between our studied population and Hmong-Mien–speaking populations; one possible reason is that Hmong-Mien–speaking populations are the dominant Indigenous populations directly descended from the ancient Neolithic rice farmer in the middle Yangtze River and may be the direct descendants of the Daxi culture, which provided the typical ancestral component for modern southwestern populations and

is also the best surrogate source populations for our studied populations. Indeed, these admixture signatures can be identified *via* admixture-f_3 statistics. Further work should be focused on the whole-genome sequencing data of more Hmong-Mien, southern Mongolic and Tungusic, and ancient DNA data from the higher time-transect to comprehensively characterize the fine-scale demographic history of southern Manchus and Mongolians and other Southeastern Asians.

CONCLUSION

We presented the first batch of genome-wide data focusing on the southern Mongolians and Manchus from Guizhou province. We used comprehensive population genetic analyses of PCA, ADMIXTURE, *qpAdm*, *qpWave*, *qpGraph*, and ALDER to explore the complex genetic history and dynamic admixture process of southwestern Chinese populations. We identified one unique genetic cline forming by our studied Mongolians and Manchus samples, which was close to the southern Hmong-Mien cline and Austronesian/Austroasiatic cline but distinct with northern Mongolic and Tungusic cline, suggesting southern Mongolians and Manchus people have experienced a differentiated demographic history since their separation from northern groups. Furthermore, allele-shared–based analysis from *f* statistics revealed that significant admixture occurred in Guizhou Manchus and Mongolians; results from admixture models demonstrated that Guizhou Mongolic and Manchus people harbored both northern ancestry and also additional gene fluxes from southern East Asians. Finally, estimates of ALDER-based admixture times from historic times demonstrated that the presented-day genetic structure observed here was caused by the massive southward migration of Mongolians empire expansion, which is consistent with the historically documented migration events.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://zenodo.org/record/4632918, doi: 10.5281/zenodo.463291.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Medical Ethics Committee of Guizhou Medical University and Xiamen University (Approval Number: XDYX2019009). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

C-CW and JH designed this study. JC, GH, and C-CW wrote the manuscript. QW, ZR, HLZ, YL, MY, JJ, and JH collected the samples. QW, ZR, HZ, JJ, YL, MY, JC, and JH conducted the experiment. JZ, GH, JG, XY, JC, KZ, RW, HM, and C-CW analyzed the data. All authors reviewed the manuscript.

FUNDING

This work was funded by the Guizhou Scientific Support Project, Qian Science Support (2021) General 448; Shanghai Key Lab of Forensic Medicine, Key Lab of Forensic Science, Ministry of Justice, China (Academy of Forensic Science), Open Project, KF202009; Guizhou Province Education Department, Characteristic Region Project, Qian Education KY No. (2021)065; Guizhou "Hundred" High-level Innovative Talent Project, Qian Science Platform Talents (2020)6012; Guizhou Scientific Support Project, Qian Science Support (2020)4Y057; Guizhou Science Project, Qian Science Foundation (2020)1Y353; Guizhou Scientific Support Project, Qian Science Support (2019)2825; Guizhou Scientific Cultivation Project, Qian Science Platform Talent (2018)5779-X; Guizhou Engineering Technology Research Center Project, Qian High-Tech of Development and Reform Commission No. (2016)1345; National Natural Science Foundation of China (NSFC 31801040), the Nanqiang Outstanding Young Talents Program of Xiamen University (X2123302), the Fundamental Research Funds for the Central Universities (ZK1144), European Research Council (ERC) grant (ERC-2019-ADG-883700-TRAM).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.685285/full#supplementary-material

Supplementary Figure 1 | Geographical location of newly collected samples.

Supplementary Figure 2 | Principal component analyses (PCA) among 18 ethnic groups in Guizhou based on the Y-chromosome haplogroup.

Supplementary Figure 3 | Principal component analyses (PCA) among 18 ethnic groups in Guizhou based on mitochondrial DNA haplogroup.

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

Bai, H., Guo, X., Narisu, N., Lan, T., Wu, Q., Xing, Y., et al. (2018). Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and East Asia. *Nat. Genet.* 50, 1696–1704. doi: 10.1038/s41588-018-0250-5

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7.

Chen, P., He, G., Zou, X., Wang, M., Luo, H., Yu, L., et al. (2018a). Genetic structure and polymorphisms of Gelao ethnicity residing in

southwest china revealed by X-chromosomal genetic markers. *Sci. Rep.* 8:14585.

Chen, P., He, G., Zou, X., Zhang, X., Li, J., Wang, Z., et al. (2018b). Genetic diversities and phylogenetic analyses of three Chinese main ethnic groups in southwest China: a Y-Chromosomal STR study. *Sci. Rep.* 8:15339.

Chen, P., Wu, J., Luo, L., Gao, H., Wang, M., Zou, X., et al. (2019). Population genetic analysis of modern and ancient DNA variations yields new insights into the formation, genetic structure, and phylogenetic relationship of Northern Han Chinese. *Front. Genet.* 10:1045. doi: 10.3389/fgene.2019.01045

Damgaard, P. B., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., Korneliussen, T., et al. (2018). 137 ancient human genomes from across the Eurasian steppes. *Nature* 557, 369–374.

de Barros Damgaard, P., Martiniano, R., Kamm, J., Moreno-Mayar, J. V., Kroonen, G., Peyrot, M., et al. (2018). The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* 360:eaar7711.

Gurdasani, D., Carstensen, T., Fatumo, S., Chen, G., Franklin, C. S., Prado-Martinez, J., et al. (2019). Uganda genome resource enables insights into population history and genomic discovery in Africa. *Cell* 179, 984–1002.e36.

He, G., Liu, J., Wang, M., Zou, X., Ming, T., Zhu, S., et al. (2021). Massively parallel sequencing of 165 ancestry-informative SNPs and forensic biogeographical ancestry inference in three southern Chinese Sinitic/Tai-Kadai populations. *Forensic Sci. Int. Genet.* 52:102475. doi: 10.1016/j.fsigen.2021.102475

He, G., Wang, M., Li, Y., Zou, X., Yeh, H. Y., Tang, R., et al. (2020a). Fine-scale north-to-south genetic admixture profile in Shaanxi Han Chinese revealed by genome-wide demographic history reconstruction. *J. Syst. Evol.* doi: 10.1111/jse.12715

He, G., Wang, Z., Guo, J., Wang, M., Zou, X., Tang, R., et al. (2020b). Inferring the population history of Tai-Kadai-speaking people and southernmost Han Chinese on Hainan Island by genome-wide array genotyping. *Eur. J. Hum. Genet.* 28, 1111–1123. doi: 10.1038/s41431-020-0599-7

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.

He, G. L., Li, Y. X., Wang, M. G., Zou, X., Yeh, H. Y., Yang, X. M., et al. (2020c). Fine-scale genetic structure of Tujia and central Han Chinese revealing massive genetic admixture under language borrowing. *J. Syst. Evol.* 59, 1–20.

Jeong, C., Ozga, A. T., Witonsky, D. B., Malmstrom, H., Edlund, H., Hofman, C. A., et al. (2016). 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.* 113, 7485–7490. doi: 10.1073/pnas.1520844113

Jeong, C., Wang, K., Wilkin, S., Taylor, W. T. T., Miller, B. K., Bemmann, J. H., et al. (2020). A dynamic 6,000-year genetic history of eurasia's eastern steppe. *Cell* 183, 890–904.e29.

Kilinc, G. M., Kashuba, N., Koptekin, D., Bergfeldt, N., Donertas, H. M., Rodriguez-Varela, R., et al. (2021). Human population dynamics and Yersinia pestis in ancient northeast Asia. *Sci. Adv.* 7:eabc4587. doi: 10.1126/sciadv.abc4587

Larena, M., Sanchez-Quinto, F., Sjodin, 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. U. S. A.* 118:e2026132118.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513, 409–413.

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, B. L., Ma, P. C., Wang, C. Z., Yan, S., Yao, H. B., Li, Y. L., et al. (2020). Paternal origin of Tungusic-speaking populations: insights from the updated phylogenetic tree of Y-chromosome haplogroup C2a-M86. *Am. J. Hum. Biol.* 33:e23462.

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

Martin, A. R., Teferra, S., Moller, M., Hoal, E. G., and Daly, M. J. (2018). The critical needs and challenges for genetic architecture studies in Africa. *Curr. Opin. Genet. Dev.* 53, 113–120. doi: 10.1016/j.gde.2018.08.005

Massilani, D., Skov, L., Hajdinjak, M., Gunchinsuren, B., Tseveendorj, D., Yi, S., et al. (2020). Denisovan ancestry and population history of early East Asians. *Science* 370:579. doi: 10.1126/science.abc1166

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528, 499–503. doi: 10.1038/nature16152

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.

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.

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, 2526–2532e2524.

Patin, E., and Quintana-Murci, L. (2018). The demographic and adaptive history of central African hunter-gatherers and farmers. *Curr. Opin. Genet. Dev.* 53, 90–97. doi: 10.1016/j.gde.2018.07.008

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

Pugach, I., Matveev, R., Spitsyn, V., Makarov, S., Novgorodov, I., Osakovsky, V., et al. (2016). The complex admixture history and recent southern origins of siberian populations. *Mol. Biol. Evol.* 33, 1777–1795. doi: 10.1093/molbev/msw055

Raghavan, M., Degiorgio, M., Albrechtsen, A., Moltke, I., Skoglund, P., Korneliussen, T. S., et al. (2014a). The genetic prehistory of the New World Arctic. *Science* 345:1255832.

Raghavan, M., Skoglund, P., Graf, K. E., Metspalu, M., Albrechtsen, A., Moltke, I., et al. (2014b). Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* 505, 87–91. doi: 10.1038/nature12736

Raghavan, M., Steinrucken, M., Harris, K., Schiffels, S., Rasmussen, S., Degiorgio, M., et al. (2015). POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* 349:aab3884.

Rasmussen, M., Anzick, S. L., Waters, M. R., Skoglund, P., Degiorgio, M., Stafford, T. W. Jr., et al. (2014). The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* 506, 225–229.

Sikora, M., Pitulko, V. V., Sousa, V. C., Allentoft, M. E., Vinner, L., Rasmussen, S., et al. (2019). The population history of northeastern Siberia since the Pleistocene. *Nature* 570, 182–188.

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

Wang, C. C., Yeh, H. Y., Popov, A. N., Zhang, H. Q., Matsumura, H., Sirak, K., et al. (2021). Genomic insights into the formation of human populations in East Asia. *Nature* 591, 413–419.

Wang, M., Zou, X., Ye, H.-Y., Wang, Z., Liu, Y., Liu, J., et al. (2020). Peopling of Tibet Plateau and multiple waves of admixture of Tibetans inferred from both modern and ancient genome-wide data. *bioRxiv* [Preprint]. doi: 10.1101/2020.07.03.185884

Wang, Q., Zhao, L., Gao, C., Zhao, J., Ren, Z., Shen, Y., et al. (2021). Ethnobotanical study on herbal market at the Dragon Boat Festival of Chuanqing people in China. *J. Ethnobiol. Ethnomedicine* 17:19.

Wei, L. H., Huang, Y. Z., Yan, S., Wen, S. Q., Wang, L. X., Du, P. X., et al. (2017a). Phylogeny of Y-chromosome haplogroup C3b-F1756, an important paternal lineage in Altaic-speaking populations. *J. Hum. Genet.* 62, 915–918. doi: 10.1038/jhg.2017.60

Wei, L. H., Wang, L. X., Wen, S. Q., Yan, S., Canada, R., Gurianov, V., et al. (2018a). Paternal origin of Paleo-Indians in Siberia: insights from Y-chromosome sequences. *Eur. J. Hum. Genet.* 26, 1687–1696. doi: 10.1038/s41431-018-0211-6

Wei, L. H., Yan, S., Lu, Y., Wen, S.-Q., Huang, Y.-Z., Wang, L.-X., et al. (2018b). Whole-sequence analysis indicates that the Y chromosome C2*-Star Cluster traces back to ordinary Mongols, rather than Genghis Khan. *Eur. J. Hum. Genet.* 26, 230–237. doi: 10.1038/s41431-017-0012-3

Wei, L. H., Yan, S., Yu, G., Huang, Y. Z., Yao, D. L., Li, S. L., et al. (2017b). Genetic trail for the early migrations of Aisin Gioro, the imperial house of the Qing dynasty. *J. Hum. Genet.* 62, 407–411. doi: 10.1038/jhg.2016.142

Weir, B. S., and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358–1370. doi: 10.2307/2408641

Weissensteiner, H. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* 44, W58–W63.

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

Yang, M. A., Gao, X., Theunert, C., Tong, H., Aximu-Petri, A., Nickel, B., et al. (2017). 40,000-year-old individual from Asia provides insight into early population structure in Eurasia. *Curr. Biol.* 27, 3202–3208.e9.

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* doi: 10.1007/s00438-021-01767-0 [Epub ahead of print].

Yunusbayev, B., Metspalu, M., Metspalu, E., Valeev, A., Litvinov, S., Valiev, R., et al. (2015). The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genet.* 11:e1005068. doi: 10.1371/journal.pgen.1005068

Zhang, M., and Fu, Q. (2020). Human evolutionary history in Eastern Eurasia using insights from ancient DNA. *Curr. Opin. Genet. Dev.* 62, 78–84. doi: 10.1016/j.gde.2020.06.009

Zhang, Y., Wu, X., Li, J., Li, H., Zhao, Y., and Zhou, H. (2018). The Y-chromosome haplogroup C3*-F3918, likely attributed to the Mongol Empire, can be traced to a 2500-year-old nomadic group. *J. Hum. Genet.* 63, 231–238. doi: 10.1038/s10038-017-0357-z

Zhao, J., Wurigemule, Sun, J., Xia, Z., He, G., Yang, X., et al. (2020). Genetic substructure and admixture of Mongolians and Kazakhs inferred from genome-wide array genotyping. *Ann. Hum. Biol.* 47, 620–628. doi: 10.1080/03014460.2020.1837952

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 September 2021
doi: 10.3389/fgene.2021.725243

Peopling History of the Tibetan Plateau and Multiple Waves of Admixture of Tibetans Inferred From Both Ancient and Modern Genome-Wide Data

Guanglin He[1,2,3,4†], Mengge Wang[5,6,7†], Xing Zou[5,8†], Pengyu Chen[9], Zheng Wang[5], Yan Liu[10], Hongbin Yao[11], Lan-Hai Wei[3], Renkuan Tang[12*], Chuan-Chao Wang[2,3,4*] and Hui-Yuan Yeh[1*]*

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

Archeologically attested human occupation on the Tibetan Plateau (TP) can be traced back to 160 thousand years ago (kya) via the archaic Xiahe people and 30∼40 kya via the Nwya Devu anatomically modern human. However, the history of the Tibetan populations and their migration inferred from the ancient and modern DNA remains unclear. Here, we performed the first ancient and modern genomic meta-analysis among 3,017 Paleolithic to present-day Eastern Eurasian genomes (2,444 modern individuals from 183 populations and 573 ancient individuals). We identified a close genetic connection between the ancient-modern highland Tibetans and lowland island/coastal Neolithic Northern East Asians (NEA). This observed genetic affinity reflected the primary ancestry of high-altitude Tibeto-Burman speakers originated from the Neolithic farming populations in the Yellow River Basin. The identified pattern was consistent with the proposed common north-China origin hypothesis of the Sino-Tibetan languages and dispersal patterns of the northern millet farmers. We also observed the genetic differentiation between the highlanders and lowland NEAs. The former harbored more deeply diverged Hoabinhian/Onge-related ancestry and the latter possessed more Neolithic southern East Asian (SEA) or Siberian-related ancestry. Our reconstructed *qpAdm* and *qpGraph* models suggested the co-existence of Paleolithic and Neolithic ancestries in the Neolithic to modern East Asian highlanders. Additionally, we found that Tibetans from Ü-Tsang/Ando/Kham regions showed a strong population stratification consistent with their cultural background and geographic terrain. Ü-Tsang

OPEN ACCESS

Edited by:
Horacio Naveira,
University of A Coruña, Spain

Reviewed by:
Yan Li,
Yunnan University, China
Jiang Huang,
Guizhou Medical University, China

***Correspondence:**
Guanglin He
Guanglinhescu@163.com
Renkuan Tang
renktang2012@163.com
Chuan-Chao Wang
wang@xmu.edu.cn
Hui-Yuan Yeh
hyyeh@ntu.edu.sg

† 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 June 2021
Accepted: 12 August 2021
Published: 03 September 2021

Citation:
He G, Wang M, Zou X, Chen P, Wang Z, Liu Y, Yao H, Wei L-H, Tang R, Wang C-C and Yeh H-Y (2021) Peopling History of the Tibetan Plateau and Multiple Waves of Admixture of Tibetans Inferred From Both Ancient and Modern Genome-Wide Data. Front. Genet. 12:725243. doi: 10.3389/fgene.2021.725243

Tibetans possessed a stronger Chokhopani-affinity, Ando Tibetans had more Western Eurasian related ancestry and Kham Tibetans harbored greater Neolithic southern EA ancestry. Generally, ancient and modern genomes documented multiple waves of human migrations in the TP's past. The first layer of local hunter-gatherers mixed with incoming millet farmers and arose the Chokhopani-associated Proto-Tibetan-Burman highlanders, which further respectively mixed with additional genetic contributors from the western Eurasian Steppe, Yellow River and Yangtze River and finally gave rise to the modern Ando, Ü-Tsang and Kham Tibetans.

Keywords: East Asian, genetic history, Sino-Tibetan, Tibetan Plateau, ancient genomes

INTRODUCTION

The Tibetan Plateau (TP), widely known as the third pole of the world, forms the high-altitude core region of Asia with an average elevation more than 4,000 meters above sea level (masl). The TP represents one of the most challenging environments for human settlements due to the perennial low temperature, extreme aridity, and severe hypoxia. However, archeological and genetic studies have demonstrated that archaic hominins who occupied the TP had well adapted to the high-altitude hypoxic environment long before the arrival of modern *Homo sapiens*. The present-day Tibetans are suggested to have uniquely adapted to the extreme high-altitude conditions since the initial colonization of the TP (Qi et al., 2013; Jeong et al., 2016; Gnecchi-Ruscone et al., 2018; Chen F. et al., 2019). However, recent linguistic evidence suggested that Tibeto-Burman populations diverged from Han Chinese approximately 5.9 thousand years ago (kya) (Zhang et al., 2019). At present, over seven million indigenous Tibetans (2016 census) are living in the TP and have successfully adapted to the high-altitude hypoxic environment. Genomic analysis found multiple variants that may jointly deliver the high-altitude fitness of the modern Tibetans which is missing in the Hans (Yi et al., 2010). For example, the positively selected haplotypes of *HIF-1α prolyl hydroxylase1 (EGLN1)* and *Endothelial PAS domain protein 1 (EPAS1)* were introduced into modern Tibetans and surrounding highlanders via the Denisovan introgression, which further promoted Tibetan's high-altitude hypoxia adaptation (Huerta-Sánchez et al., 2014). Compared to the well-established population prehistory in other parts of East Asia (He et al., 2020; Ning et al., 2020; Yang et al., 2020; Wang C. C. et al., 2021), the population history of the TP's was far from clear due to the lack of excavated archeological sites and human remains. For example, there are a limited amount of zooarchaeological and archaeobotanical data for reconstructing the subsistence strategy and ancient DNA (aDNA) data for dissecting the genomic correlation between ancient individuals and modern Tibetan-like highlanders.

To date, when, where, and how the early human colonizers conquered the TP, and who were the ancestors of the modern Tibetans remain unanswered. Archeological, paleo-anthropological, and genetic studies focusing on the peopling processes of the TP and demographic history of Tibetan Highlanders are still in developmental stages (Aldenderfer, 2011). As revealed by the archeological evidence, handprints and footprints of *Homo sapiens* found at the Quesang site in southern TP (4,200 masl) suggested that the intermittent human presence on the TP could trace back to at least 20 kya (Zhang and Li, 2002), and the permanent human occupation was dated to the early Holocene (Meyer et al., 2017). The Nwya Devu site, located nearly 4,600 masl in Central Tibet, could be dated to at least 30 kya, which deepened considerably the history of the peopling of the TP and the antiquity of human high-altitude adaptations (Zhang et al., 2018). The palaeo-proteomic analysis of a Xiahe Denisovan mandible indicated that the prehistoric colonization of archaic hominins on the TP could be traced back to the Middle Pleistocene epoch (around 160 kya) (Chen F. et al., 2019). This Pleistocene colonization of archaic humans was recently evidenced via the Denisovan type of mtDNA found in Xiahe site (Zhang et al., 2020). Additionally, modern human genomic data also provided supporting evidence that humans did exist on the TP before the Last Glacial Maximum (LGM), and the genetic relics of the Upper Paleolithic inhabitants in modern Tibetans indicated some extent of genetic continuity between the initial Paleolithic settlers and modern Tibetan highlanders (Zhao et al., 2009; Qin et al., 2010; Qi et al., 2013; Li et al., 2015; Lu et al., 2016). The archaeogenetic investigation of prehistoric Himalayan populations provided supporting evidence for the high-elevation East Asian origin of the first inhabitants of the Himalayas, indirectly indicating the pre-Neolithic human activities on the TP (Jeong et al., 2016).

In contrast to the Late Pleistocene Hunter-Gatherer colonization, the timing and dynamics of the Holocene permanent human occupation of the TP have also provoked many debates (Ding et al., 2020; Liu W. et al., 2020). Recent archeological and genomic findings suggested that the permanent settlement on the TP was a relatively recent occurrence along with the establishment of farming and pastoralism on the Plateau (Chen et al., 2015; Li et al., 2019). Chen et al. reported archaeobotanical and zooarchaeological data from 53 archeological sites in the northeastern TP (NETP) and illustrated that the novel agropastoral subsistence strategy facilitated year-round living on the TP after 3.6 kya (Chen et al., 2015). The first comprehensive and in-depth genomic investigation of the Tibet sheep also revealed a stepwise pattern of recent permanent human occupation on the TP through the Tang-Bo Ancient Road (from northern China to the NETP ∼3,100 years ago and from the NETP to southwestern areas of the TP ∼1,300 years ago) (Hu et al., 2019). However, it remains unknown who brought the

cold-tolerant barley agriculture and livestock to the TP, and how indigenous foragers interacted with the incoming farmers. The archeological observations demonstrated that incoming farmer groups did not replace the local foragers, but co-existed with them for extended periods (Gao et al., 2020; Ren et al., 2020). The mitochondrial evidence and radiocarbon dates of the cereal remains also revealed that millet farmers adopted and brought barley agriculture to the TP around 3.6–3.3 kya. Contemporary Tibetans could trace their main ancestry back to the Neolithic millet farmers (Li et al., 2019). Moreover, the genetic variations of modern Tibetan groups have also been explored based on the forensically available markers (Wang Z. et al., 2018; Zou et al., 2018; He et al., 2019). However, the low resolution of these markers hindered the comprehensive understanding of prehistoric human activities on the TP and impeded the dissection of the ancestral component of Tibetans. Lu and Zhang et al. conducted a series of typical population genomic studies focusing on the demographic history of modern Tibetans and other high-altitude highlanders (Lu et al., 2016; Zhang et al., 2017). They found that Tibetans arose from a mixture of multiple ancestral genetic sources with the co-existence of Paleolithic and Neolithic ancestries.

Collectively, previous studies paved the way toward a better understanding of the Middle Pleistocene arrival, Paleolithic colonization and Neolithic permanent settlement on the TP. However, most of the previous archeological investigations have primarily focused on the NETP (< 4000 masl). Besides, the lack of discussion of ancient samples from the TP and incomprehensive analysis of ancient/modern individuals from East Asia hindered our ability to spatiotemporally connect dispersed ancient East Asians and modern Tibetans. Thus, we comprehensively meta-analyzed the genetic variations of ancient/modern highlanders from the TP and surrounding lowland eastern Eurasians with the aims to (I) portray the genetic landscape of the East Asian highlanders, (II) study the genetic similarities and differences between highlanders and lowlanders, (III) explore the genetic substructure among geographically/culturally different Tibetans, (IV) reconstruct their deep evolutionary history and the corresponding migration and admixture processes. By analyzing genome-wide data of modern Tibetans and Neolithic-to-historic individuals from East Asia, we shed light on the genetic transition, turnover or continuity, ancestral composition, and demographic history of Tibetan highlanders.

MATERIALS AND METHODS

Publicly Available Dataset

We collected 2,444 individuals from 183 geographically/culturally different populations (Patterson et al., 2012; Lipson et al., 2018a; Jeong et al., 2019; Liu D. et al., 2020) belonging to fifteen language families or groups: Altai (also referred to as Trans-Eurasian including Mongolic, Japonic, Koranic, Tungusic, and Turkic), Sino-Tibetan (Sinitic and Tibeto-Burman), Hmong-Mien, Austronesian, Austroasiatic, Uralic, Caucasian, Chukotko-Kamchatkan, Eskimo-Aleut, Indo-European and Tai-Kadai. The 383 modern East Asian individuals genotyped via the Affymetrix Human Origins array were also used here (Wang C. C. et al., 2021). To explore the genomic history of modern Tibetans and elucidate the peopling process of the TP, we focused on the genome-wide data of 98 modern Tibetans collected from eleven geographically different regions with different cultural backgrounds, which includes five Ü-Tsang Tibetan groups from Tibet Autonomous Region, three Ando Tibetan groups from Qinghai and Gansu, four Kham Tibetan groups from Sichuan, Yunnan, and Tibet (**Figure 1A**). Raw data were quality-controlled using the PLINK v.1.9 (Chang et al., 2015) following the standard threshold (Wang C. C. et al., 2021; Yao et al., 2021). Besides, Paleolithic-to-historic published ancient genomes from East Eurasia (Russia, China, Mongolia, Nepal and Southeast Asia) were collected from recent ancient DNA studies or from Allen Ancient DNA Resource (AADR) released by Reich Lab (Jeong et al., 2016; Yang et al., 2017, 2020; Ning et al., 2020; Wang C. C. et al., 2021). A total of 161 Paleolithic to historic East Asians and eight Nepal ancients were collected and first comprehensively meta-analyzed and discussed (Jeong et al., 2016; Yang et al., 2017, 2020; Ning et al., 2020; Wang C. C. et al., 2021). Detailed information of key ancient populations is presented in **Table 1**.

Principal Component Analysis

We performed principal component analysis (PCA) with the *smartpca* program of the EIGENSOFT package (Patterson et al., 2006) using the default settings with additional parameters: lsqproject: YES and numoutlieriter: 0. Population data of modern East Asia were used to reconstruct the genetic background of PCA, in which modern samples were mainly sampled from Altaic, Sino-Tibetan, Hmong-Mien, Austronesian, Austroasiatic, and Tai-Kadai language families. Ancient genomes were projected onto the first two components. The projected ancient populations included eight individuals from Nepal (Jeong et al., 2016) (Chokhopani, Samdzong, and Mebrak cultures), eighty-four samples from the Yellow River (Ning et al., 2020; Yang et al., 2020; Wang C. C. et al., 2021), Amur River and West Liao River in the coastal and inland northern East Asia (including Houli, Yangshao, Longshan, Qijia, Hongshan, Yumin and other cultures), fifty-eight individuals (Ning et al., 2020; Yang et al., 2020; Wang C. C. et al., 2021) belonging to Tanshishan and other cultures in the coastal southeast East Asia (Fujian and Taiwan).

F_{ST} Calculation and TreeMix Analysis

We used the Plink 1.9 and an in-house script to estimate the pairwise F_{ST} genetic distance (Purcell et al., 2007) among 82 modern populations with a sample size large than five. We also calculated F_{ST} values among 31 ancient populations. We ran TreeMix v.1.13 (Pickrell and Pritchard, 2012) with migration events ranging from 0 to 8 to construct the topology among eastern Eurasians with the maximum likelihood tree.

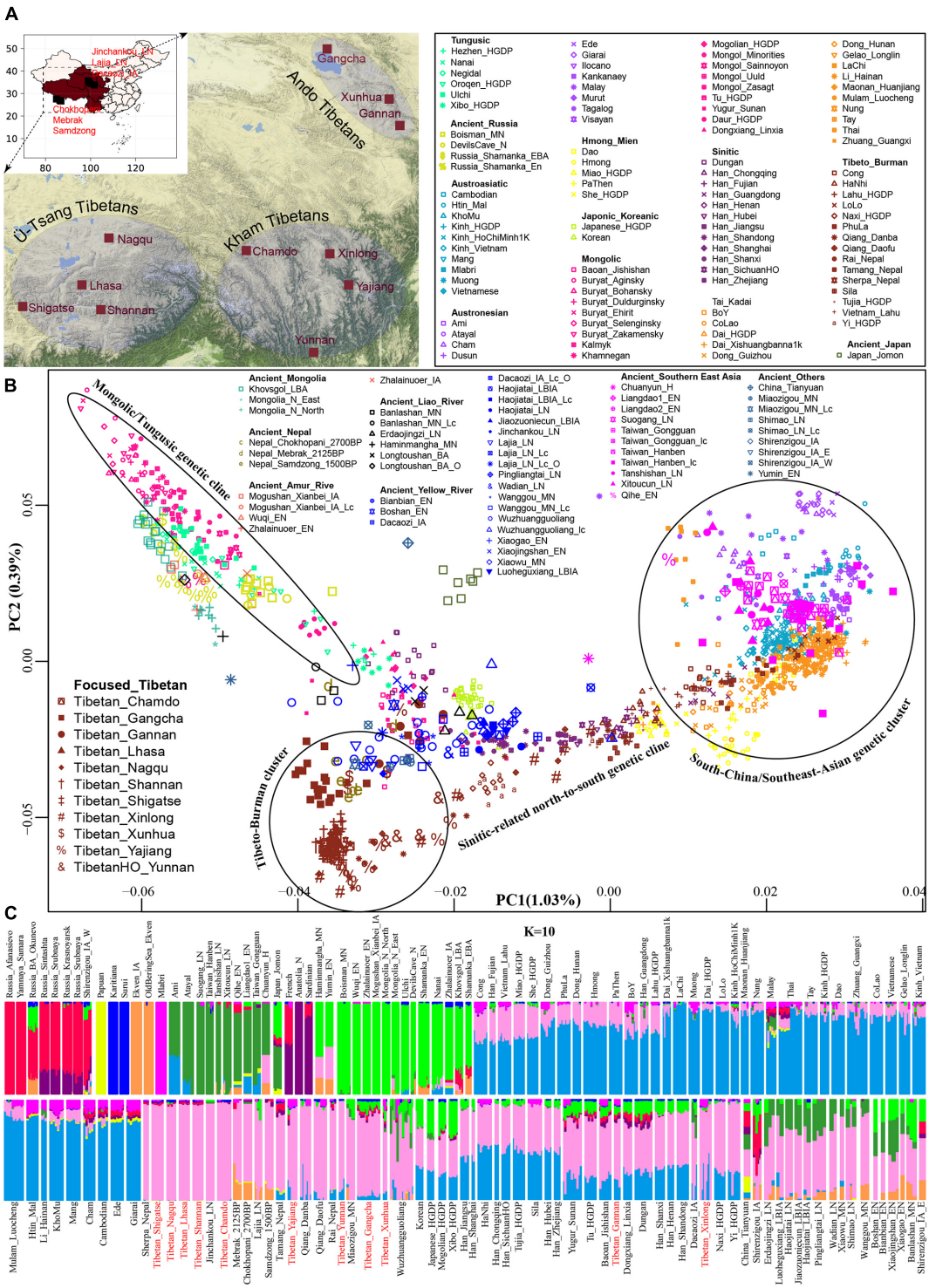

FIGURE 1 | The geographical position of the focused Tibetans and genetic patterns of East Asians. **(A)** Sampling place of eleven geographically different modern Tibetan populations mainly discussed in the present study from the five provinces (Tibet Tibetan Autonomous Region, Qinghai, Gansu, Sichuan, and Yunnan) from western China. China map is presented in the top-left of **A** and five studied western provinces were zoom-in as the presented Google map. **(B)** Principal component analysis (PCA) showed the genetic similarities and differences between the ancient/modern East Asians from geographically/linguistically/culturally different populations. Spatial-temporally diverse ancient populations were projected onto the two-dimensional genetic background of modern East Asians. **(C)** Admixture ancestry estimation based on the model-based ADMIXTURE. Here, the optimal predefined ten ancestral populations were used. EN, early Neolithic; MN, middle Neolithic; LN, late Neolithic; IA, iron age; BA, bronze Age; LBIA, late bronze age and iron age; Lc, loc coverage; O, outlier.

TABLE 1 | The detailed information of included ancient Chinese populations.

Ancient populations	Time period	Sample size	Archeological site	Testing platform	Y haplogroup types	MtDNA haplogroup types	Reference
China_Xinjian_IA	Iron Age	11	Shirenzigou	Shotgun	O2a, Q1a1, Q1a2a2b1∼, R1a1a1, R1b and R1b2b2	A17, D4j1b, G3b, H15b1, I1b, T1a1b, U4, U4′9, U5a2 and U5b2c	Ning et al., 2019
Wuzhuangguoliang_LN	Late Neolithic	12	Wuzhuangguoliang	Exome.capture	F(3), C	A7, A + 152 + 16362, B4a2b1, B4a4 (2), D4q, D5a3, F1g1, G2a1, M11a and R11a	Wang C. C. et al., 2021
China_AR_EN		2	Wuqi	1240K	.	C4a1 and C5	Ning et al., 2020
China_AR_IA	Iron Age	1	Zhalainuoer	1240K	.	N9a9	Ning et al., 2020
China_AR_Xianbei_IA	Iron Age	3	Mogushan Xianbei	1240K	.	C5a1, C4a1a4 and Z3a1	Ning et al., 2020
China_HMMH_MN	Middle Neolithic	1	Haminmangha	1240K	.	D4j	Ning et al., 2020
China_Miaozigou_MN	Middle Neolithic	3	Miaozigou	1240K	.	A14, C4a2a1 and D4b2	Ning et al., 2020
China_NEastAsia_Inland_EN	Early Neolithic	1	Yumin	1240K	.	.	Yang et al., 2020
China_SEastAsia_Coastal_EN	Early Neolithic	1	Qihedong	1240K	.	.	Yang et al., 2020
China_SEastAsia_Coastal_His	Historic	1	Chuanyundong	1240K	O1a1a1a1a1a1	.	Yang et al., 2020
China_SEastAsia_Coastal_LN	Late Neolithic	11	Xitoucun and Tanshishan	1240K	F, K2, NO, O, O1a2, O1b1a1a1a and O2a	.	Yang et al., 2020
China_SEastAsia_Island_EN	Early Neolithic	4	Liangdao	1240K	O1a, O and O2a2b	.	Yang et al., 2020
China_Shimao_LN	Late Neolithic	3	Shengedaliang	1240K	.	G2a1, D5a2a1b, M80′D	Ning et al., 2020
China_Upper_YR_IA	Iron Age	4	Dacaozi	1240K	.	D4b2b, G2b1b, F1g and Z3	Ning et al., 2020
China_Upper_YR_LN	Late Neolithic	7	Jinchankou and Lajia	1240K	.	B4c1b2c2, G3a2, A18, F1g, G1c1, F1a1a and F1g	Ning et al., 2020
China_WLR_BA		3	Longtoushan	1240K	.	D4m1, D4j14 and B4c1a2	Ning et al., 2020
China_WLR_LN	Late Neolithic	3	Erdaojingzi	1240K	.	B5b1a, A22 and N9a1	Ning et al., 2020
China_WLR_MN	Middle Neolithic	3	Banlashan	1240K	.	D5a3a1 and D5a3a1	Ning et al., 2020
China_YR_LBIA	Late Bronze Age/Iron Age	6	Haojiatai, Jiaozuoniecun, Luoheguxiang	1240K	.	M8a2b, B4d1′2′3, F4a2, C4a1a2 and A5b1b	Ning et al., 2020
China_YR_LN	Late Neolithic	8	Pingliangtai, Haojiatai and Wadian	1240K	.	D4b1a (3), D5a2a, F2h, N9a2, D4 and D4e1a	Ning et al., 2020
China_YR_MN	Middle Neolithic	8	Xiaowu and Wanggou	1240K	.	F1, M8a2, D4g2a1, B4d1 and C4a1a1	Ning et al., 2020
Nepal_Chokhopani_2700BP	Iron Age	1	Chokhopani	Shotgun	O2a2b1a1a1a4a1	D4j1b	Jeong et al., 2016
Nepal_Mebrak_2125BP	Iron Age	3	Mebrak	Shotgun	O2a2b1	Z3a1a, M9a1a1c1b1a and M9a1a2	Jeong et al., 2016
Nepal_Samdzong_1500BP	Historic	4	Samdzong	Shotgun	O2a2b1a1a1a4a1 (2), D1a1a1	M9a1a1c1b1a, M9a1a, M9a1a, F1c1a1a and F1d	Jeong et al., 2016
Russia_DevilsCave_N	Early Neolithic	6	Devil's Gate Cave	Shotgun	C2a	D4m (3) and D4(2)	Sikora et al., 2019
Taiwan_Gongguan	Late Neolithic	2	Gongguan	1240K	.	Y2a1	Wang et al., 2020
Taiwan_Hanben_IA	Iron Age	45	Hanben	1240K	O2a2b2a2b(4), O1a1a1a1(3), O1a1a1a(2), O1a2(2), F(2), O2a2b2b(1), O2a2a1a2(1), O(1), O1a1a1a1a1a(1), O2a2b(1), O1a(1)	F4b1(5), R(5), E1a1a1(4), F3b1a + 16093(4), B4a1a(3), E1a1a(3), D6a2(2), E1a(2), M7b1a2a1(2), E2a(2), F1a3a(1), B4b1a2(1), F3b1(1), E2b(1), R30(1), F3b1a2(1), R9c1b2(1), M7c1c3(1), F4b(1), F3b1a(1), B4b1a2f(1), E1a1(1) and B5a2a1 + 16129(1)	Wang et al., 2020
AR33K	Paleolithic	1	AR33K	1240K	.	.	Mao et al., 2021
Longlin	Paleolithic	1	Longlin	1240K	.	M71 + 151	Yang et al., 2020
China_Tianyuan	Paleolithic	1	Tianyuan	1240K	K2b	.	Yang et al., 2017
Russia_Altai.DG	Paleolithic	1	Altai	1240K	.	.	Reich et al., 2010

ADMIXTURE Analysis

We carried out the model-based clustering analysis using the ADMIXTURE (v.1.3.0) (Alexander et al., 2009) after pruning SNPs with a strong linkage disequilibrium via the PLINK v.1.9 (Chang et al., 2015) with the parameters of – indep-pairwise 200 25 0.4. We ran ADMIXTURE with the 10-fold cross-validation (−cv = 10). The predefined number of ancestral populations ranging from $K = 2$ to $K = 20$ with 100 bootstraps and different random seeds were used. We chose the best-fitted model with the minimum cross-validation errors. The smallest cross-validation error was obtained (0.4176) when we used 10 predefined ancestral sources.

F-Statistics and Admixture Modeling Graph

We conducted two different forms of the three-population tests using the *qp3Pop* program implemented in the ADMIXTOOLS (Reich et al., 2009; Patterson et al., 2012). Outgroup-f_3-statistics were performed in the form of f_3*(Reference Eurasians, targeted Tibetans; Mbuti)* to assess the shared genetic drift between our focused Tibetans and their reference populations. A central African population Mbuti was used as the outgroup. Admixture-f_3*(Surrogate population1, Surrogate population2; Targeted populations)* were performed to test whether our targeted population was an admixture of two sources related to our used surrogate populations. Negative f_3-values with a Z-score smaller than −3 indicated that two source populations were admixed to form the targeted populations. Four-population comparisons were conducted using *qpDstat* programs implemented in the ADMIXTOOLS (Reich et al., 2009; Patterson et al., 2012) with the additional parameter (f_4: YES) in three different forms. The first one was conducted in the form of f_4*(Tibetan1, Tibetan2; Eurasian reference, Mbuti)* to test whether two Tibetans form one clade relative to the used Eurasian reference. Non-statistically significant f_4 values showed two left populations formed one clade. Other two f_4-statistics in the forms f_4*(Eurasian, Source; Eurasian2, Mbuti)* and f_4*(Eurasian1, Eurasian2; Source, Mbuti)* were conducted to examine whether the used ancestral source shared more alleles with one of the Eurasians compared with others. We assessed standard errors using the weighted block jackknife approach. We next used the *qpGraph* program implemented in the ADMIXTOOLS (Reich et al., 2009; Patterson et al., 2012) to reconstruct the deep population history of modern Tibetans and other modern and ancient East Asians based on the combined results of the f_2, f_3 and f_4*-statistics*. The absolute Z-scores smaller than 3 indicated better-fitted models.

Streams of Ancestry and Inference of Mixture Proportions

We used the *qpWave/qpAdm* programs implemented in the ADMIXTOOLS (Haak et al., 2015) to estimate mixture coefficient and corresponding standard errors according to a basic set of outgroup populations: Mbuti, Ust_Ishim, Russia_Kostenki14, Papuan, Australian, Mixe, MA1, and Mongolia_N_East.

RESULTS

Close Genetic Affinity Between Ancient/Modern Tibetans With NEAs

Descriptive analyses of PCA and ADMIXTURE were first used to provide an overview of the genetic structure. All modern Tibetans and Neolithic-to-historic East Asians were grouped in the East-Asian genetic cline along with the second component in the Eurasian-PCA. To focus on the genetic variations of East Asians, we constructed East-Asian-PCA among 106 modern populations (**Figure 1B**) and found that modern East Asians grouped into four genetic clines or clusters: Mongolic/Tungusic genetic cline consisting of populations from northeast Asia; south-China/Southeast-Asian genetic cluster comprising of Austronesian, Austroasiatic, Tai-Kadai, and Hmong-Mien speakers; Sinitic-related north-to-south genetic cline, and Tibeto-Burman cluster, which were consistent with the linguistic/geographical divisions. Tibetan populations were grouped and showed a relatively close relationship with some of the Mongolic/Tungusic speakers in northern China, and they were also grouped closely with northern Han and other lowland Tibeto-Burman speakers. Focused on the population substructures within Tibetans, we further observed three different sub-clusters: the high-altitude Tibet-Ü-Tsang cluster (Lhasa, Nagqu, Shannan and Shigatse), Gan-Qing-Ando cluster in northeastern TP (Xunhua, Gangcha and Gannan) and Tibetan-Yi-corridor cluster (Chamdo, Xinlong, Yajiang and Yunnan), which were also consistent with the geographical positions of sampling places and cultural backgrounds.

We subsequently explored the patterns of genomic affinity between ancient populations and modern East Asians by projecting all included ancient individuals (243 eastern Eurasian ancients) onto the genetic background of modern populations. Here, we found four ancient population genetic clusters. Neolithic-to-historic SEAs (including Hanben and Gongguan from Taiwan, Late-Neolithic mainland Tanshishan and Xitoucun people) grouped together and clustered with modern Tai-Kadai, Austronesian, and Austroasiatic speakers. Neolithic-to-Iron Age NEAs (both coastal Shandong Houli and inland Yangshao, Longshan, and Qijia people) grouped together and were projected closely to the juncture position of three main East Asian genetic lines and the northmost end of Han Chinese genetic cline. We observed a close genetic relationship between early Neolithic Houli individuals associated with the main subsistence strategy of hunter-gathering and the Henan Middle/Late-Neolithic Yangshao/Longshan farmers, which indicated the genetic continuity in the Neolithic transition from foragers to millet farmers in the early Neolithic northern China. We also identified the subtle genetic differences within these Neolithic-to-Iron Age individuals from northern China. These Shandong Houli individuals were localized closely with modern Mongolic-speaking Baoan, Tu, Yugur, and Dongxiang, while the early Neolithic Xiaogao individuals were posited closely with modern Tungusic-speaking Hezhen and Xibo. All Shandong Neolithic ancient populations were localized distantly from the modern Shandong Han Chinese and shifted to modern northern Chinese minorities, which indicated that

modern northern Han received additional gene flow from SEA related ancestral lineage or ancient Houli individuals harbored more Siberian-associated ancestry. Late-Neolithic Longshan individuals (Pingliangtai, Haojiatai, and Wadian) and Bronze/Iron Age individuals (Haojiatai, Jiaozuoniecun, and Luoheguxiang) in Henan province were grouped together and shifted to the Han Chinese genetic cline and partially overlapped with Han Chinese from Shanxi and Shandong provinces. This observed genetic similarities among the Late Neolithic to present-day NEAs from the Central Plain (Henan, Shanxi, and Shandong) indicated a genetic stability in the core region of Chinese civilization since the Late-Neolithic period. Middle-Neolithic Yangshao individuals (Xiaowu and Wanggou) in Henan province grouped with some of the Wuzhuangguoliang_LN individuals collected from Shaanxi province and were shifted to more northern modern minorities. The inland Middle/Late-Neolithic NEAs from Shaanxi (Shimao), Inner Mongolia (Miaozigou) and upper Yellow River (Lajia and Jinchankou) clustered together and were shifted toward modern Tibetans and ancient Nepal samples (Mebrak, Samdzong and Chokhopani).

For ancient populations from the West Liao River, three genetic-affinity clusters could be identified in the projected PCA results: northern cluster (Haminmangha_MN and Longtoushan_BA_O) showed a genetic affinity with Shamanka and Mongolia Neolithic people; middle Hongshan cluster was localized between Mongolia minorities and modern Gangcha Tibetan; southern cluster (Upper Xiajiadian Longtoushan_BA and Erdaojingzi_LN) possessed close relationship with the Yellow River farmers, which suggested that both Neolithic ancients associated with steppe pastoralists from Mongolia Plateau and millet farmers from Yellow River Basin had participated in the formation of the Late Neolithic and subsequent populations in the West Liao River Basin. These population movements, interactions, and admixture processes have recently been fully elucidated by Ning et al. (2020). Here, we observed that the Late Neolithic populations in the southern cluster were localized between the coastal early Neolithic NEAs and inland Neolithic Yangshao and Longshan individuals, which indicated that millet farmers from the middle/lower Yellow Rivers (Henan and Shandong) had played an important role in the formation of Hongshan people or their descendants via both inland and coastal northward migration routes. For ancient populations from Mongolia Plateau, Russia Far East, Trans-Baikal-Region, and Amur River Basin, all included forty-six individuals (Neolithic-to-Bronze Shamanka, Mongolian, DevilsCave, Boisman, and others) clustered closely to modern Tungusic language speakers (Nanai and Ulchi) and also to some Mongolic speakers. Jomon individuals were grouped together in the intermediate position between the northern Russian coastal Neolithic people and southern Iron Age Taiwan Hanben and coastal Neolithic SEAs, but localized far away from modern Japanese populations.

Patterns of genetic relationship revealed from the top two components (extracting 1.42% variation: PC1: 1.03% and PC2: 0.39%) showed a genomic affinity between modern Tibetans, ancient Nepal populations, and ancient/modern East Asians and Siberians. To further explore the genetic structure and corresponding population relationships, we estimated the ancestry composition and cluster patterns according to the model-based maximizing likelihood clustering algorithm (**Figure 1C** and **Supplementary Figure 1**). We observed two northern and two southern East Asian dominant ancestries. The coastal NEA ancestry (light green) maximized in Neolithic northeast Asians (Boisman_MN, Wuqi_EN, Zhalainuoer_EN, Mongolia_N_North, Mongolia_N_East, DevilsCave_N and Shamanka_EN) and modern Tungusic speakers (Ulchi and Nanai). This light green ancestry also existed in the Bronze Age to present-day populations from northeastern China and Russia, and reached at a high proportion in the coastal Early Neolithic NEAs from Shandong. The other type of northern ancestry was enriched in modern highland Tibetans and Qijia culture-related Late Neolithic Lajia and Jinchankou populations, which also maximized in Nepal Bronze Age to historic individuals and ancient NEAs, as well as the lowland modern Sino-Tibetan speakers, inland Hmong-Mien and Tai-Kadai language speakers. We named this Tibetan-associated ancestry as inland NEA ancestry, which was the direct indicator of the close genetic affinity between Tibetan and ancient/modern NEAs. Dark green ancestry was enriched in the coastal Early Neolithic SEAs, Iron-Age Hanben, and modern Austronesian Ami and Atayal. Therefore, we referred to this dark green component as the coastal SEA ancestry. The blue component maximized in LaChi samples as the counterpart of the coastal ancestry that was widely distributed in Hmong-Mien and Tai-Kadai-speaking populations. This blue inland SEA ancestry also existed in the lowland Tibetans with a relatively high proportion in all Kham and Ando Tibetans except for Chamdo Tibetans. Besides, we found that Tibetans collected from the northeast TP harbored more coastal NEA ancestry. Some Austroasiatic-associated dark pink ancestry maximized in Mlabri also appeared in Yajiang, Xinlong Kham, and Xunhua and Gannan Ando Tibetans. The Steppe pastoralist-like red component was enriched in Bronze Age Afanasievo and Yamnaya, which was also identified in Qinghai and Gansu Ando Tibetans.

Population Differentiation Between Highland and Lowland East Asians and Substructure Among Tibetans

To further explore the genetic differentiation between eleven modern Tibetan populations and ancient/modern reference populations, we first calculated the pairwise F_{ST} genetic distances among 82 modern populations (**Supplementary Table 1**, modern dataset) and 32 ancient/modern populations (**Supplementary Table 2**, ancient dataset). We found a strong genetic affinity among geographically close populations. As shown in **Supplementary Figures 2, 3**, the high-altitude Tibetans from the south (Shigatse and Shannan), central (Lhasa), north or northeast (Nagqu and Chamdo) of Tibet Autonomous Region had the smallest F_{ST} genetic distances with their geographical neighbors, followed by lowland Ando Tibetans from the northeastern TP (Qinghai and Gansu) and the Kham Tibetans from the southeastern region of the TP (Sichuan

and Yunnan) and other Tibeto-Burman-speaking populations (Qiang, Tu and Yi). For Ando Tibetans from the Ganqing region, Gangcha Tibetan harbored a close genetic affinity with northern or northeastern Tibet Tibetans (Chamdo and Nagqu) with the smallest F_{ST} genetic distances, followed by Qiang, Yugur, and Tu or other geographically close Tibetans (**Supplementary Figure 4**). Different patterns were observed in Gangcha and Xunhua Tibetans, which showed the closest relationship with each other, and then followed by Tu and Yugur. We also found relatively small genetic distances between Tibetans (Gannan and Xunhua) and the Turkic-speaking Kazakh population, suggesting a western Eurasian affinity of Tibetans from the northeastern region of the TP relative to the Tibetans from the central region. **Supplementary Figure 5** presented the patterns of genetic differentiation between lowland Kham Tibetans and their reference populations. We found that Yajiang and Xinlong Tibetans from Sichuan province harbored a close genetic affinity with the geographically close populations (Tibetan, Qiang, Yugur and Tu). Yunnan Tibetans had the smallest genetic distance with Gangcha and Chamdo Tibetans, followed by Qiang, Yi, and Tu. Among Tibetans and Neolithic to Iron Age East Asians (**Supplementary Figure 6**), we also found Iron Age Hanben population from Taiwan and some southern Siberian ancients showed a closer relationship with modern Tibetans relative to other ancient East Asians. We should note there might be statistical bias in the F_{ST}-based analyses because of the different sample sizes in different populations.

Phylogenetic relationships were further reconstructed based on the genetic variations of modern Eurasian populations and ancient eastern Eurasians using TreeMix software based on genetic distances. As shown in **Figure 2**, a phylogenetic tree with no migration events showed that modern populations from similar language families tended to cluster into one clade. Altaic-speaking (Turkic and Mongolic) populations clustered with Uralic speakers. Southern Austronesians first clustered with Tai-Kadai speakers and then clustered with Hmong-Mien and Austroasiatic speakers. Tibetans first clustered with each other, especially for high-altitude Ü-Tsang Tibetans, and then clustered with the lowland East Asians. The observed geographical affinity showed that the genetic differentiation between modern highland Tibetans and lowland East Asians could be identified although they both derived majority of their ancestry from Neolithic Yellow River farmers. We further analyzed the population splits and gene flow events between modern Tibetans and 26 ancients from eastern Eurasia (except for Anatolia_N from Near East) with three predefined admixture events. Modern Tibetans (except for Gannan and Xinlong Tibetans) first clustered with the highland Nepal ancients and then clustered with the lowland Neolithic-NEAs and Neolithic to Bronze Age southern Siberians. The cluster patterns also showed a distant relationship between northern and southern East Asians, as well as the genetic distinction between the highland ancient/modern Tibetans and the lowland SEAs, which further provided evidence for some special connections or close genetic relationships between Tibetans and NEAs.

Genetic affinity was further evaluated via the outgroup-f_3-statistics in the form f_3(*modern Tibetans, ancient/modern Eurasians, Mbuti*). We found a close genetic affinity within Tibetan populations and identified the genetic connection between Tibetan and Han Chinese. Among 184 modern populations (**Figure 3** and **Supplementary Table 3**), the top allele sharing population for each Tibet Tibetan was another geographically close Tibetan group. Shannan Tibetan shared the most alleles with Lhasa/Shigatse/Nagqu Tibetans, and similar patterns of population affinity were identified in southern Shigatse Tibetan and central Lhasa Tibetan. However, Nagqu Tibetan shared the most alleles with the northeastern Chamdo Kham Tibetan (followed by Tibetan-Burman-speaking Qiang from Sichuan province and other Tibetans or Sherpa), and these patterns of genetic affinity were consistent with that of Chamdo Tibetan and others. Following the genomic affinity within Tibetans, we also found that these five Tibet Tibetans shared the strongest genetic affinity with the lowland Han Chinese, which was consistent with the common origin of Sino-Tibetan speakers from the Upper and Middle Yellow River Basin (YRB). For Sichuan/Yunnan lowland Kham Tibetans, Xinlong Tibetan shared the most genetic drift with Han Chinese and other lowland Tibeto-Burman-speaking Qiang and Tujia. Being different from Xinlong Tibetan, geographically close Yajiang and Yunnan Tibetans shared the most genetic drifts with Qiang and geographically close Tibetans (Chamdo and Xinlong), followed by Han Chinese and other Tibetans. These lowland Han/SEA affinities of Kham Tibetans suggested that lowland Tibetans from southwestern China harbored ancestry that derived from SEAs via the massive migrations and admixtures in the prehistoric/historic times. Gangcha Ando Tibetan not only showed the genetic affinity with Sinitic and Tibeto-Burman speakers but also showed the signals of genetic affinity with Turkic-speaking populations. Allele sharing results from Gannan and Xunhua Tibetans showed that the Han Chinese groups shared the most ancestry components with them.

Levels of allele sharing between modern Tibetans and 106 Paleolithic to historic Eurasian ancients (including 33 populations from Russia, 41 from China, 29 from Mongolia, and 3 from Nepal) inferred from the outgroup-f_3-statistics showed that modern Tibetans had a clear connection with ancient Neolithic to Iron Age NEAs, which was consistent with the patterns observed in the PCA, F_{ST}, ADMIXTURE and modern population-based affinity estimations (**Supplementary Table 3**). Middle-altitude Chamdo Tibetan shared the most genetic drift with Neolithic Wuzhuangguoliang_LN (low coverage samples), upper Yellow River Late Neolithic farmers (Jinchankou and Lajia, which are the represented typical source populations for Qijia culture), followed by Iron Age Dacaozi people, Shimao people from Shaanxi, Middle-Neolithic Banlashan associated with Hongshan culture in northern China and other NEAs from lower and middle YRB (**Supplementary Figure 7**). Neolithic people from Russia and Mongolia and Bronze to historic Nepal ancients showed a relatively distant genetic relationship with modern Chamdo Tibetan (**Supplementary Table 3**). Different from the pattern of Chamdo Tibetan, southern and central Ü-Tsang Tibetans showed increased ancestry associated with Nepal

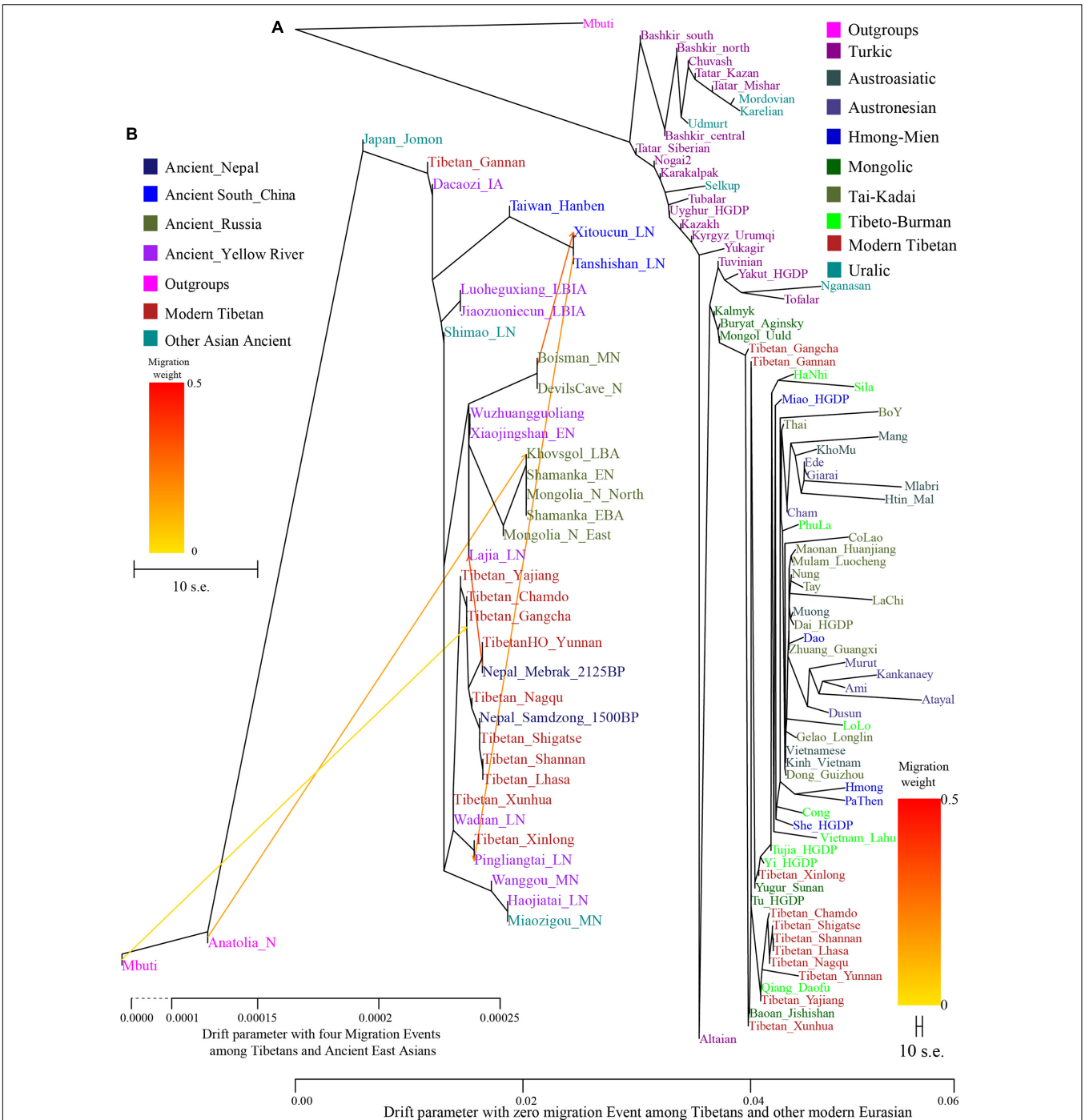

FIGURE 2 | Maximum likelihood phylogeny reconstruction based on the genetic variation from both modern Tibetan and Eurasian modern reference populations. **(A)**, modern Tibetan and Neolithic-to-historic East Asian **(B)**. Mbuti was used as the root. Focused on the phylogenetic relationship among all modern populations, we used the patterns of genetic relationship with zero migration events. And evaluating the evolutionary history among modern Tibetan and ancient Chinese, we included three migration events. To better present our result, the drift branch length of Mlabri was shortened as the third of the truth drift branch length due to the strong genetic drift that occurred in Mlabri.

ancient people, and northern Nagqu Tibetan showed the intermediate trend of population affinity with 2700-year-old Chokhopani. As showed in **Supplementary Figures 8, 9**, lowland Tibetans from southwestern China and northeastern China showed a similar population affinity with NEA ancients. The genomic affinity between modern Tibetans and some southern East Asians (such as Oakaie_LNBA) could be also identified in **Figure 3**.

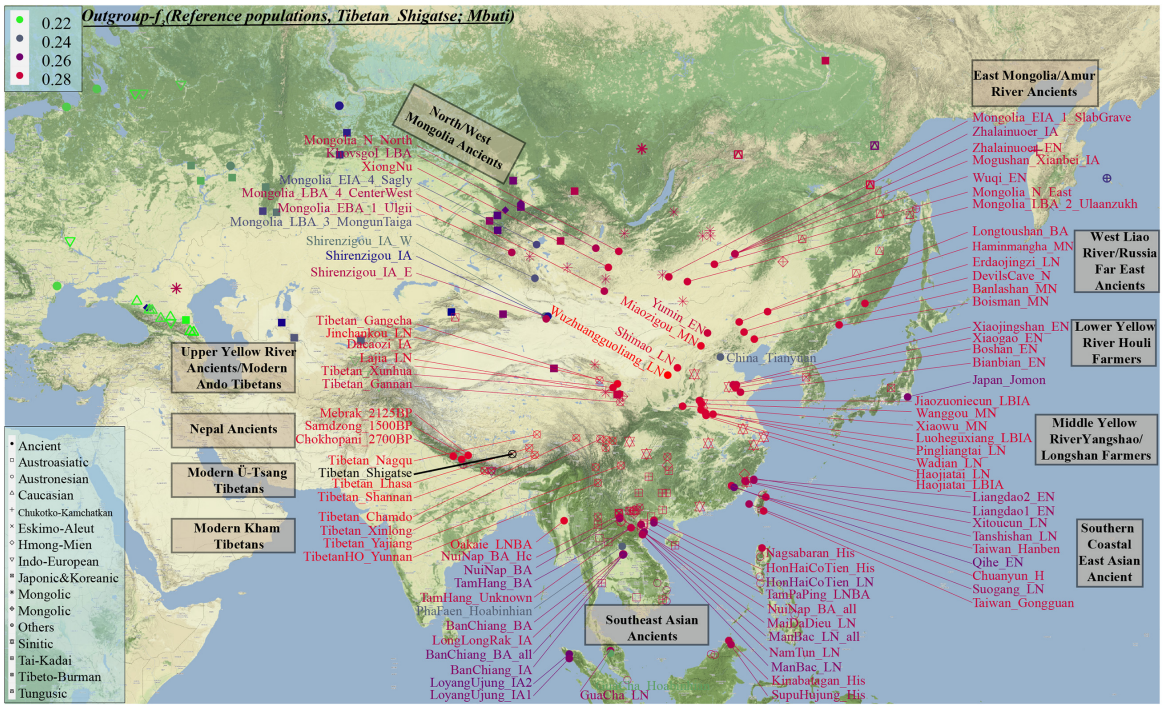

FIGURE 3 | The genomic affinity between our Shigatse Tibetan populations and other modern and ancient spatial-temporally different eastern Eurasian populations. The red color denoted a stronger genetic affinity with Shigatse Tibetans, and the blue color showed a lower genetic affinity.

Admixture Signatures of Modern Tibetans and Ancient Populations From Tibetan Plateau

We carried out admixture-f_3-statistics in the form f_3*(source population1, source population2; Targeted Tibetan)* to detect the signals of recent genetic admixture in Tibetans. We also re-evaluated the admixture signatures in the eight ancient individuals from Nepal and eleven ancient individuals from Qinghai province using this three-population comparison testing and our comprehensive ancient/modern reference dataset. We found different patterns of admixture signals and source populations in the highland/lowland ancient/modern Tibetans (**Supplementary Tables 4–18**). Besides, we also identified small but significant differences within geographically/culturally different Tibetans. By setting the statistically significant threshold at Z-score < -3, no admixture signals were observed in southern Tibetans (Shannan and Shigatse) over forty thousand tested pairs, and only four pairs in central Lhasa Tibetan with one source from 1500-year-old Samdzong and other from Kham Tibetan/Qiang, or the combination of southern Tibet Tibetan with Neolithic-NEAs or Baikal ancients (**Supplementary Tables 4–6**). It was interesting to find that 188 tested population pairs showed statistically significant f_3-statistic values with one source from Tibeto-Burman speakers and the other from Western Eurasian Steppe pastoralists (Alan, Andronovo, Sintashta, Poltavka, and Yamnaya) in f_3*(Source1, Source2; Nagqu Tibetan)*. Tibetans from southern and central Tibet combined with the lowland modern East Asians, but not with ancient East Asians, could also produce significant admixture signals for Nagqu Tibetan (**Supplementary Table 7**). Chamdo Tibetan at the junction regions between Ü-Tsang Tibetan and Kham Tibetan had the potential possibility of cultural contact and population admixture, but only one pair of source populations could give a significant admixture signal in Chamdo Tibetans: f_3*(Lhasa Tibetan, Yajiang Tibetan; Chamdo Tibetan)* $= -3.49^*SE$ (**Supplementary Table 8**). Three Tibetans from the Gansu-Qinghai region possessed admixture signatures from over several thousand population pairs with one from modern or ancient East Asians and the other from Western Eurasians (**Supplementary Tables 9–11**). Results from f_3*(Yumin_EN, Austronesian/Tai-Kadai; Gansu-Qinghai Tibetan)* showed that the combination of inland Neolithic NEA of Yumin_EN as northern ancestral source with Austronesian/Tai-Kadai speakers as the southern ancestral source could produce significant negative f_3-values, and these admixture signals could also be identified in f_3*(Neolithic NEAs, Neolithic-Russian/modern Turkic/Mongolic/Indo-European speakers; Gansu-Qinghai Tibetan)*. Tibetans from Sichuan province only showed significant signals as an admixture between northern and southern East Asians or the highland Tibeto-Burman speakers and lowland East Asians, i.e., f_3*(highland Tibeto-Burman speakers, lowland Tibeto-Burman speakers; Sichuan Tibetan)* $< -3^*SE$ (**Supplementary Tables 12, 13**). Similar to the southern Tibet Tibetans, no obvious admixture signals were observed in Yunnan Tibetans, which may be caused by the genetic isolation or obvious genetic drift that occurred recently (**Supplementary Table 14**). The statistics focused on the ancient populations from the TP showed

seven pairs can give admixture signals for modeling Qinghai Iron Age Dacaozi samples (**Supplementary Tables 15–18**), which are the pairs of ancient NEAs and modern SEAs, or Chamdo Tibetan-related source and Taiwan Iron Age Hanben-like populations.

Intra Population Differentiation Amongst High-Altitude and Low-Altitude Residing Tibetans Inferred From f_4-Statistics

To gain insights into the population substructures among modern Tibetans, we first conducted symmetry-f_4-statistics in the form f_4*(modern Tibetan1, modern Tibetan2; modern Tibetan3, Mbuti)*, in which we expected to observe the non-significant f_4-values if no significant differences existed between different Tibetan groups. As shown in **Supplementary Table 19** and **Supplementary Figure 10**, we observed that Chamdo Tibetan formed a clade with Nagqu/Yunnan Tibetans compared with others in f_4*(Tibetan1, Chamdo Tibetan; Tibetan2, Mbuti)* and all included Tibetans shared more alleles with Chamdo Tibetan compared with Ando Tibetans. Compared to the low-altitude Sichuan Tibetans, Chamdo Tibetan had more high-altitude Tibetan-related ancestry, while Gannan Tibetan shared more alleles with Xinlong Tibetan compared with Chamdo Tibetan. Compared with high-altitude Tibetans, Chamdo Tibetan shared more alleles with other low-altitude Tibetans. Results from the symmetry-f_4*(Shigatse/Shannan/Lhasa Tibetans, Shigatse/Shannan/Lhasa Tibetans; Tibetan2, Mbuti)* with non-significant Z-scores showed clear genetic homogeneity among Tibet central/southern-Ü-Tsang Tibetans (**Supplementary Figures 11, 12**). Negative-f_4-values in f_4*(Gansu-Qinghai Ando Tibetans, Shigatse/Shannan/Lhasa Tibetan; Tibetans, Mbuti)* showed that all included Tibetans shared more alleles with southern Tibet Tibetans relative to Gansu-Qinghai Ando Tibetans. However, northern Tibet Tibetans formed a clade with Chamdo and Yunnan Tibetans and received more high-altitude Tibetan-related derived alleles compared with Gansu-Qinghai and Sichuan Tibetans. For lowland Tibetans, northwestern Tibetans in Gangcha and Xunhua formed one clade, i.e., all absolute Z-scores of f_4*(Gangcha, Xunhua Tibetan; Tibetan2, Mbuti)* were less than three (**Supplementary Figure 13**). Compared with Gannan Tibetans, Qinghai Tibetans had more ancestry sharing with Tibet Tibetans. We did not find Tibetan populations shared more alleles with Gannan Tibetans relative to other Tibetans, as all values in f_4*(Tibetan1, Gannan Tibetan; Tibetan2, Mbuti)* were larger than zero. Southwestern Yunnan Tibetan formed one clade with Chamdo/Xinlong/Yajiang Tibetans, all of them belonged to Kham Tibetans (**Supplementary Figures 14, 15**). Lowland Sichuan/Yunnan Tibetans harbored increased Tibetan-related derived alleles compared with Gansu-Qinghai Tibetans and more ancestry related to highland Tibetans compared with other highland Tibetans.

We additionally explored genetic affinity and population substructure among highland and lowland Tibetans using ancient Eurasian populations via f_4*(Modern Tibetan1, Modern Tibetan2; Ancient Eurasians, Mbuti)*. The non-significant Z-scores in f_4*(Ü-Tsang Tibetans1, Ü-Tsang Tibetans2; Ancient Eurasians, Mbuti)* confirmed the genomic homogeneity within the four high-altitude Ü-Tsang Tibetans. We could also identify the more allele sharing between the Nepal ancients and Ü-Tsang Tibetans compared to Ando and Kham Tibetans (**Supplementary Figures 16–19**). Compared with Shannan Tibetan, Nagqu Tibetan harbored increased ancestry associated with the lowland ancient populations. Compared to Qinghai Ando Tibetans, Nagqu Tibetan possessed both increased Nepal ancients-related ancestry and increased Late Neolithic Lajia-related ancestry relative to Xunhua Tibetan. Nagqu Tibetan also harbored additionally increased ancestry related to the coastal Late Neolithic SEAs, middle Yellow River Middle-Neolithic to Iron Age ancient populations, Upper Xiajiadian culture-related Bronze Age populations, inland Neolithic NEAs and other upper Yellow River Late Neolithic and Iron Age populations. Significant negative-f_4-values were observed in Ando Tibetans via f_4*(modern Tibetan1, Gansu-Qinghai Ando Tibetans; Bronze Age stepped pastoralists, Mbuti)*, which suggested that Ando Tibetans harbored increased ancestry related to steppe pastoralists, such as Sintashta, Yamnaya, Afanasievo, Srubnaya, Andronovo and Xinjiang Iron Age Shirenzigou populations. Although strong genetic affinity within Ando Tibetans was confirmed with the similar patterns of f_4-based sharing alleles and non-significant statistical results in symmetry-f_4 statistics. Statistically significant negative f_4-values in f_4*(Gangcha Tibetan, Gannan Tibetan; Ami/Atayal/Hanben/Gongguan/Tanshishan_LN/Qihe_EN, Mbuti)* showed that Gannan Tibetan harbored increased SEA ancestry related to modern Austronesian or Proto-Austronesian-related Neolithic to present-day southeastern coastal/island populations (**Supplementary Figures 20–22**). A similar SEA affinity of Gannan Tibetan was also identified compared with Tibet Ü-Tsang Tibetans. Results of the four-population comparison analysis focused on Kham Tibetans are presented in **Supplementary Figures 23–25**, which suggested that Kham Tibetans had increased both northern and SEA ancestry.

Spatiotemporal Comparison Analysis Among Modern Tibetans and All Paleolithic-to-Historic East Asians Showed the Genetic Admixture and Continuity of Modern Tibetans

We nest used f_4-statistics to elucidate the patterns of genomic structure and population dynamic of East Asians and provide new insights into the origin of culturally/geographically diverse Tibetans. Focused on four early coastal Neolithic NEAs from Shandong province, f_4*(coastal Neolithic NEA1, coastal Neolithic NEA2; Modern Tibetans/Ancient East Asians, Mbuti)* revealed the similar genetic relationship between modern Tibetans and these different Neolithic NEAs (**Supplementary Figure 26**). Results from f_4*(Bronze/Iron Age Henan populations, Neolithic-to-Iron-Age Henan populations; Eastern Modern Tibetan/Ancient East Asians, Mbuti)* only revealed Luoheguxiang people had increased ancestry associated with modern Austronesian-speaking Ami (**Supplementary Figures 27–29**) relative to Wanggou_MN. The

Late Neolithic Haojiatai population had more SEA-like ancestry related to Xitoucun_LN and Iron Age Hanben people compared with Wanggou_MN (**Supplementary Figure 30**). The genetic affinity with southern coastal populations (Ami/Atayal/Hanben-related) was also observed in Pingliangtai_LN, but not in Wadian_LN and Middle Neolithic Wanggou_MN and Xiaowu_EN (**Supplementary Figures 31–34**). Focused on ancients from Shaanxi and Inner Mongolia, we found that modern Tibetans and northern and southern EAs from the Yellow River and south China shared more alleles with Late Neolithic Shimao populations (**Supplementary Figure 35**). Temporal analysis among upper Yellow River ancients showed all modern Tibetans showed a similar relationship with them, although Iron Age Dacaozi people harbored more SEA ancestry. These results suggested that population movements from southern China have a significant influence on the gene pool formation of northeastern populations on the TP at least from the Iron Age (**Supplementary Figure 36**). Symmetrical relationships among East Asians with temporally different Nepal ancient populations were shown in **Supplementary Figure 37**.

Next, we also explored the similarities and differences of the shared genetic profiles related to northern Neolithic East Asians via the spatial comparison analysis with modern Tibetans and all available ancient East Asians as reference. We conducted a series of symmetry f_4-statistics to compare all eleven modern Tibetan populations and other ancient East Asians against the geographically different ancient NEAs and ancient Tibetans. **Figure 4** and **Supplementary Figures 38–41** showed the shared alleles between the targeted populations and the lowland early Neolithic NEAs and others. The f_4*(NEAs, Chokhopani; Modern Tibetan/Neolithic-to-historic East Asians, Mbuti)* was used to determine the lowland and highland East Asian affinity. Compared with four coastal Neolithic Shandong populations, we found that Ü-Tsang Tibetans had a strong highland East Asian affinity. Besides, comparison against the coastal and inland ancients revealed that modern Tibetans had a strong inland-NEA-affinity, especially with Late Neolithic Lajia people from the upper Yellow River. This Lajia-affinity or inland-NEA-affinity persisted when we substituted inland Yumin_MN with the coastal Neolithic NEAs (**Supplementary Figure 42**), but disappeared when we substituted the latter Neolithic groups with the early Neolithic NEAs (**Supplementary Figures 43–48**). We summarized the overall highland/lowland East Asian affinities of Tibetans in **Supplementary Figure 49**, which showed the Ando and Kham Tibetans had lowland NEA affinity, and Ü-Tsang Tibetans possessed additional Nepal ancient affinity.

Our genomic studies have identified population substructures within modern Tibetans. Modern Tibetans can be classified into three subgroups by their different affinities with NEAs, SEAs and Siberians, which were confirmed by the negative values in f_4*(Reference populations, modern Tibetans; northern/southern EAs and Siberians, Mbuti)*. We further tested if one single source could explain the observed genetic variations in Tibetans. We first assumed that modern Tibetans were the direct descendants of SEAs which is associated with the Yangtze Rice farmers. As shown in **Supplementary Figures 50–58**, we observed significant negative f_4-values in f_4*(SEAs, modern Tibetans; Reference populations, Mbuti)* when we used NEAs/Siberians as the reference populations, which indicated obvious gene flow events from these reference populations into modern Tibetans. We then assumed that Tibetans' direct ancestor was coastal Neolithic-NEAs, we conducted f_4*(Shandong ancients, modern Tibetans; Neolithic-to-historic East Asians, Mbuti)* and found only Nepal ancients showed the negative-f_4-values, which was consistent with the common origin of the Sino-Tibetan speakers from YRB (**Supplementary Figures 59–62**). The patterns were confirmed when we assumed Yangshao and Longshan farmers or their related populations (**Supplementary Figures 63–71**), Shaanxi ancients (**Supplementary Figures 72–74**), and other ancient NEAs and southern Siberians (**Supplementary Figures 75–88**) as the direct ancestor of modern Tibetans. As shown in **Supplementary Figures 75–88**, when assuming Yumin or Ulchi as the direct ancestor of Tibetans, we identified additional gene flows from the SEAs (Hanben and Tanshishan et al.) and Yellow River farmers into Tibetans. Assuming the Nepal ancients as direct ancestors, we detected obvious additional gene flow from the lowland ancient East Asians to Kham Tibetans (**Supplementary Figures 89–91**). Additional predefined ancestral populations from Russia and Chinese Xinjiang further confirmed the strong northern East Asian affinity (**Supplementary Figures 92–104**). Thus, f_4-statistics showed that the formation of modern Tibetans had involved multiple admixture events.

Ancestry Compositions of Ancient/Modern Tibetans via *qpWave*/*qpAdm* and *qpGraph*

From the autosomal perspective, we found the close connections of modern Tibetans and Neolithic NEAs. From a paternal Y chromosomal perspective, Tibetan shared a genetic affinity with Andamanese Onge and Jomon hunter-gatherers from the Japanese archipelago (Shi et al., 2008). Onge and Jomon were suggested to be an early Asian lineage with a close relationship with 7700-year-old Hoabinhian from southeast Asia (McColl et al., 2018). We further explored the number of ancestral populations of modern Tibetans, Nepal ancients and Jomon using the *qpWave* and estimated their corresponding ancestry proportions under one-way, two-day and three-way admixture models. The *qpWave* results (p_rank < 0.05) showed that at least two ancestral populations were needed to explain the observed genetic variations in targeted populations. We first employed the two-way model of Onge and six inland/coastal early Neolithic-NEAs and found inland Yumin failed to fit our targeted populations' genetic variations (all p_values < 0.05). The two-way model "Xiaogao_EN-Onge" could fitted all modern Tibetans well except for Gannan Tibetan with the Xiaogao-related ancestry proportion ranging from 0.846 in Shannan Tibetan to 0.906 in Xinlong Tibetan. The 2700-year-old Chokhopani, like geographically close Shigatse Ü-Tsang Tibetans, could be fitted as an admixture of 0.861 NEA Xiaogao-related ancestry and 0.139 Onge-related ancestry (**Supplementary Table 20** and **Figure 5**). Younger Nepal ancient could be modeled as major ancestry from Onge-related ancestry and minor ancestry

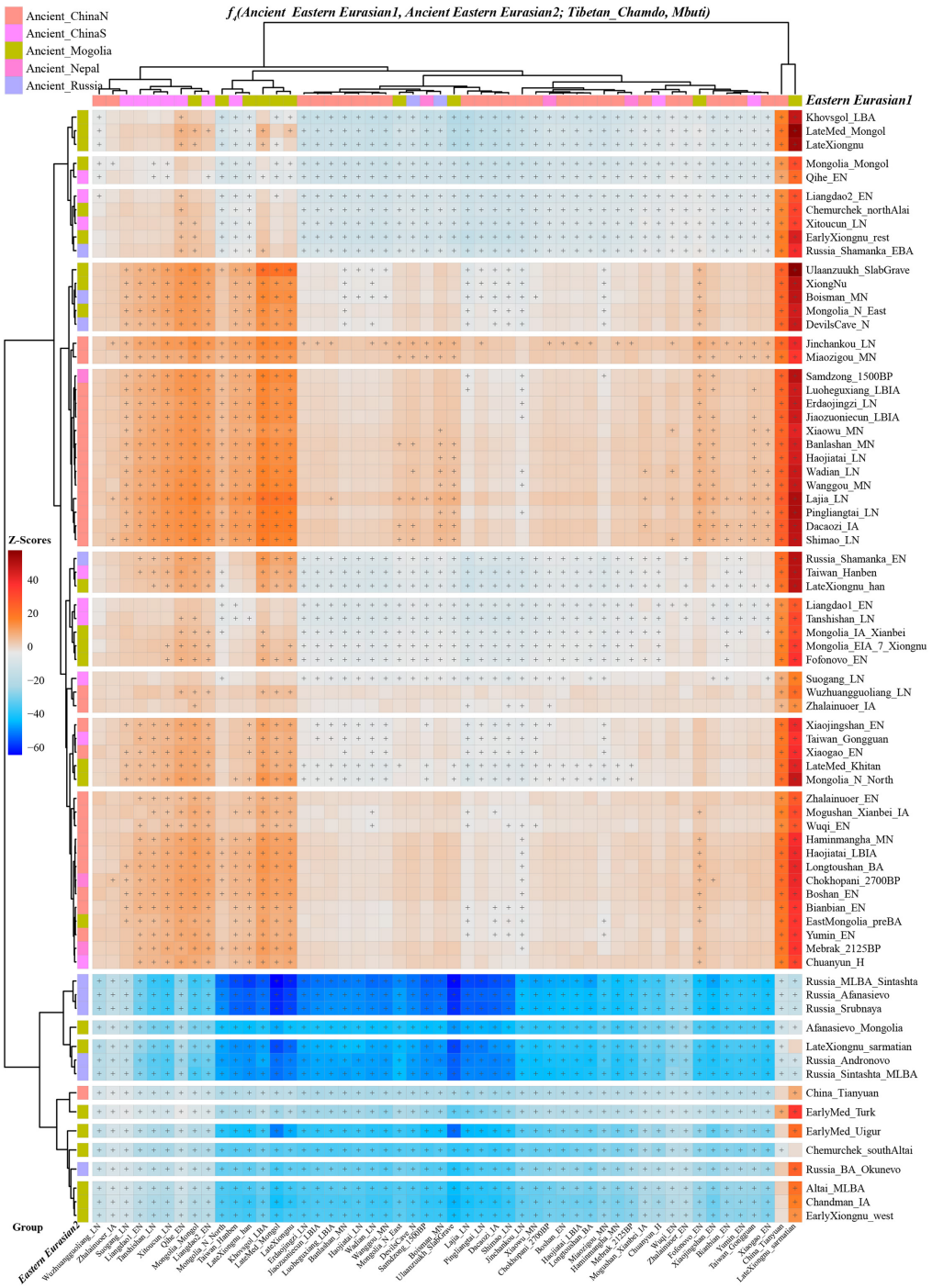

FIGURE 4 | The genomic affinity between Chamdo Tibetans and other eastern Eurasian ancient populations inferred from four population affinity-f_4 statistics of the form f_4(Ancient Eastern Eurasian1, Ancient Eastern Eurasian; Tibetan_Chamdo, Mbuti). Red color with statistically significant f_4-values (marked with "+") demoted Chamdo Tibetans shared more derived alleles with Ancient Eastern Eurasian1 (right population lists) compared with Ancient Eastern Eurasian2 (bottom population lists). Blue color with significant f_4-values denoted Chamdo Tibetans shared more Ancient Eastern Eurasian1-related derived alleles relative to their counterpart.

associated with NEA lineage. Jomon could be modeled as deriving 0.484 of its ancestry from populations related to Xiaogao_EN and 0.516 from groups related to Onge with marginal statistical significance. We substituted Boshan_EN and Bianbian_EN with Xiaogao_EN, we could obtain similar results, however, when we substituted Xiaojingshan_EN with Boshan_EN, 1500-year-old Samdzong failed to fit our two-way model (p_rank1 = 0.00007). The "Zhalainuoer_EN-Onge"

model could be successfully fitted highland Tibet Tibetans and Yunnan Tibetan with high Onge-related ancestry but failed to fit other Ando and Kham Tibetans. Using Middle-Neolithic East Asian as the source, the "Xiaowu_MN-Onge" model failed to all targets, and the "DevilsCave_N-Onge" model could only fit the Sichuan Tibetans, Jomon, and Chokhopani with a high proportion of Onge-related ancestry. Except for populations with a western Eurasian affinity (Ando Tibetans and Samdzong), all remaining ancient/modern populations could be fitted as the admixture between Onge and Middle Neolithic Wanggou_MN, Banlashan_MN, or Miaozigou_MN. We additionally substituted Onge with Hoabinhian as the southern source representative for deep lineage and used early Neolithic to Late-Neolithic NEAs as the other source to perform the two-way admixture model for estimating the ancestry proportion of modern Tibetan without Gangcha and Gannan Tibetans and Nepal ancients except for ancient Samdzong and Jomon. As shown in **Figure 5**, a good fit could be acquired with slightly variable ancestry composition compared with Onge-based two-way models. We finally employed the Afanasievo (significant negative-f_3 value in admixture-f_3-statistics) as the western Eurasian source in a three-way admixture model to fit the genetic variations in Ando Gangcha and Gannan Tibetans and Samdzong. All three populations could be successfully fitted when we introduced the Bronze Age steppe pastoralists' related ancestry.

Finally, to comprehensively summarize the phylogenetic relationships and reconstruct the population history between Neolithic East Asians and modern Tibetans in one phylogenetic framework, we built a series of admixture graph models via *qpGraph*. The core model of our admixture graph included archaic Denisovan and central African Mbuti as the roots, Loschbour as the representative of western Eurasian, modern Onge hunter-gatherer from Andaman island and 40,000-year-old Tianyuan (3% ancestry from Denisovan) as representatives of deep lineages of southern East Eurasian and northern East Eurasian. As shown in **Figure 6A**, East Asians diverged into northern lineage (represented by East Mongolia Neolithic population with 1% gene flow from western Eurasian) and southern lineage (represented by Liangdao2_EN with 35% ancestry deriving from lineages close to Onge). Here, Late Neolithic Qijia-related Lajia people could be fitted as an admixture of 84% from a lineage related to NEAs and 16% from a lineage associated with Andamanese Onge. Ancient Chokhopani in Nepal could be modeled as driving 86% of the ancestry from Lajia_LN and 14% from the Onge side. Our model provided ancient genomic evidence of the co-existence of both Paleolithic hunter-gatherer ancestry associated with the indigenous TP people and Neolithic NEA ancestry in Chokhopani culture-related ancient Tibetans and Late Neolithic Lajia people. We subsequently added all eleven modern Tibetan populations to this scaffold model and found all Ü-Tsang and Kham Tibetans except for Xinlong Tibetan could be fitted as direct descendants from Chokhopani with additional gene flow from one NEA related population, which also contributed additional 33% ancestry to Iron Age Hanben people. This gene flow could be regarded as the epitome of the second wave of Neolithic expansion into TP. Thus, results from **Figure 6** suggested that seven Tibetans could be well fitted with three sources of ancestry: Onge-related, Lajia_LN-related and second wave of NEA lineage-related, in respective proportion of 0.1235, 0.8265, and 0.0500 (Shannan); 0.1440, 0.8160, and 0.0400 (Shigatse); 0.1344, 0.8256, and 0.0400 (Lhasa), 0.1176, 0.7224, and 0.1600 (Nagqu); 0.1001, 0.6699, and 0.2300 (Chamdo); 0.1106, 0.6794, and 0.2100 (Yunnan); 0.1232, 0.7568, and 0.1200 (Yajiang). We could obtain a good fit when considering one gene flow event for Gansu-Qinghai Ando Tibetans with the Loschbour-related ancestry proportion varying from 2 to 3% (**Figure 7**). To further explore the best ancestral source proximity of the second migration wave, extended admixture graphs introducing inland/coastal northern and SEA Neolithic populations were reconstructed. As shown in **Figure 8**, the second wave into lowland Kham Tibetans with Neolithic SEA affinity could be well fitted as directly deriving from Hanben-related ancestral population with the proportion ranging from 5 to 11%. We then added northern coastal early Neolithic Houli Boshan people, Middle Neolithic Xiaowu Yangshao people, Late Neolithic Wadian people, and Bronze to Iron Age Haojiatai Shangzhou people to our core model in **Figure 6** and then fitted all Tibetans on it. We found that Yunnan Kham Tibetan harbored 33% additional ancestry associated with Longshan people, and Sichuan Yajiang Kham Tibetan with 26% additional Longshan-related ancestry (**Figure 9**). It was interesting to find that the gene pool of the Lhasa Ü-Tsang Tibetan was also influenced by the second population migration associated with the Longshan people. This second gene flow event persisted when we substituted Longshan people with other Neolithic or Bronze to Iron Age populations with acceptable ancestry proportions (**Supplementary Figures 105–107**). These phenomena may be caused by the genetic stability of the main ancestry in the Central Plain (Henan and Shandong provinces).

Recent genetic studies have evidenced that Denisovan-like haplotypes have contributed to the high-altitude adaptation of modern Tibetans (Huerta-Sánchez et al., 2014). Morphologic evidence from Xiahe people and mitochondrial DNA from Baishiya Karst Cave's ancient remains further suggested that archaic people related to Denisovan had arrived at TP during the Pleistocene (Chen F. et al., 2019; Zhang et al., 2020). However, our best-fitted *qpGraph*-based phylogenetic frameworks (**Figures 6–9**) did not show the expected genetic contribution from archaic hominid into modern Tibetans, probably due to the limited number of the available SNPs and the relatively low proportion of Denisovan ancestry in East Asians. Ancient genomes from ∼34,000-year-old Mongolia Salkhit and 40,000-year-old Tianyuan people have been evidenced for carrying genomic segments of Denisovan ancestry (Massilani et al., 2020). We also identified the archaic admixture into Tianyuan-related people in our reconstructed models. Next, we conducted the f_4-statistics in the form of f_4(modern and ancient East Asians, Tibetans; Denisovan, Chimpanzee) and did not identify significant f_4-values, which suggested that highland and lowland East Asians formed a clade relative to the Denisovan and both harbored equal levels of Denisovan related ancestry. Similar patterns of archaic admixture between Neolithic East Asians and modern East Asians were also evidenced in a recent

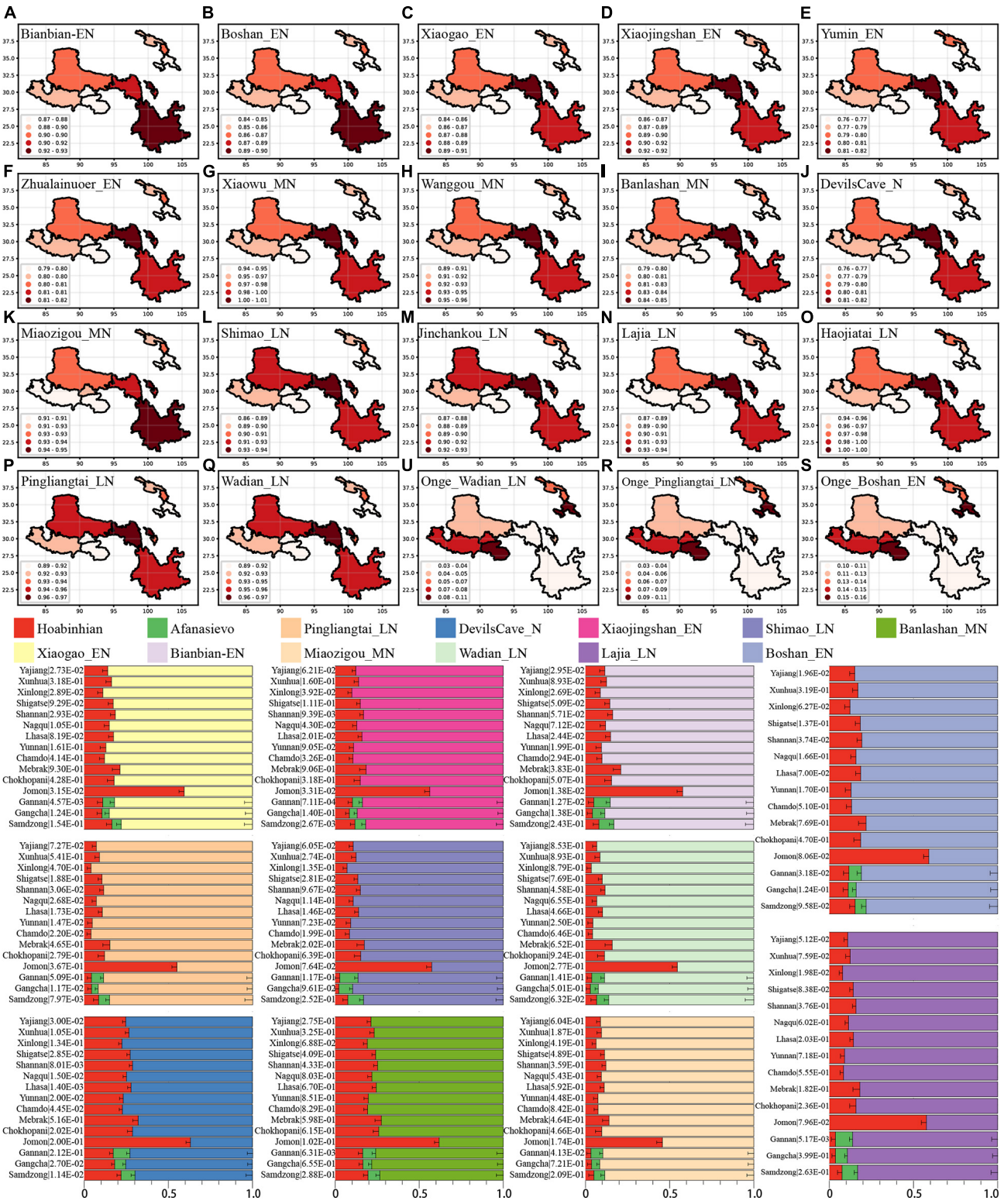

FIGURE 5 | Results of qpAdm showed the main ancestry composition of ancient/modern Tibetans and Jomon Hunter-Gatherer were the results of the mixing of ancient NEA and one deep lineage associated with South Asian Hunter-Gatherer Onge or Southeast Hunter-Gatherer Hoabinhian (the early Asian). Heatmap showed the NEA-related ancestry in the two-way admixture model of Onge and the early Neolithic East Asian **(A–F)**, Middle-Neolithic NEA **(G–K)**, and Late-Neolithic NEA **(L–Q)**. Onge-related ancestry was presented with three cases **(R,S,U)**. Bar plots showed the ancestry composition of the two-way model of Hoabinhian and East Asian for modern Tibetan, Jomon and Ancient Nepal Mebrak and Samdzong people, and three-way model for Qinghai and Gansu Tibetans.

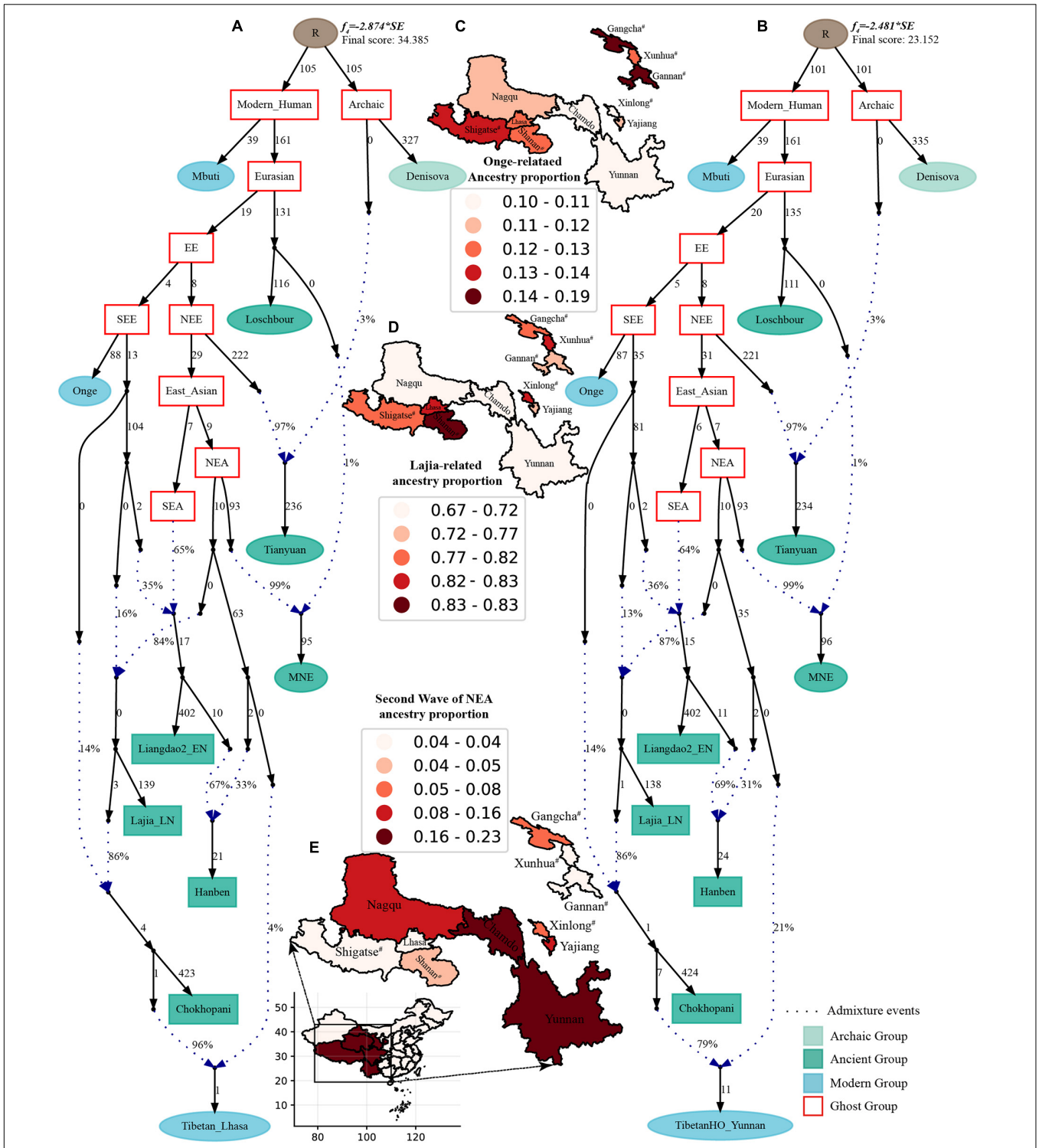

FIGURE 6 | Admixture graph model of East Asians and modern Tibetans based on the Human Origin dataset. Admixture history of highland Tibetan from Lhasa **(A)** and lowland Tibetan from Yunnan **(B)**. Heatmap showed the ancestry composition of modern Tibetans from three source populations: deep hunter-gatherer One-related ancestry **(C)**, the first batch of Neolithic farmer-associated ancestry **(D)** and the second batch of Neolithic farmer related ancestry **(E)**. Denisovan and Central African of Mbuti were used as the Archaic and modern roots respectively. Western Eurasian was represented by Loschbour. Deep southern Eurasian (SEE) and northern Eurasian (NEE) were represented by South Asian Hunter-Gatherers of Onge and 40,000-year-old Tianyuan people. East Asian subsequently diverged as NEA (NEA) and SEA (SEA). All f_4-statistics of included populations are predicted to within 3 standard errors of their observed values. Branch lengths are given in units of 1000 times the f_2 drift distance (rounded to the nearest integer). Pound signs denoted the modern populations added to the basic model of A,B with larger Z-scores or Zero internal branch length. Blue dotted lines denoted admixture events with admixture proportions as shown.

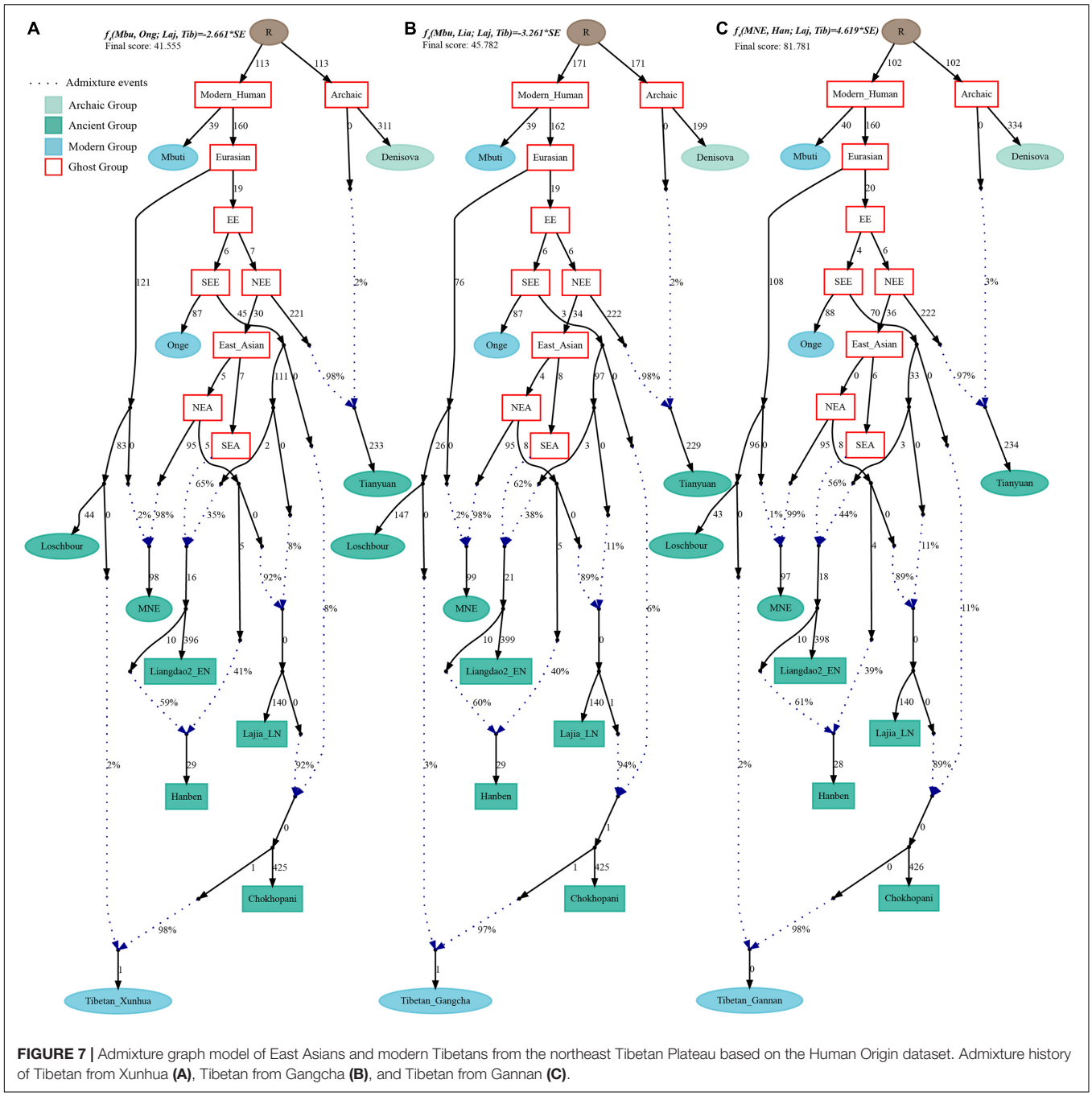

FIGURE 7 | Admixture graph model of East Asians and modern Tibetans from the northeast Tibetan Plateau based on the Human Origin dataset. Admixture history of Tibetan from Xunhua **(A)**, Tibetan from Gangcha **(B)**, and Tibetan from Gannan **(C)**.

study on ancient DNA (Yang et al., 2020). We further tried to explore the highly differentiated 5-SNP EPAS1 haplotype motif (AGGAA) in our studied Tibetan populations, and we found that these five SNP loci were not included in the array we used. To further explore the possible bias caused by the lower SNP density in the HO dataset, we reconstructed a new *qpGraph* model based on the 1240K dataset focusing on both East Asians and Oceanians since Australians and Papuans have been suggested to possess a higher proportion of Denisovan related ancestry (Browning et al., 2018). We successfully identified an additional Denisovan-related gene flow into modern Oceanian populations (4%), but the obtained best-fitted model also did not include the genetic contribution from archaic people into modern Tibetans (**Figure 10**). Our *qpGraph*-based phylogeny showed that Tibetan was modeled as an admixture of 74% ancestry from the upper Yellow River farmers and 26% from Guangxi pre-Neolithic Longlin people. We should note that the obvious archaic gene flow into modern Tibetans had been documented in the proposed two-wave model of 'Admixture of Admixture' based on the phased haplotype via Lu's whole-genome sequenced study (Lu et al., 2016). Besides, Browning et al. (2018) also documented

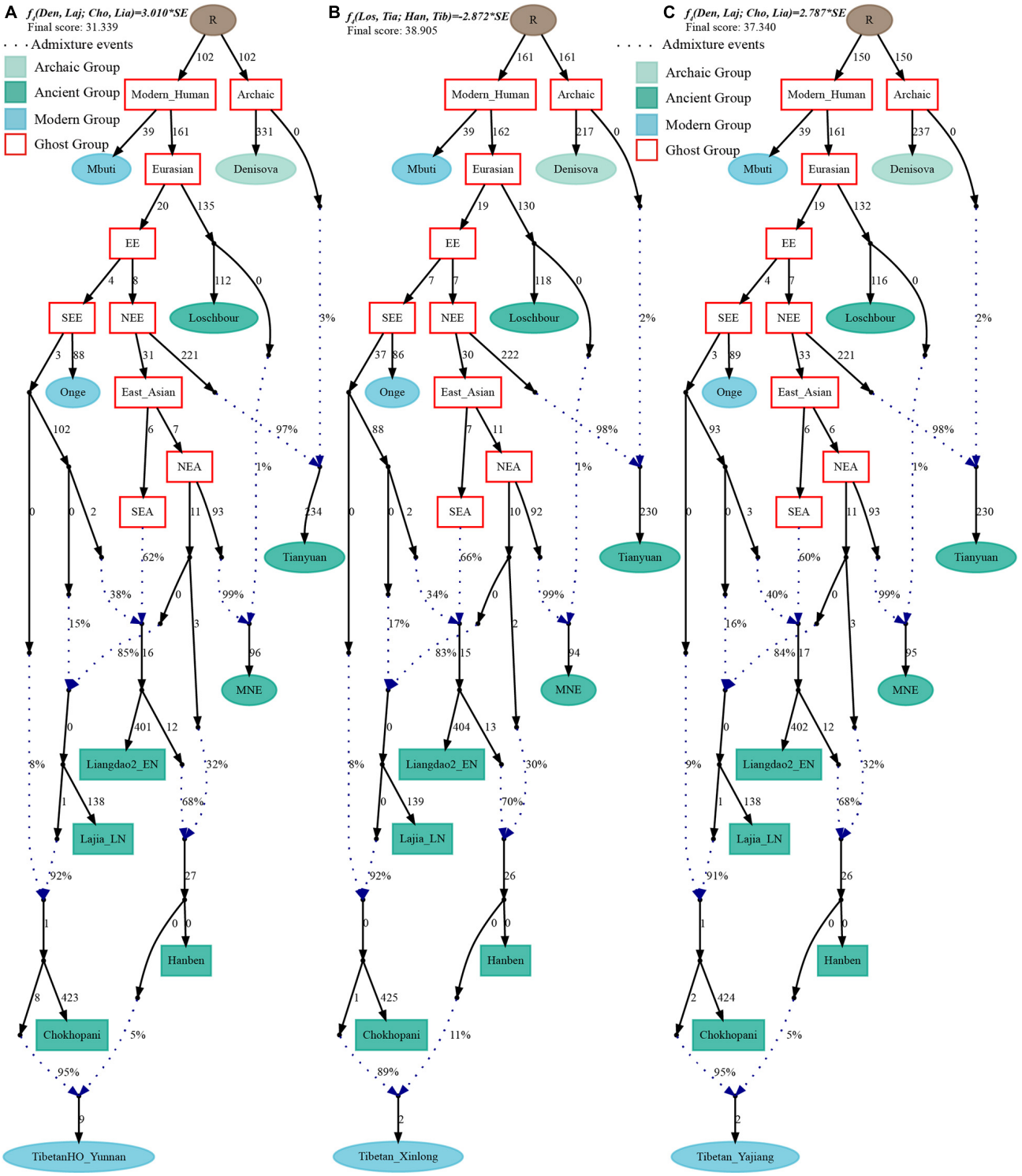

FIGURE 8 | Admixture graph model of East Asians and modern lowland Tibetans based on the Human Origin dataset. Admixture history of lowland Tibetan from Yunnan **(A)**, Tibetan from Xinlong **(B)**, and Tibetan from Yajiang **(C)**.

two different pluses of archaic Denisovan admixture into East Asians. Thus, our reconstructed phylogenetic modeling graph without Denisovan archaic gene flow into modern Tibetans may be caused by the low admixture introgression levels at whole-genome scale, or by the enrichment of archaic genes in just certain specific regions, such as the EPAS1. The actual

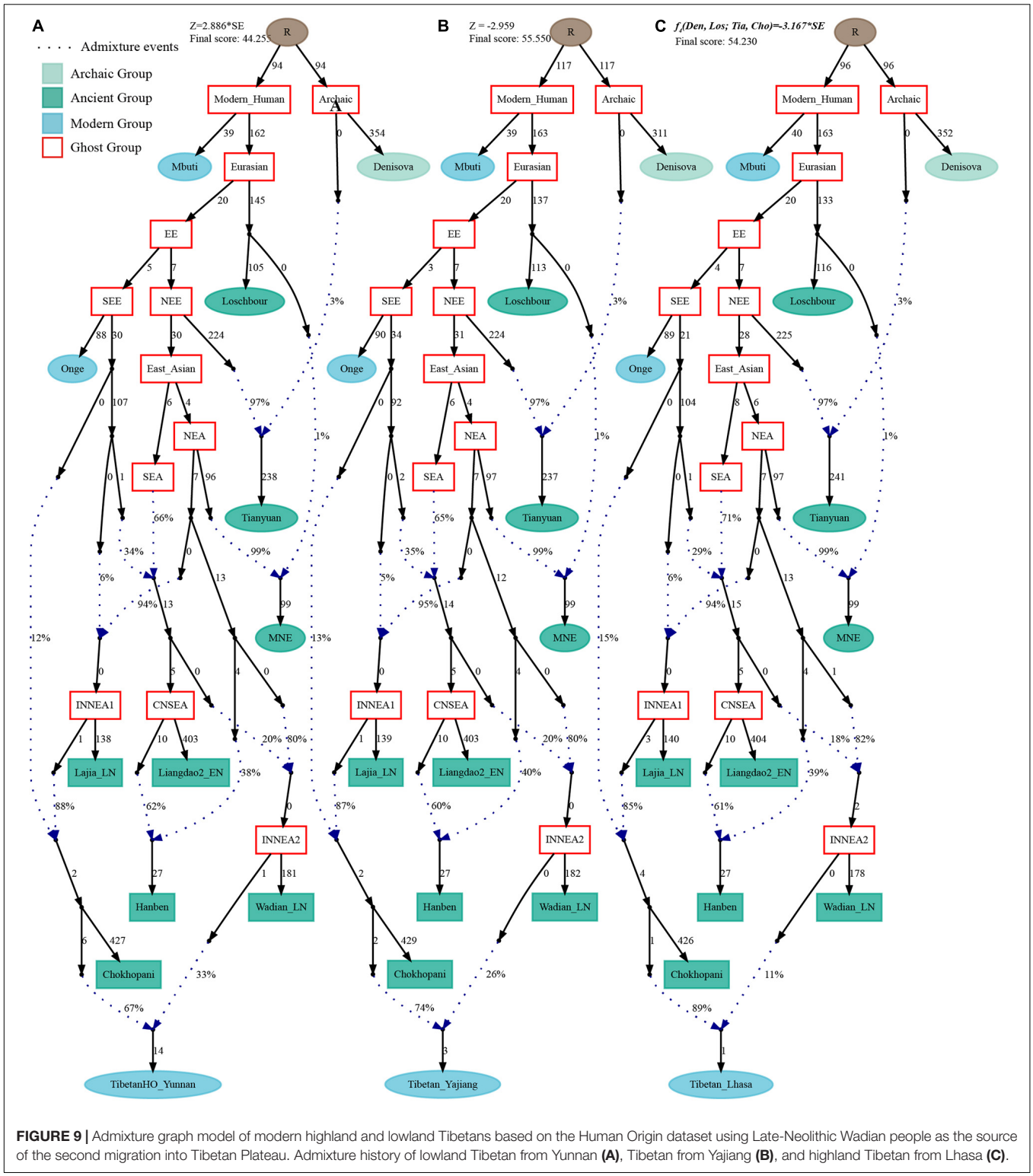

FIGURE 9 | Admixture graph model of modern highland and lowland Tibetans based on the Human Origin dataset using Late-Neolithic Wadian people as the source of the second migration into Tibetan Plateau. Admixture history of lowland Tibetan from Yunnan **(A)**, Tibetan from Yajiang **(B)**, and highland Tibetan from Lhasa **(C)**.

genetic interaction and introgression between lowland/highland anatomically and behaviorally modern humans and archaic people may be more complex. More powerful statistical computational methods and long-read sequencing data may provide new insights into the archaic admixture landscape of ancestral Tibetan populations. Thus, deep-whole-genome sequencing of modern and ancient highland East Asians needs to be conducted to further explore, simulate and validate the complete landscape of Denisovan gene diversity in modern Tibetans.

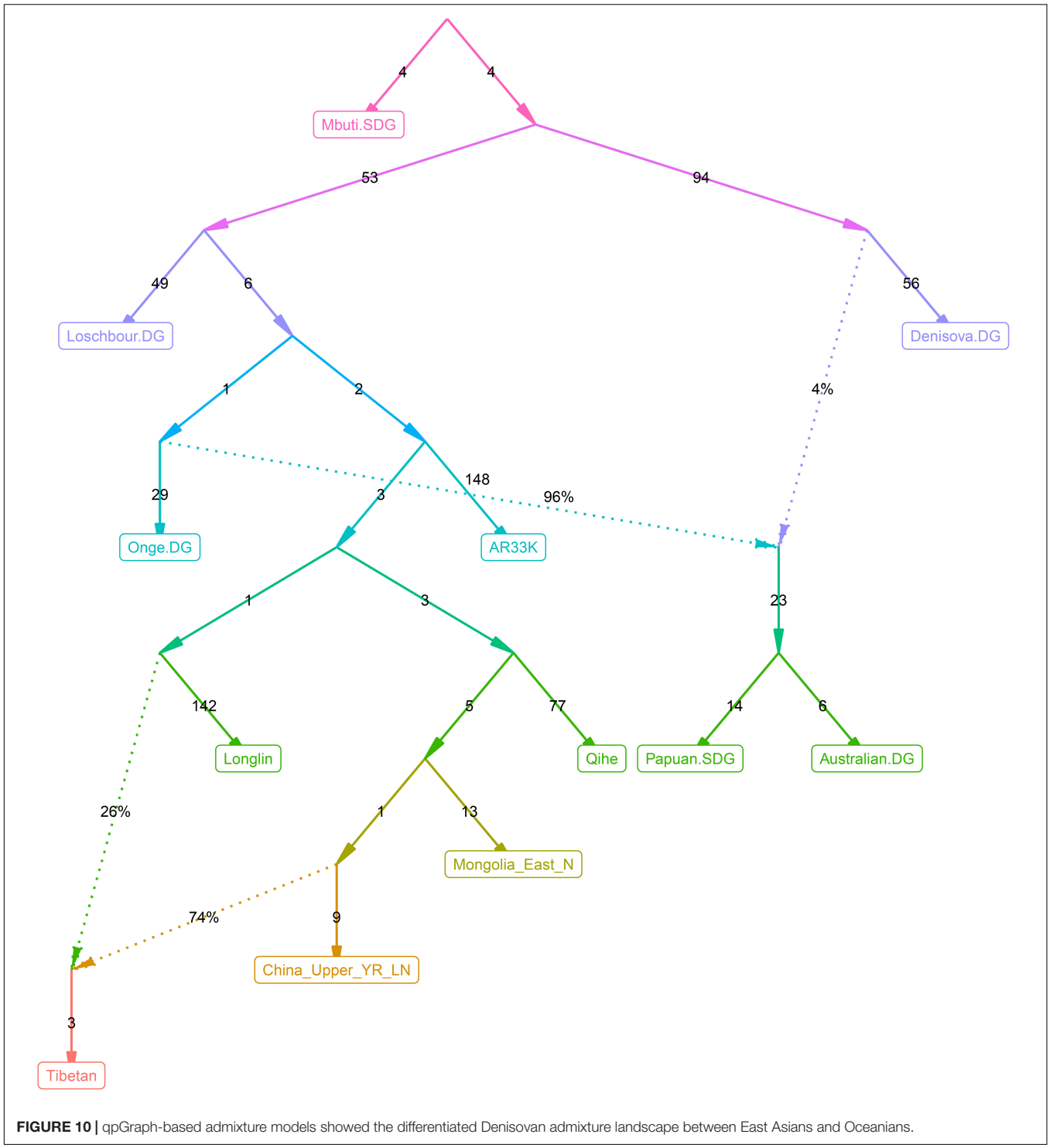

FIGURE 10 | qpGraph-based admixture models showed the differentiated Denisovan admixture landscape between East Asians and Oceanians.

DISCUSSION

Prehistoric human activities and the origin of the high-altitude Tibetans are the research topic in a variety of disciplines, mainly including genetics, archeology, anthropology, history and literature. Recent genome-wide sequencing and paleo-genomic researches have revolutionized our knowledge of the peopling history of Europe (Olalde et al., 2018), Central/South Asia (Narasimhan et al., 2019), America (Nakatsuka et al., 2020), Africa (Skoglund et al., 2017), and Oceania (Lipson et al., 2018b). More and more ancient genomes from the surrounding regions of East Asia have been reported to explore the population dynamics in Southeast Asia (Lipson et al., 2018a; McColl et al., 2018) and South Siberia or Eurasia's

Eastern Steppe (Lazaridis et al., 2014; Raghavan et al., 2014; Mathieson et al., 2015; Damgaard et al., 2018; Sikora et al., 2019), but the ancient genomes in China are still lacking. Fortunately, eight ancient DNA studies from China have been conducted to characterize the deep population history of East Asians. Yang et al. (2017) sequenced 40,000-year-old Tianyuan individual from Beijing and found that the early Asian population substructures have existed before the divergence between East Asians and Native Americans and the peopling of America by anatomically modern human populations. Late Pleistocene and Holocene genomes from the Amur River reported by Mao et al. (2021) demonstrated the genetic transformation among the paleolithic population and their genetic stability in the Neolithic period. Yang et al. (2020) recently conducted another ancient DNA work focused on 24 ancient genomes from Neolithic northern East Asia (eight samples), Neolithic southern East Asia (fifteen samples), and one historic Chuanyun sample, and they found the north-south genetic differentiation among East Asians persisted since the early Neolithic period due to the observed significant genetic differences between Neolithic Shandong and Fujian samples (Yang et al., 2020). Besides, they also identified southward migrations from Shandong Houli populations and northward migrations from Fujian Tanshishan populations, as well as a Neolithic coastal connection from southeastern Vietnam to Russia Far East, and a Proto-Austronesian connection between SEAs and southeast Pacific Vanuatu islanders. The 11,000-year population dynamic documented in Guangxi province showed the extensive admixture between Guangxi, Fujian and Vietnam ancients, which contributed to the formation of pre-agriculture populations (Baojianshan and Dushan) and the affinity between historic Guangxi people and modern Tai-Kadai and Hmong-Mien people (Wang T. et al., 2021). Ning et al. (2020) reported the population history of northern China using fifteen ancient genomes from the Yellow River, West Liao River, and Amur River and discovered that the subsistence strategy changes were associated with the population movement and admixture. Ning and Wang et al. also reported the genomes of ten Iron Age Shirenzigou samples and found the Yamnaya-related steppe pastoralists mediated the population communications between East Asia and western Eurasia, and probably dispersed Indo-European language into Northwest China (Wang et al., 2020). Although these signs of progress have been achieved, the population history, genetic relationship, and genetic differentiation between the highland and lowland modern/ancient East Asians kept in their infancy and remained to be clarified. Thus, we collected nineteen TP-related Neolithic to historic ancients, seventy-eight modern Tibetans from Ü-Tsang, Ando and Kham Tibetan regions, as well as all available eastern Eurasian ancients with different prehistoric human cultural backgrounds as well as modern Eurasians from Indo-European, Altaic, Uralic, Sino-Tibetan, Austronesian, Austroasiatic, Hmong-Mien and Tai-Kadai language families and conducted one comprehensive Paleolithic to present-day ancient/modern genomic meta-analysis. We provided new insights into the peopling of TP and clarify the relationships between high-altitude and lowland ancient/modern East Asians.

There are three hypotheses proposed to elucidate the origin of the Sino-Tibetan language family based on linguistic diversity and others (Zhang et al., 2019). The three hypotheses are North China origin associated with Yangshao/Majiayao hypothesis, Southwest Sichuan origin hypothesis, and Northeast India origin hypothesis. Ancient/modern genomes from the TP showed a clear connection with the northern modern Han Chinese and Neolithic-NEAs, especially with the coastal Houli people from Shandong, inland Yangshao and Longshan people from Henan, and Qijia people from Ganqing region, which supported the northern China origin of modern Tibeto-Burman-speaking populations. Shared ancestry revealed by our PCA, pairwise F_{ST} and outgroup-f_3-values, ADMIXTURE, and f_4-statistics among ancient/modern highlanders and NEA lowlanders showed their close relationship, which was consistent with genetic similarities revealed by the forensic low-density genetic markers and uniparental haplotype/haplogroup data (Zou et al., 2018; Chen P. et al., 2019; He et al., 2019). Direct evidence supported and confirmed this proposed common origin of the Sino-Tibetan (North China origin hypothesis) that was provided by the phylogenetic relationship reconstruction. Both *TreeMix*- and *qpGraph*-based phylogenetic framework supported that the main ancestry in modern Tibetans and ancient TP samples (Nepal and Qijia ancients) was derived from the common NEA lineage related to East Mongolia Neolithic people and Yangshao/Longshan/Houli people from the Central Plain in northern China. Thus, our results in this meta-genomic analysis supported the main lineage that contributed to TP people originated from the Upper and Middle Yellow River with the Neolithic expansion of millet farmers. Our analysis confirmed the origin, diversification, and expansion of the modern Sino-Tibetan populations revealed by the mitochondrial and Y-chromosome variations (Wang L. X. et al., 2018; Li et al., 2019).

Although strong evidence for the common origin of Sino-Tibetan speakers was provided, we still identified the differences in their ancestry composition. Compared with the highlanders on the TP, lowland Late Neolithic to present-day East Asians harbored more ancestry related to Neolithic SEAs and Siberians. Iron Age Dacaozi people from the Gansu-Qinghai region also showed a close genetic affinity with southern people from Tanshishan culture, which indicated the northward dispersal of rice farmers. Compared with the lowland Yangshao/Longshan or coastal Houli populations, the highland populations harbored a certain (8∼14%) proportion of Paleolithic hunter-gatherer ancestry related to the early diverged deep eastern Eurasian lineages (Onge or Hoabinhian related lineages). Lu et al. (2016) illuminated the co-existence of Paleolithic and Neolithic ancestry in modern Tibetans based on the shared haplotypes. Here, we further evidenced the Neolithic and Pre-Neolithic ancestries co-existed in highland East Asians using the allele frequency spectrum in the *f*-statistics (especially for the admixture models of *qpAdm* and *qpGraph*). Thus, our meta-analysis provided new robust evidence for the co-existence of both Paleolithic and Neolithic ancestries in the gene pool of East Asian highlanders as well as the Paleolithic colonization and Neolithic expansion of TP people, which was previously clarified via the modern whole genomes, mitochondrial and

Y-chromosomal data (Qi et al., 2013; Wang L. X. et al., 2018; Li et al., 2019).

Additionally, we also found obvious population substructures among modern Tibetans: Ü-Tsang Tibetans in Tibet core region had predominant original Paleolithic and Neolithic ancestries; Ando Tibetans from Gansu-Qinghai region in northwest China had 2∼3% western Eurasian related ancestry via *qpGraph*-based model; Kham Tibetans from Sichuan and Yunnan provinces possessed a strong southern Neolithic East Asian affinity. Thus, population substructures observed in modern Tibetans were consistent with the geographic and cultural divisions, which suggested that the complex cultural background and terrain to some extent served as the barriers for population movement and admixture. Our *qpGraph*-based phylogeny revealed the gene flow from southern Iron-Age East Asians into Kham Tibetans, from Neolithic NEAs into Kham and Ü-Tsang Tibetans, from western Eurasians into Ando Tibetans, which demonstrated multiple waves of migrant influx from the Siberia, northern and southern East Asia had shaped the gene pool of Tibetan highlanders.

CONCLUSION

We performed a comprehensive genomic meta-analysis focused on Neolithic to present-day people to clarify the relationship between the TP highlanders and lowland East Asians and to explore the peopling of TP. We found a strong genetic affinity between Tibetans and Neolithic to present-day NEAs, which suggested Tibeto-Burman speakers originated from the Upper and Middle YRB in northern China. The observation of the shared ancestry between Han Chinese and Tibetans was consistent with the co-dispersal of millet farmers and Sino-Tibetan languages. Although the shared ancestry persisted between ancient Tibetans and lowland Neolithic people (Yangshao/Longshan/Houli culture), we also found genetic differentiation between them: highland Tibetans harbored more deeply diverged eastern Eurasian Onge-related hunter-gatherer ancestry, but the lowland Neolithic to present-day NEAs possessed more ancestry related to the Neolithic SEAs and Siberians, which not only suggested the co-existence of Paleolithic and Neolithic ancestries in ancient/modern Tibetans but also illuminated the population history of Paleolithic colonization and Neolithic expansion. Besides, consistent with the geographic/linguistic divisions, we identified population substructures in modern Tibetans: more Onge/Hoabinhian related ancestry in Ü-Tsang Tibetans, much more western Eurasian related ancestry in Ando Tibetans, and more Neolithic SEA related ancestry in Kham Tibetan. In short, modern East Asian highlanders derived their ancestry from at least five waves of population admixture: Hoabinhian as the oldest Paleolithic layer; additional gene flow from two Neolithic expansions (inland and coastal) from NEAs, one Neolithic SEA northwestward expansion and one western Eurasian eastward expansion.

SIGNIFICANCE STATEMENT

The Tibetan Plateau has a harsh and extreme high-altitude hypoxic environment, which is inhospitable for human permanent settlement. The population genomic history of modern Tibetans and the population dynamic demographic history of their predecessors need to be comprehensively characterized. We used one large-scale modern and ancient Eurasian meta-dataset to perform genomic analyses focusing on the fine-scale population structure of modern and ancient East Asian highlanders. Firstly, we identified the genomic affinity between highlanders and Neolithic-to-modern Northern East Asians, which was in accordance with the archeologically documented phenomena of Neolithic millet farmer expansion from the Yellow River Basin with the dissemination of Tibeto-Burman languages. Secondly, we identified the obvious population substructure in modern Tibetans along with their cultural division. Thirdly, we documented multiple waves of peopling the Tibetan Plateau and the complex admixture history of East Asian highlanders via the *qpGraph*-based phylogenetic frameworks.

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 were reviewed and approved by this project was inspected and approved by the Medical Ethics Committee of the Xiamen University. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

GH, RT, C-CW, H-YY, and MW conceived the idea for the study. GH, MW, XZ, PC, ZW, YL, HY, and L-HW performed or supervised wet laboratory work and analyzed the data. GH, MW, and XZ wrote and edited the manuscript. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

This study was supported by the China Postdoctoral Science Foundation (2021M691879), National Natural Science Foundation of China (31801040 and 31760309), Nanqiang

Outstanding Young Talents Program of Xiamen University (X2123302), and Fundamental Research Funds for the Central Universities (ZK1144), Foundation for Humanities and Social Sciences Research of the Ministry of Education (18YJAZH116), Scientific Research Project of Colleges and Universities in Gansu Province (2017B-34), Gansu University of Political Science and Law Major Scientific Research Projects (2017XZD10), Lanzhou Talent Innovation and Entrepreneurship Project (2018-RC-113), and Gansu Province Guides Science and Technology Innovation Special Project (2018ZX03).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.725243/full#supplementary-material

REFERENCES

Aldenderfer, M. (2011). Peopling the Tibetan plateau: insights from archaeology. *High Alt. Med. Biol.* 12, 141–147. doi: 10.1089/ham.2010.1094

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

Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S., and Akey, J. M. (2018). Analysis of human sequence data reveals two pulses of archaic denisovan admixture. *Cell* 173, 53–61. doi: 10.1016/j.cell.2018.02.031

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7.

Chen, F., Welker, F., Shen, C. C., Bailey, S. E., Bergmann, I., Davis, S., et al. (2019). A late middle pleistocene denisovan mandible from the Tibetan Plateau. *Nature* 569, 409–412. doi: 10.1038/s41586-019-1139-x

Chen, P., Wu, J., Luo, L., Gao, H., Wang, M., Zou, X., et al. (2019). Population genetic analysis of modern and ancient DNA variations yields new insights into the formation, genetic structure, and phylogenetic relationship of Northern Han Chinese. *Front. Genet.* 10:1045. doi: 10.3389/fgene.2019.01045

Chen, F. H., Dong, G. H., Zhang, D. J., Liu, X. Y., Jia, X., An, C.-B., et al. (2015). Agriculture facilitated permanent human occupation of the Tibetan Plateau after 3600 BP. *Science* 347, 248–250. doi: 10.1126/science.1259172

Damgaard, P. B., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., Korneliussen, T., et al. (2018). 137 ancient human genomes from across the Eurasian steppes. *Nature* 557, 369–374.

Ding, M., Wang, T., Ko, A. M., Chen, H., Wang, H., Dong, G., et al. (2020). Ancient mitogenomes show plateau populations from last 5200 years partially contributed to present-day Tibetans. *Proc. Biol. Sci.* 287:20192968. doi: 10.1098/rspb.2019.2968

Gao, J. Y., Hou, G. L., Wei, H. C., Chen, Y. C., Chongyi, E., Xiaoliang, C., et al. (2020). Prehistoric human activity and its environmental background in Lake Donggi Cona basin, northeastern Tibetan Plateau. *Holocene* 30, 657–671. doi: 10.1177/0959683619895583

Gnecchi-Ruscone, G. A., Abondio, P., De Fanti, S., Sarno, S., Sherpa, M. G., Sherpa, P. T., et al. (2018). Evidence of polygenic adaptation to high altitude from Tibetan and sherpa genomes. *Genome Biol. Evol.* 10, 2919–2930.

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522, 207–211.

He, G., Wang, Z., Su, Y., Zou, X., Wang, M., Chen, X., et al. (2019). Genetic structure and forensic characteristics of Tibeto-Burman-speaking U-Tsang and Kham Tibetan Highlanders revealed by 27 Y-chromosomal STRs. *Sci. Rep.* 9:7739.

He, G. L., Li, Y. X., Wang, M. G., Zou, X., Yeh, H. Y., Yang, X. M., et al. (2020). 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

Hu, X. J., Yang, J., Xie, X. L., Lv, F. H., Cao, Y. H., Li, W. R., et al. (2019). The genome landscape of tibetan sheep reveals adaptive introgression from argali and the history of early human settlements on the Qinghai-Tibetan Plateau. *Mol. Biol. Evol.* 36, 283–303. doi: 10.1093/molbev/msy208

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., et al. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512, 194–197. doi: 10.1038/nature13408

Jeong, C., Balanovsky, O., Lukianova, E., Kahbatkyzy, N., Flegontov, P., Zaporozhchenko, V., et al. (2019). The genetic history of admixture across inner Eurasia. *Nat. Ecol. Evol.* 3, 966–976.

Jeong, C., Ozga, A. T., Witonsky, D. B., Malmstrom, H., Edlund, H., Hofman, C. A., et al. (2016). 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.* 113, 7485–7490. doi: 10.1073/pnas.1520844113

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513, 409–413.

Li, Y. C., Tian, J. Y., and Kong, Q. P. (2015). A dual origin of Tibetans: evidence from mitochondrial genomes. *J. Hum. Genet.* 60, 403–404. doi: 10.1038/jhg.2015.40

Li, Y.-C., Tian, J.-Y., Liu, F.-W., Yang, B.-Y., Gu, K.-S.-Y., Rahman, Z. U., et al. (2019). Neolithic millet farmers contributed to the permanent settlement of the Tibetan Plateau by adopting barley agriculture. *Natl. Sci. Rev.* 6, 1005–1013. doi: 10.1093/nsr/nwz080

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

Lipson, M., Skoglund, P., Spriggs, M., Valentin, F., Bedford, S., Shing, R., et al. (2018b). Population turnover in remote oceania shortly after initial settlement. *Curr. Biol.* 28, 1157–1165. doi: 10.1016/j.cub.2018.02.051

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, W., Chen, F., Cheng, T., Gao, X., Xia, H., Zhang, D., et al. (2020). New advances in the study of prehistoric human activity on the Tibetan Plateau. *Chinese Sci. Bull.* 65, 475–482. doi: 10.1360/tb-2019-0382

Lu, D., Lou, H., Yuan, K., Wang, X., Wang, Y., Zhang, C., et al. (2016). Ancestral origins and genetic history of Tibetan Highlanders. *Am. J. Hum. Genet.* 99, 580–594. doi: 10.1016/j.ajhg.2016.07.002

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

Massilani, D., Skov, L., Hajdinjak, M., Gunchinsuren, B., Tseveendorj, D., Yi, S., et al. (2020). Denisovan ancestry and population history of early East Asians. *Science* 370:579. doi: 10.1126/science.abc1166

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528, 499–503. doi: 10.1038/nature16152

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.

Meyer, M. C., Aldenderfer, M. S., Wang, Z., Hoffmann, D. L., Dahl, J. A., Degering, D., et al. (2017). Permanent human occupation of the central Tibetan Plateau in the early Holocene. *Science* 355, 64–67. doi: 10.1126/science.aag0357

Nakatsuka, N., Lazaridis, I., Barbieri, C., Skoglund, P., Rohland, N., Mallick, S., et al. (2020). A paleogenomic reconstruction of the deep population history of the andes. *Cell* 181, 1131–1145.

Narasimhan, V. M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., et al. (2019). The formation of human populations in South and Central Asia. *Science* 365:eaat7487.

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.

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.e2524. doi: 10.1016/j.cub.2019.06.044

Olalde, I., Brace, S., Allentoft, M. E., Armit, I., Kristiansen, K., Booth, T., et al. (2018). The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* 555, 190–196.

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

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

Qi, X., Cui, C., Peng, Y., Zhang, X., Yang, Z., Zhong, H., et al. (2013). Genetic evidence of paleolithic colonization and neolithic expansion of modern humans on the tibetan plateau. *Mol. Biol. Evol.* 30, 1761–1778. doi: 10.1093/molbev/mst093

Qin, Z., Yang, Y., Kang, L., Yan, S., Cho, K., Cai, X., et al. (2010). A mitochondrial revelation of early human migrations to the Tibetan Plateau before and after the last glacial maximum. *Am. J. Phys. Anthropol.* 143, 555–569. doi: 10.1002/ajpa.21350

Raghavan, M., Skoglund, P., Graf, K. E., Metspalu, M., Albrechtsen, A., Moltke, I., et al. (2014). Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* 505, 87–91. doi: 10.1038/nature12736

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., et al. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468, 1053–1060. doi: 10.1038/nature09710

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

Ren, L., Dong, G., Liu, F., D'alpoim-Guedes, J., Flad, R. K., Ma, M., et al. (2020). Foraging and farming: archaeobotanical and zooarchaeological evidence for Neolithic exchange on the Tibetan Plateau. *Antiquity* 94, 637–652. doi: 10.15184/aqy.2020.35

Shi, H., Zhong, H., Peng, Y., Dong, Y. L., Qi, X. B., Zhang, F., et al. (2008). Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. *BMC Biol.* 6:45. doi: 10.1186/1741-7007-6-45

Sikora, M., Pitulko, V. V., Sousa, V. C., Allentoft, M. E., Vinner, L., Rasmussen, S., et al. (2019). The population history of northeastern Siberia since the Pleistocene. *Nature* 570, 182–188.

Skoglund, P., Thompson, J. C., Prendergast, M. E., Mittnik, A., Sirak, K., Hajdinjak, M., et al. (2017). Reconstructing prehistoric African population structure. *Cell* 171, 59–71.

Wang, C.-C., Yeh, H.-Y., Popov, A. N., Zhang, H.-Q., Matsumura, H., Sirak, K., et al. (2020). The genomic formation of human populations in East Asia. *bioRxiv* 2020:e004606.

Wang, C. C., Yeh, H. Y., Popov, A. N., Zhang, H. Q., Matsumura, H., Sirak, K., et al. (2021). Genomic insights into the formation of human populations in East Asia. *Nature* 591, 413–419.

Wang, T., Wang, W., Xie, G., Li, Z., Fan, X., Yang, Q., et al. (2021). 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

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

Wang, Z., He, G., Luo, T., Zhao, X., Liu, J., Wang, M., et al. (2018). Massively parallel sequencing of 165 ancestry informative SNPs in two Chinese Tibetan-Burmese minority ethnicities. *Forensic Sci. Int. Genet.* 34, 141–147. doi: 10.1016/j.fsigen.2018.02.009

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

Yang, M. A., Gao, X., Theunert, C., Tong, H., Aximu-Petri, A., Nickel, B., et al. (2017). 40,000-year-old individual from Asia provides insight into early population structure in Eurasia. *Curr. Biol.* 27, 3202–3208. doi: 10.1016/j.cub.2017.09.030

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. Genom.* 296, 631–651. doi: 10.1007/s00438-021-01767-0

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X., Pool, J. E., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329, 75–78.

Zhang, C., Lu, Y., Feng, Q., Wang, X., Lou, H., Liu, J., et al. (2017). Differentiated demographic histories and local adaptations between Sherpas and Tibetans. *Genome Biol.* 18:115.

Zhang, D., Xia, H., Chen, F., Li, B., Slon, V., Cheng, T., et al. (2020). Denisovan DNA in late pleistocene sediments from Baishiya Karst Cave on the Tibetan Plateau. *Science* 370, 584–587. doi: 10.1126/science.abb6320

Zhang, D. D., and Li, S. H. (2002). Optical dating of Tibetan human hand- and footprints: An implication for the palaeoenvironment of the last glaciation of the Tibetan Plateau. *Geophys. Res. Lett.* 29, 16.11–16.13.

Zhang, M., Yan, S., Pan, W., and Jin, L. (2019). Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic. *Nature* 569, 112–115. doi: 10.1038/s41586-019-1153-z

Zhang, X. L., Ha, B. B., Wang, S. J., Chen, Z. J., Ge, J. Y., Long, H., et al. (2018). The earliest human occupation ofthe high-altitude Tibetan Plateau 40 thousand to 30 thousand years ago. *Science* 362, 1049–1051. doi: 10.1126/science.aat8824

Zhao, M., Kong, Q. P., Wang, H. W., Peng, M. S., Xie, X. D., Wang, W. Z., et al. (2009). Mitochondrial genome evidence reveals successful Late Paleolithic settlement on the Tibetan Plateau. *Proc. Natl. Acad. Sci. U S A* 106, 21230–21235. doi: 10.1073/pnas.0907844106

Zou, X., Wang, Z., He, G., Wang, M., Su, Y., Liu, J., et al. (2018). Population genetic diversity and phylogenetic characteristics for high-altitude adaptive kham tibetan revealed by DNATyper(TM) 19 amplification system. *Front. Genet.* 9:630.

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.

The reviewer JH declared a past co-authorship with the authors GH, XZ, C-CW, L-HW, H-YY to the handling editor.

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 September 2021
doi: 10.3389/fgene.2021.735084

Genomic Insight Into the Population Structure and Admixture History of Tai-Kadai-Speaking Sui People in Southwest China

Xiaoyun Bin[1†], Rui Wang[2,3,4†], Youyi Huang[1], Rongyao Wei[1], Kongyang Zhu[2,3,4], Xiaomin Yang[2,3,4], Hao Ma[2,3,4], Guanglin He[2,3,4], Jianxin Guo[2,3,4], Jing Zhao[2,3,4], Meiqing Yang[5], Jing Chen[5], Xianpeng Zhang[6], Le Tao[2,3,4], Yilan Liu[2,3,4], Xiufeng Huang[1*] and Chuan-Chao Wang[2,3,4*]*

[1] College of Basic Medical Sciences, Youjiang Medical University for Nationalities, Baise, China, [2] State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen, China, [3] Department of Anthropology and Ethnology, Institute of Anthropology, School of Sociology and Anthropology, National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China, [4] State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen, China, [5] Department of Forensic Medicine, Guizhou Medical University, Guiyang, China, [6] Jinzhou Medical University, Jinzhou, China

OPEN ACCESS

Edited by:
Alison G. Nazareno, Federal University of Minas Gerais, Brazil

Reviewed by:
Shaoqing Wen, Fudan University, China
Jatupol Kampuansai, Chiang Mai University, Thailand

***Correspondence:**
Rui Wang
17786126601@163.com
Xiufeng Huang
hxflcw@163.com
Chuan-Chao Wang
wang@xmu.edu.cn

[†] These authors have contributed equally to this work

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

Received: *02 July 2021*
Accepted: *23 August 2021*
Published: *20 September 2021*

Citation:
Bin X, Wang R, Huang Y, Wei R, Zhu K, Yang X, Ma H, He G, Guo J, Zhao J, Yang M, Chen J, Zhang X, Tao L, Liu Y, Huang X and Wang C-C (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

Sui people, which belong to the Tai-Kadai-speaking family, remain poorly characterized due to a lack of genome-wide data. To infer the fine-scale population genetic structure and putative genetic sources of the Sui people, we genotyped 498,655 genome-wide single-nucleotide polymorphisms (SNPs) using SNP arrays in 68 Sui individuals from seven indigenous populations in Guizhou province and Guangxi Zhuang Autonomous Region in Southwest China and co-analyzed with available East Asians *via* a series of population genetic methods including principal component analysis (PCA), ADMIXTURE, pairwise Fst genetic distance, *f*-statistics, *qpWave*, and *qpAdm*. Our results revealed that Guangxi and Guizhou Sui people showed a strong genetic affinity with populations from southern China and Southeast Asia, especially Tai-Kadai- and Hmong-Mien-speaking populations as well as ancient Iron Age Taiwan Hanben, Gongguan individuals supporting the hypothesis that Sui people came from southern China originally. The indigenous Tai-Kadai-related ancestry (represented by Li), Northern East Asian-related ancestry, and Hmong-Mien-related lineage contributed to the formation processes of the Sui people. We identified the genetic substructure within Sui groups: Guizhou Sui people were relatively homogeneous and possessed similar genetic profiles with neighboring Tai-Kadai-related populations, such as Maonan. While Sui people in Yizhou and Huanjiang of Guangxi might receive unique, additional gene flow from Hmong-Mien-speaking populations and Northern East Asians, respectively, after the divergence within other Sui populations. Sui people could be modeled as the admixture of ancient Yellow River Basin farmer-related ancestry (36.2–54.7%) and ancient coastal Southeast

Asian-related ancestry (45.3–63.8%). We also identified the potential positive selection signals related to the disease susceptibility in Sui people *via* integrated haplotype score (iHS) and number of segregating sites by length (nSL) scores. These genomic findings provided new insights into the demographic history of Tai-Kadai-speaking Sui people and their interaction with neighboring populations in Southern China.

Keywords: genetic substructure, genetic admixture, population history, Tai-Kadai-speaking Sui people, East Asia

INTRODUCTION

Southwest China is home to diverse ethnic minorities and linguistic families. Previous population genetic studies based on genetic markers including mitochondrial DNA (mtDNA) and Y-chromosome haplogroups, short tandem repeats (STR), insertion/deletion polymorphisms (InDels), and genome-wide single-nucleotide polymorphisms (SNPs) shed light on the genetic profile and demographic history of ethnolinguistic groups from southern China and Southeast Asia (Wen et al., 2004; Kutanan et al., 2017; Liu C. et al., 2018; He et al., 2020; Huang et al., 2020; Liu D. et al., 2020; Liu J. et al., 2020; Liu Y. et al., 2020; Wang C. C. et al., 2021). A SNP chip-based population study from the genomic perspective demonstrated that the genetic profile of Tai-Kadai-speaking Hainan Li from southernmost China (referred to as Hlai) was less affected by the Neolithic farming expansion or historical migration compared with other mainland Tai-Kadai-speaking populations; Hlai-related lineage contributed a large proportion of the ancestry to the mainland Tai-Kadai-speaking populations (He et al., 2020). Huang et al. (2020) found that the bidirectional gene flow between Tai-Kadai- and Hmong-Mien-speaking populations formed the Hmong-Mien/Tai-Kadai cline; Tai-Kadai-related groups also had a strong impact on the genetic makeup pattern of populations in Mainland Southeast Asia in the recent two millennia. From the ancient genomic perspective, Wang C. C. et al. (2021) reconstructed the deep population history of East Asia and found a kind of ancestry probably related to the Neolithic Yangtze River farmers, which contributed widely to present-day Austronesian speakers and Tai-Kadai speakers. Yang et al. (2020) revealed that Neolithic Fujian-related ancestry contributed substantially to the present-day Southern Chinese and Southeast Asians; during the Early Neolithic period, Northern East Asians related to Coastal Shandong-related ancestry migrated southward and shifted the genetic makeup of populations from southern China. Additionally, Wang T. et al. (2021) recently reported that Guangxi Longlin-related ancestry, Fujian Qihe-related ancestry, and deep East Asian Hoabinhian-related ancestry participated in the formation of Early Neolithic Guangxi ancients (represented by Dushan/Baojianshan) but limitedly contributed to present-day Southeast Asians; the historical Guangxi samples possessed a genetic profile similar to that of present-day Hmong-Mien- and Tai-Kadai speaking populations.

Our studied population, Tai-Kadai-speaking Sui people, is officially recognized as one of the 56 ethnic groups in China. The Chinese "Sui" means "Water," reflecting the living environment and lifestyles of Sui people. More than 80% of Sui people inhabited Guizhou, one of the most ethnolinguistically diverse provinces in southwest China; the rest of the Sui resided in adjacent provinces in China, such as Guangxi, Yunnan, and Sichuan (2010 Census). According to the historical accounts, the ancestors of Tai-Kadai-speaking populations were ancient Baiyue tribes, the indigenous people living in southern China. Forced by warfare and famine during the Qin dynasty (circa second century B.C.), Chinese Han continuously expanded toward the south for a long time. A great many Baiyue people migrated to southwest China and then formed the Tai-Kadai-speaking people (Gao, 2002; Zhang and Zhang, 2018). Published genetic evidence invalidated the origin of Tai-Kadai-speaking Sui people. The maternally inherited mtDNA HVS-1 region analysis showed Sui had high frequencies of mtDNA haplogroups, which were dominant in southern China [B (B4a, 3.3%; B4b1, 6.7%; and B5a, 20%), M7 (M*, 6.7%; M7b*, 6.7%; M7b1, 6.7%; and M8a, 6.7%), F (F*, 3.3%; F1a, 20%; and F3, 13.3%), and R (R9b, 3.3%)] (Li et al., 2007a). From the paternal Y-chromosome side, Li et al. investigated the haplotype network of Y-STRs, showing that the haplotype O1a-M119 was the dominating haplogroup in Tai-Kadai-speaking Sui (F, 8%; K, 10%; O1a*, 18%; O2a*, 44%; and O3a5, 20%), as well as in Taiwan aborigines, but not in other East Asians, indicating the Sui was native to southern China (Li et al., 2008).

Previous genetic findings were predominately based on autosomal/X/Y STRs, mainly aimed to evaluate the forensic characterization of STR markers and investigate the genetic relationships between the Guizhou Sui people and the surrounding Tai-Kadai-speaking, Hmong-Mien-speaking Miao, and Sinitic populations living in southern China (Yang et al., 2012; Ji et al., 2017; Chen et al., 2018; Guo et al., 2019). Thus, the fine-scale genetic structure and admixture history, the potential positive selection signals of Tai-Kadai-speaking Sui populations, especially Guangxi Sui, are still underrepresented owing to a lack of genome-wide data. In this study, we genotyped 498,655 SNPs of a total of 68 Sui individuals from seven populations in Southwest China and compared them with the published SNP dataset of present-day and ancient East Asians in order to advance the understanding of the demographic history of Sui people from a genomic perspective.

Abbreviations: SNP, single-nucleotide polymorphism; HO, Human Origin; PC, principal component; N, Neolithic; EN, Early Neolithic; MN, Middle Neolithic; LN, Late Neolithic; BA, Bronze Age; IA, Iron Age; H, Historical; SEA, Southeast Asia; EA, East Asia; TK, Tai-Kadai; AN, Austronesian; AA, Austroasiatic; HM, Hmong-Mien; YR, Yellow River.

MATERIALS AND METHODS

Sampling, Genotyping, and Quality Control

A total of 68 blood samples were collected from the seven populations from the Guangxi Zhuang Autonomous Region and Guizhou province with written informed consent and genotyped using Affymetrix WeGene V1 array. These samples were collected randomly from unrelated participants whose parents and grandparents are indigenous people and have a non-consanguineous marriage of the same ethnical group for at least three generations. The ethnicities of all participates were used as their self-declaration based on their family migration history and corresponding family records. Our study and sample collection were reviewed and approved by the Medical Ethics Committee of Youjiang Medical University for Nationalities and Xiamen University (approval number: XDYX2019009) and followed the recommendations provided by the revised Helsinki Declaration of 2000. After removing batch effects and missing sites, we genotyped 498,655 genome-wide SNPs. We listed the detailed sample information in **Supplementary Table 1** and plotted the geographic sampling locations in **Figure 1**.

We further calculated the kinship coefficient *via* the GCTA software (Yang et al., 2011) using options "--autosome --make-grm" and then conducted PLINK1.9 (Purcell et al., 2007) with the option "--missing" to calculate the SNP calling rate for each individual and "--remove" to exclude the individuals with the lowest SNP calling rate and had up to third-degree kinship with other collected samples (kinship coefficient > 0.125). The genetic relationship matrix (GRM) was displayed in **Supplementary Figure 1**. Finally, we got 58 unrelated Sui individuals for further study.

Merging Data

We merged our newly collected samples with published genome-wide SNP data of present-day and ancient East Asian and Southeast Asian populations (Patterson et al., 2012; Huang et al., 2018, 2020; Lipson et al., 2018; McColl et al., 2018; Liu Y. et al., 2020; Ning et al., 2020; Yang et al., 2020; Wang T. et al., 2021;[1]). Two datasets were used in subsequent population genetic analysis: (1) we merged our newly collected Sui individuals with a 1240K capture dataset to create the high-SNP-density "merged-1240K" dataset harboring 373,933 SNP sites; and (2) we merged our newly published Sui people data with 82 present-day populations or nine meta-populations and 40 ancient groups from the Affymetrix Human Origins (HO) panel to generate the "merged-HO" dataset, covering lower SNP sites (119,349) but maximum the number of populations and individuals. Data merging was done by *mergeit* from EIGENSOFT (Patterson et al., 2006).

Principal Component Analysis

We carried out PCA on the "merged-HO" dataset by the *smartpca* program of EIGENSOFT (Patterson et al., 2006) with the default parameters and lspproject: YES. We only used modern populations to construct PCs and then projected the ancient samples onto the top two PCs.

ADMIXTURE

Before the ADMIXTURE analyses, we pruned SNPs on the "merged-HO" dataset in strong linkage disequilibrium with each other using PLINK1.9 (Purcell et al., 2007) by parameters --indep-pairwise 200 25 0.4 and then ran ADMIXTURE (Alexander et al., 2009) with default parameters from $K = 2$ to 12. The cross-validation error reached the lowest point at $K = 4$ (**Supplementary Figure 2**).

Pairwise Fst Genetic Distance

Modern populations which harbored genomic information of more than five individuals on the "merged-HO" dataset were used to calculate pairwise Fst following Weir and Cockerham (1984). We estimated Fst by *smartpca* using EIGENSOFT (Patterson et al., 2006) with default parameters and inbreed: YES and fsonly: YES. The neighbor-joining (N-J) phylogenetic relationship was constructed using Mega 7.0 (Kumar et al., 2016). Populations in the same clade or branches indicated that they had closer relationships than populations in different clades.

f3-Statistics and *f4*-Statistics

We used the *qp3pop* and *qpDstat* packages implemented in ADMIXTOOLS (Patterson et al., 2012) with default parameters to calculate the *f*-statistics. Statistical significance was assessed using the default blocked jackknife approach implemented in ADMIXTOOLS. Outgroup-f_3*(X, Y; outgroup)* calculated the shared genetic drift between X and Y since their divergence from the outgroup. *Admixture-*f_3*(source1, source2; target)* evaluated the admixture signals in the targets (Z-score < -3). In the form of f_4*(outgroup, W; X, Y)*, a Z-score > 3 implied that W shared more alleles with Y than with X; a Z-score less than -3 suggested that W shared extra alleles with X compared with Y; a $|Z\text{-score}| < 3$ indicated that X and Y formed a clade in relation to the outgroup and W. We used an African population Yoruba as the outgroup.

Genetic Homogeneity Testing by *qpWave*

Pairwise *qpWave* Test

We used pairwise *qpWave* as implemented in ADMIXTOOLS (Patterson et al., 2012) on the "merged-1240K" dataset to test whether pairwise populations were genetically homogeneous in relation to a set of outgroups. We used Mbuti, Mongolia_N_East, DevilsCave_N, Ami, Liangdao2, and Vietnam_LN as outgroups because those groups were unlikely to have recent gene flow with our studied Sui people and might be differently related to the ancestral sources of Sui people. A p-value > 0.05 for rank = 0 suggested that pairwise populations were homogeneous genetically relative to outgroups. A p-value < 0.05 for rank = 0 indicated that a minimum of two streams of ancestry were needed to relate pairwise groups to the outgroups.

Outgroup-Dropping Pairwise *qpWave* Test

We did the "outgroup-dropping" test in which we dropped one of the populations in the outgroup set by turn to investigate

[1]https://reich.hms.harvard.edu/downloadable-genotypes-present-day-and-ancient-dna-data-compiled-published-papers

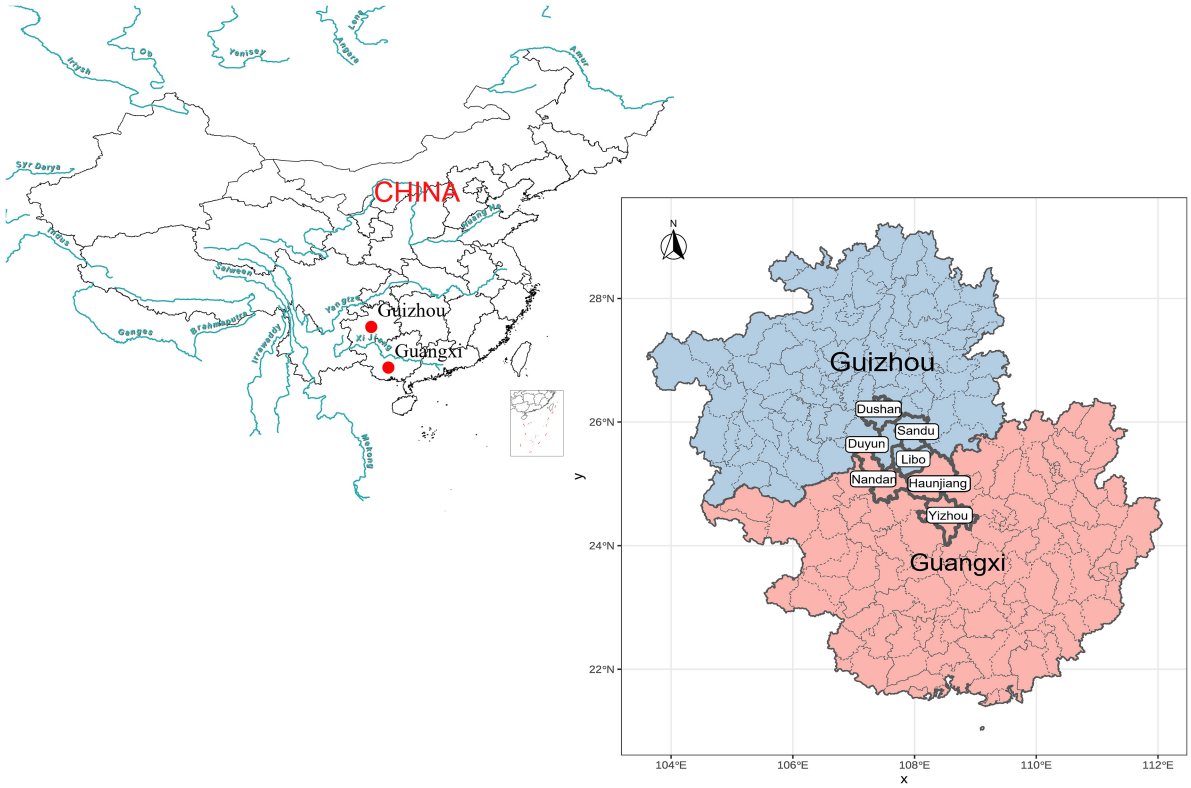

FIGURE 1 | Geographical positions of new collected Sui people from Guangxi and Guizhou Province, Southwest China.

which outgroups may lead to the nonhomogeneity between the pairwise-tested populations.

Admixture Coefficient Modeling by *qpAdm*

We used *qpAdm* as implemented in ADMIXTOOLS (Patterson et al., 2012) with default parameters and all snps: YES to estimate admixture proportions for one target population as a combination of N-specified source populations by exploiting the shared genetic drift with a set of outgroups. The models with a p-value > 0.05, nested p-value < 0.05, and admixture proportions estimated between (0, 1) were accepted.

Two-Way Admixture Model

We used ancient Northern East Asian-related ancestry (represented by YR_LN) and ancient Southeast Asian-related ancestry (represented by Liangdao2) as sources. Seven populations (Mbuti, Tianyuan, Papuan, Onge, DevilsCave_N, Japan_Jomon, and Mongolia_N_East) were used as outgroups.

Three-Way Admixture Model

We further used YR_LN, Ami, Vietnam_N as proxies for ancient Northern East Asian, ancient coastal Southeast Asian, and ancient inland Southeast Asian-related ancestries. Nine populations (Mbuti, Tianyuan, Papuan, Onge, Liangdao2, DevilsCave_N, Japan_Jomon, Mongolia_N_East, and Malaysia_LN) were used as outgroups.

Detecting the Positive Natural Selection Signals

Before identifying the natural selection, we used PLINK1.9 (Purcell et al., 2007) to remove individuals whose SNP-missing-rate is greater than 10% with parameter --geno 0.1 and then applied ShapeIT with default parameters (Delaneau et al., 2013) to phase autosomal SNP data of the Sui people. We calculated the integrated haplotype score (iHS) (Voight et al., 2006) and number of segregating sites by length (nSL) (Ferrer-Admetlla et al., 2014) for each phased SNP site (377,197) *via* the selscan software with default parameters (Szpiech and Hernandez, 2014). Then we used selscan's norm package to normalize the scores within every 100 bins of allele frequency. A total of 1,829 SNPs with both absolute normalized iHS and an nSL greater than the threshold (top 1% iHS: 2.566430; top 1% nSL: 2.533570) were regarded as the candidate sites under natural selection. We then perform (1) the gene annotation *via* 3DSNP (Lu et al., 2016) and (2) KEGG analysis *via* Kobas (Bu et al., 2021).

RESULTS

Investigating the Population Structure of Studied Sui Populations

We first carried out PCA to uncover an overview of the genetic structure of East Asians and Southeast Asians (**Figure 2**). The

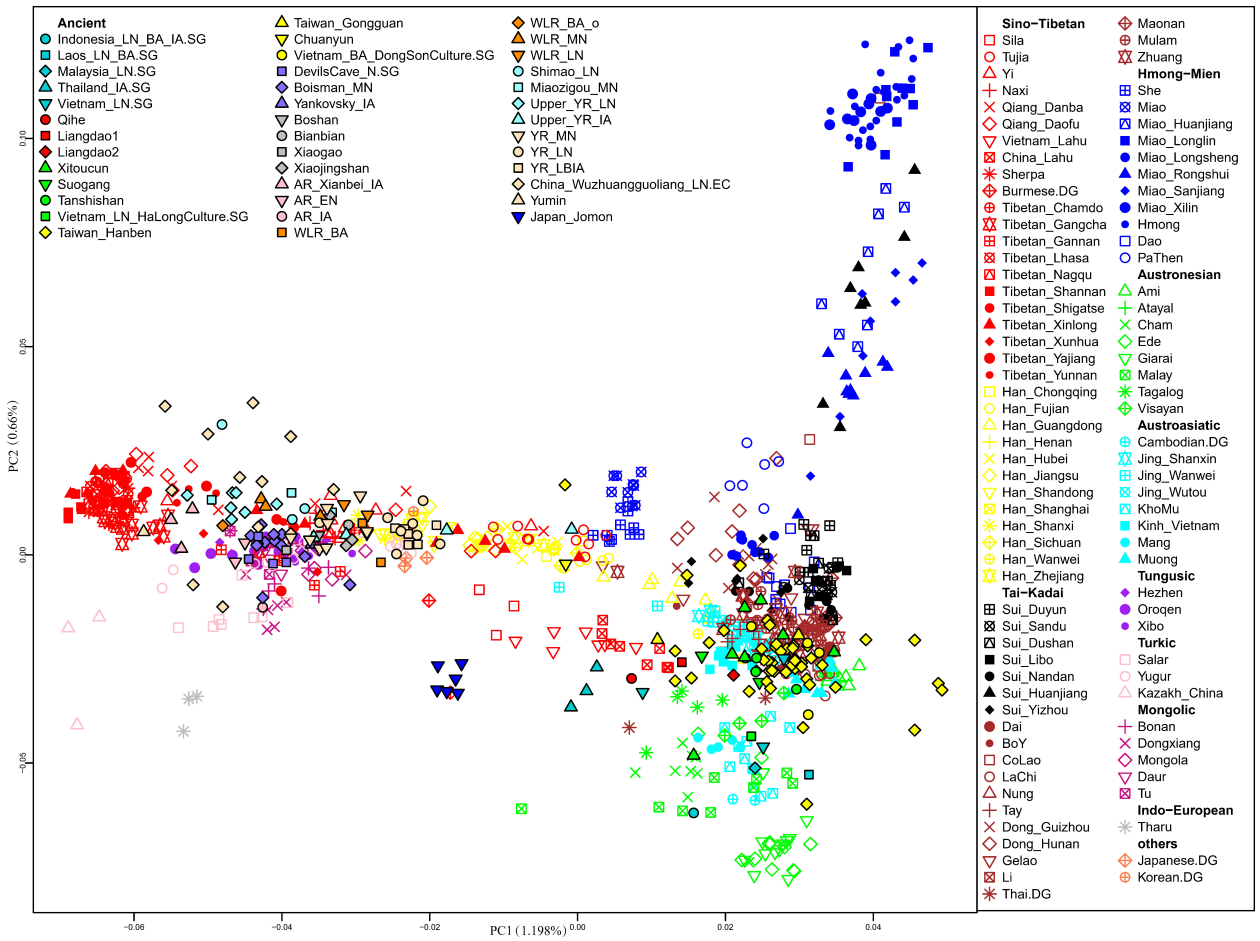

FIGURE 2 | Principal component analysis (PCA) was constructed based on our studied Sui groups and published present-day East Asian and Southeast Asian populations, then projected ancient East Asians onto the first two principal components.

top two PCs (PC1 and PC2) revealed that within present-day populations, individuals from the same linguistic classification and geographic locations were mostly placed together, displayed as the following genetic clusters: Altaic-related (Tungusic speaking, Turkic speaking, and Mongolic speaking); Tibetan-Burman-related; Japanese- and Korean-related; Han-related North-South cline; and Southeast Asian-related clusters which comprised Hmong-Mien-related, Austronesian-related, Austroasiatic-related, and Tai-Kadai-related populations.

The distributions of the studied Sui people were relatively scattered, with two major clusters: one genetic cline consisting of Huanjiang Sui individuals from Guangxi, which was placed on the Hmong-Mien-related cline; and the other comprising the rest of the newly studied Sui groups, clustered together with published Tai-Kadai-speaking populations and displaying closer genetic relationships with AA, AN, and Sinitic-speaking populations from Southern China and Southeast Asia. Yizhou Sui from Guangxi slightly shifted toward the Han-related cline. Furthermore, we detected that the Sui-related cluster tended to deviate to HM-related cline compared with other published TK speakers; Tibetan-Burman-related and Altaic-related groups exhibited significant genetic differentiation with the studied Sui people. After projecting the ancient EA and SEA individuals onto the top two PCs, we observed that most ancient individuals fall relatively close with geographically close modern populations. Coastal Fujian_EN (Liangdao1, Liangdao2, and Qihe), Fujian_LN (Xitoucun, Suogang, and Tanshishan), Taiwan_IA (Taiwan_Hanben and Taiwan_Gongguan), Inland Vietnam Bronze Age (Vietnam_BA_DongSonCulture) individuals plotted relatively close with our studied Sui populations than other published ancient samples.

Model-based ADMIXTURE with an optimal K of 4 (**Figure 3**) suggested that four major EA and SEA ancestral components can adequately explain the genetic makeup of the studied populations: (1) The Inland Southeast Asian-related component (noted as pink) was maximized in Late Neolithic ancient SEAs from Vietnam, Malaysia, Laos, and modern Austroasiatic speakers Mang from Vietnam. This lineage also reached high proportions in the populations from southern China and Southeast Asia, such as Tai-Kadai speakers Dai and Austroasiatic speakers Jing. (2) The Hmong-Mien-related ancestry (denoted as yellow) was dominant in Hmong from

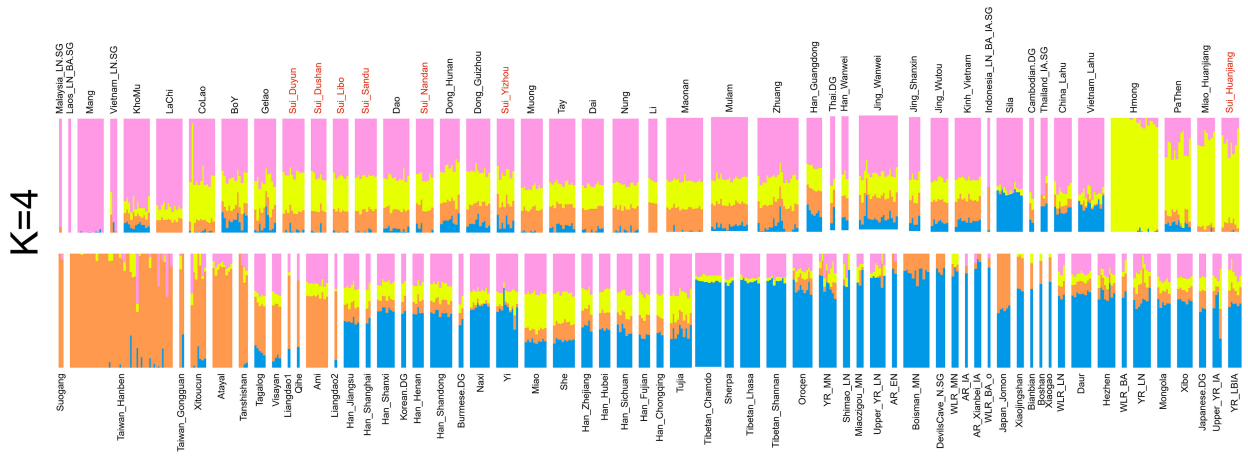

FIGURE 3 | Model-based ADMIXTURE results showed the individual ancestry component composition of studied Sui and reference East Asian and Southeast Asian individuals with the predefined four ancestral sources (K = 4).

Vietnam and Huanjiang_Miao from southern China. (3) The coastal Southeast Asian-related ancestry (denoted as orange) was enriched in the Neolithic Fujian populations, Iron Age Gongguan, and Hanben individuals from Taiwan as well as present-day indigenous Austronesian-speaking Ami and Atayal. (4) The Northern East Asian-related (denoted as blue) ancestry maximized in Tibetans and was also widely distributed in Sinitic and Altaic speakers and ancient Northeast Asians. The ADMIXTURE model in K = 4 revealed the difference in genetic compositions within Sui populations. The studied Huanjiang Sui individuals were characterized by a considerable amount of Hmong-Mien-related ancestry (∼70%), with ∼22% Inland Southeast Asian-related ancestry, ∼7% coastal Southeast Asian-related ancestry, and very few Northern East Asian-related ancestry (less than 1%). While the primary ancestry component assigned to other studied Sui groups as well as neighboring TK and AA speakers was the Inland southeast Asian-related component, they harbored less Hmong-Mien-related ancestry but higher proportions of coastal Southeast Asian-related and Northern East Asian-related ancestry compared with Huanjiang Sui. In addition, Yizhou Sui harbored a significantly higher Northern East Asian-related component than each Sui population (Yizhou Sui: mean = 8.5% versus Sui_Dushan: mean = 1.43%; Sui_Duyun: mean = 1.9%; Sui_Huanjiang: mean = 0.28%; Sui_Libo: mean = 0.56%; Sui_Nandan: mean = 2.79%; Sui_Sandu: mean = 0.88%; $p < 0.02102$, Student's t-test).

Population Relationships Between Studied Sui Groups and Worldwide Reference Populations

To explore the genetic affinity between the studied Sui and reference populations, we first constructed the unrooted N-J phylogenetic tree based on Wright's fixation index pairwise Fst genetic distance among 79 modern populations (**Supplementary Figure 3**). We identified two main genetic branches highly correlated to the geographic locations; the sub-clades also corresponded well to the linguistic classifications: (1) the Northern East Asian-related one, which was made up of the Altaic-, Tibetan- Burman-, and Sinitic-speaking populations; and (2) the Southeast Asian-related cluster composed of the HM-, TK-, AN-, and AA-speaking populations. The Sui people showed a relatively close phylogenetic relationship with neighboring Tai-Kadai-speaking populations, such as Maonan, Dong, and CoLao. We observed the strong genetic assimilation within Sui groups except for Huanjiang Sui, which first clustered with HM speakers (Huanjiang Miao, Hmong, PaThen, and Hunan Miao) and then with TK-speaking CoLao and Dong, followed by other studied Sui populations.

We subsequently conducted outgroup-f_3-statistics in the form of f_3*(X, Studied Sui; Yoruba)* to measure the shared genetic drift. The cluster patterns in the heatmap (**Figure 4**) confirmed that the studied Sui had a striking genetic affinity with each other and with present-day populations from southern China and Southeast Asia, especially Tai-Kadai speakers Maonan and Dong, Hmong-Mien speaker Hmong, and Austroasiatic speaker Jing. Huanjiang Sui individuals shared the highest genetic drift with Huanjiang Miao (f_3: 0.237811), followed by Hmong (f_3: 0.235770) and then by other Sui people (f_3: 0.222949–0.225170). Focusing on the results of *outgroup-f_3(Studied Sui, ancient individuals; Yoruba)* (**Figure 5**), we found that each Sui group displayed similar patterns of genetic affinity with ancient reference populations. The top highest shared drift with the studied Sui was provided by Iron Age Hanben and Gongguan samples from Taiwan ($f_3 > 0.223575$), followed by Fujian Historical Chuanyun individuals ($f_3 > 0.220979$), Fujian_LN Xitoucun individuals ($f_3 > 0.220718$), and inland YR basin farmer populations ($f_3 > 0.215268$).

To quantitatively evaluate the genetic similarity and differentiation among the studied Sui populations compared with the worldwide reference populations, we performed symmetrical f_4-statistics in the form of f_4*(Yoruba, X; Sui population 1, Sui*

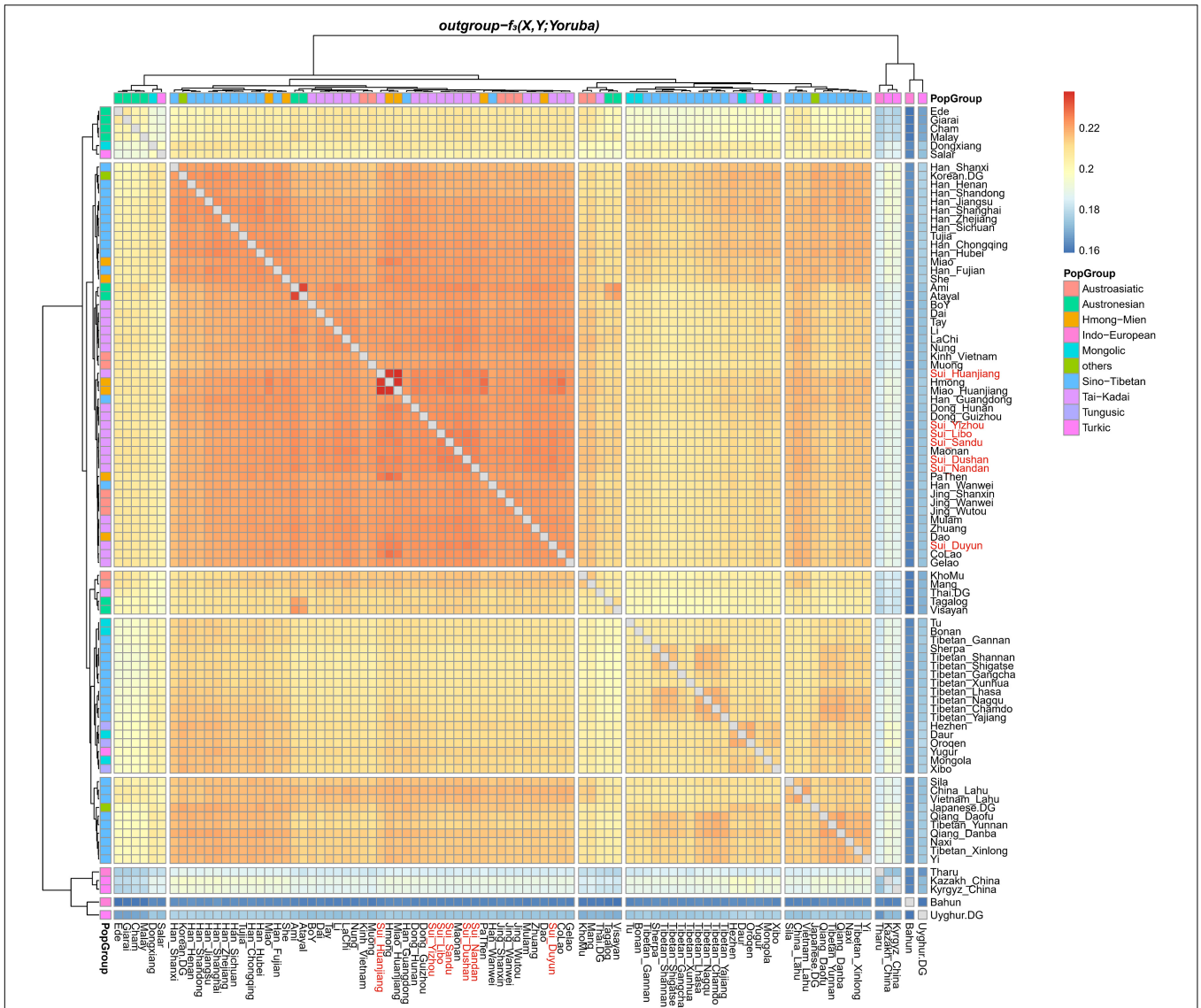

FIGURE 4 | Heatmap showed the sharing genetic drift between the present-day East Asians estimated *via outgroup-f*$_3$ *(modern East Asian 1, modern East Asian2; Yoruba)*. Red indicated more shared genetic drift among the pairwise populations since their divergence from an African outgroup Yoruba.

population 2) shown in **Supplementary Table 2**. The observed significant negative f_4 values with absolute *Z*-scores larger than 3 indicated that X shared more genetic drift with *Sui population 1* relative to *Sui population 2*; otherwise, significant positive f_4 values indicated more shared alleles between X and *Sui population 2* rather than *Sui population 1*. *Z*-scores ranging from (−3, 3) denoted that Sui 1 and Sui 2 formed one clade in relation to X and outgroup Yoruba, respectively. The observed significant *Z*-scores in f_4*(Yoruba, X; Sui_Huanjiang, Sui population 2)* suggested that HM-speaking Hmong, Miao_Huanjiang, and PaThen shared excess alleles with Huanjiang Sui individuals ($-18.139 \leq Z\text{-scores} \leq -4.691$). Tai-Kadai-speaking Li, Mulam, Nung, and Tay and AA-speaking Kinh and Muong shared fewer alleles with Huanjiang Sui than with other Sui people ($1.138 \leq Z\text{-scores} \leq 3.738$). The significant negative *Z*-scores in f_4*(Yoruba, Northern East Asian; Sui_Yizhou, Sui population 2)* indicated that there are significantly more derived alleles shared between Northern East Asians (such as Iron Age Amur River Basin-related_Xianbei populations, coastal Siberia Boisman_MN, DevilsCave_N, and Tibetan_Shannan) and Guangxi Yizhou Sui compared with Sui_Libo (*Z*-scores = −3.143, −3.045), Sui_Dushan (*Z*-scores = −3.085), and Sui_Duyun (*Z*-scores = −3.000). AN-speaking Ede shared excess alleles with Sui_Duyun (*Z*-scores = 3.101) and Sui Libo (*Z*-scores = 3.048) compared with Sui Yizhou. HM-speaking Hmong shared more derived alleles with Sui_Duyun than with Sui_Yizhou (*Z*-scores = 3.355). We did not identify the genetic difference among Guizhou Sui groups and Guangxi Sui_Sandu as no f_4*(Yoruba, X; Guizhou Sui population 1, Guizhou Sui population 2/Sui_Sandu)* with significant *Z*-scores were observed.

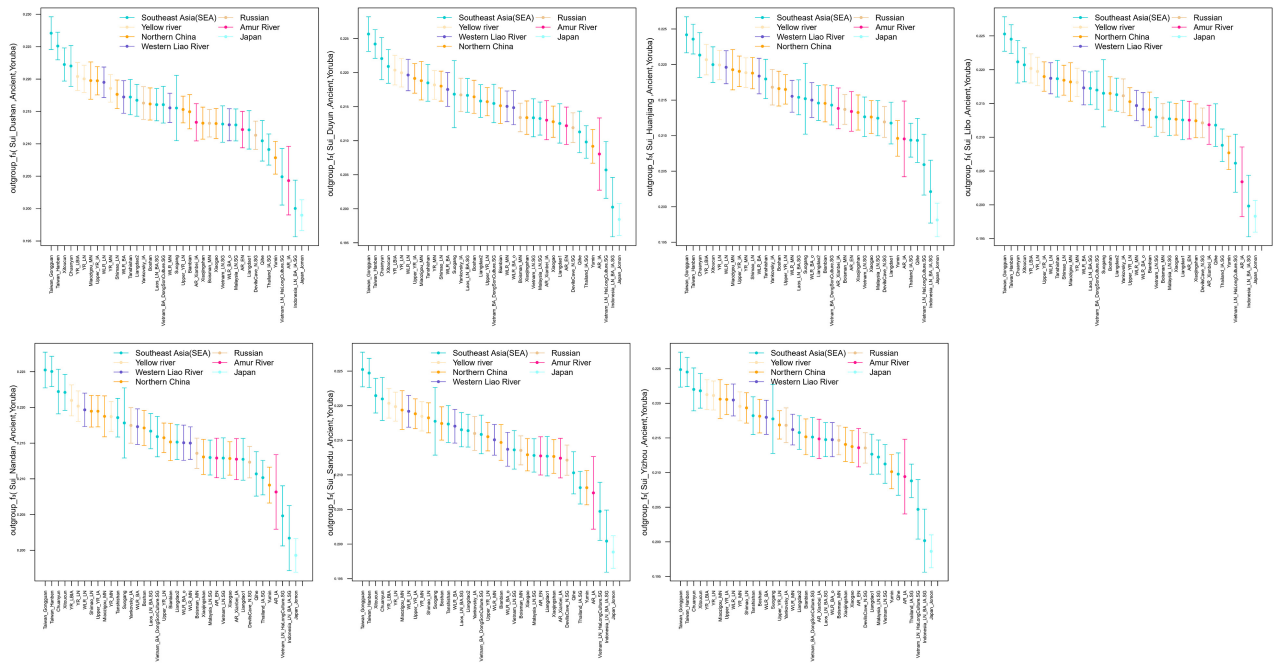

FIGURE 5 | Error bar plot showed the genetic affinity between studied Sui populations and ancient East Asians revealed by *outgroup-f*$_3$ *(ancient reference East Asians, Studied Sui; Yoruba)*. The error bar is marked as the standard deviation.

We then performed pairwise *qpWave* analysis, which was more accurate than *symmetry* f_4*-statistics*, to further test whether seven Sui groups were genetically homogeneous (**Figure 6**). Each pair of Guizhou Sui populations had a p-value > 0.05 for rank = 0. We observed p-values < 0.05 (rank = 0) in the one-way admixture model when Guangxi Nandan, Huanjiang, and Yizhou Sui were used as one of the test populations. Pairwise *qpWave* results confirmed that (1) there is genetic homogeneity within Guizhou Sui individuals; (2) there is genetic heterogeneity within Guangxi Sui individuals; and (3) Guizhou Sui and Guangxi Sui were not derived from a single homogenous population in relation to the outgroups we used.

We next did the outgroup-dropping pairwise *qpWave* test (**Supplementary Figure 4**). Pairwise test populations which had genetic heterogeneity relative to the full set of outgroups showed a p-value > 0.05 (rank = 0) in the "outgroup-dropping pairwise *qpWave* test" instead, suggesting that the "dropped outgroup"-related ancestry might have a unique gene flow with one of the test groups, explaining the nonhomogeneity between the pairwise test populations. The results suggested that different levels of Ami-related gene flow may lead to the heterogeneity between Guangxi Sui and Guizhou Sui; coastal Amur River Neolithic DevilsCave-related ancestry may drive the nonhomogeneity between Yizhou Sui and other Sui groups.

The Ancestry Inference of the Studied Sui People

We performed all possible f_4-statistics in the form of f_4*(Yoruba, X; studied group, Y)* for each Sui group to explore the possible extra gene flow that Sui people/Y received from X after the divergence between the specific population Y and studied Sui.

When Tai-Kadai-speaking Hlai (representing the unadmixed form of Tai-Kadai-speaking populations) was same as Y (**Figures 7A,C**), the observed significant negative Z-scores suggested that Hmong-Mien speakers (Hmong and Miao_Huanjiang, $-17.174 \leq Z\text{-scores} \leq -2.139$) and Tibetan-Burman-related populations (such as Sherpa, $-4.103 \leq Z\text{-scores} \leq -2.120$) shared more derived alleles with our newly reported Sui people, indicating that the Sui people received the extra Hmong-Mien-related ancestry and northern East Asian sources after the separation of the Hlai and Sui groups or explaining that the Hlai received the ancestry which was a deeper linage than the common ancestor of Hmong-Mien-speaking populations and Sui people. No significant positive f_4 values were observed.

As shown by the Fst-based N-J tree and *outgroup-f*$_3$ statistics, the geographically close Tai-Kadai-speaking Maonan firstly clustered with Sui groups. Thus, we further did the formal test in the form of f_4*(Yoruba, reference population; Sui, Maonan)* to explore the fine-scale genetic differentiation between the studied Sui and Maonan (**Figures 7B,D**, respectively). We observed that Guizhou Sui_Libo, Sui_Sandu, Sui_Dushan, and Guangxi Sui_Nandan had similar genetic profiles with Maonan (all $|Z\text{-scores}| < 3$). Hmong-Mien-speaking Miao groups from Vietnam and Guangxi showed additional gene flow with Guangxi Sui_Huanjiang ($-17.884 \leq Z\text{-scores} \leq -17.838$). Hmong shared more alleles with Sui_Duyun than with Maonan (Z-score = -3.044). Some of the Tai-Kadai speakers (such as Dai, Zhuang, Nung, Tay, and Lachi), Austroasiatic

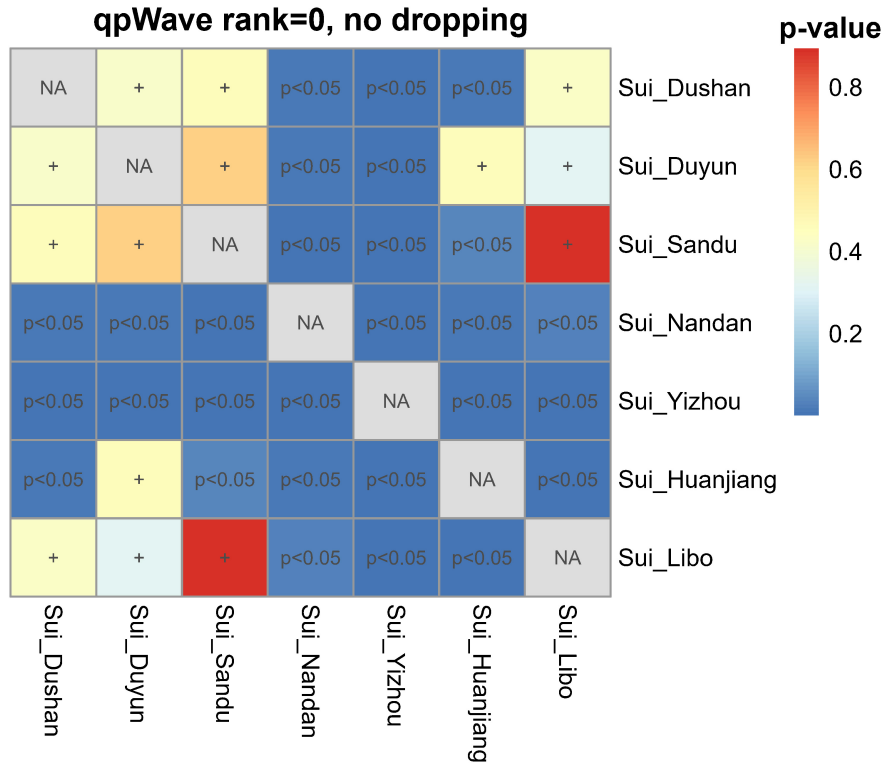

FIGURE 6 | Testing the genetic homogeneity of pairwise Sui populations. Heatmap showed *p*-values (rank = 0) of pairwise *qpWave* in Guizhou and Guangxi Sui populations. The *p*-value > 0.05 was noted as +, indicating this pair of studied Sui groups derived from a single homogeneous population relative to a set of outgroups.

speakers (such as Jing), and Austronesian speakers (such as Ede) shared more alleles with Maonan than with the Sui_Huanjiang/Sui_Yizhou/Sui_Duyun (*Z*-scores $\geq$ 3.027). When X = ancient EA and SEA individuals, we observed that ancient Fujian_EN (Liangdao2) individuals shared more genetic drift with Maonan than with each newly studied Guangxi Sui (i.e., Sui_Huanjiang, Sui_Yizhou, and Sui_Nandan, *Z*-scores $\geq$ 3.611).

We exhausted all possible pairs of reference populations as genetic sources to estimate admixture signals in each studied Sui population *via admixture-f_3-statistics*. The statistically significant negative f_3 values with *Z*-scores less than −3 suggested that the target population might be an admixture of two source-related populations. Here, we reported the top 10 lowest f_3 value results for each Sui group in the form of f_3*(modern source1, modern source2; Studied Sui)* and f_3*(ancient source1, ancient source2; Studied Sui)*, respectively, in **Supplementary Tables 3, 4**.

When focused on f_3*(modern source1, modern source2; Studied Sui)*, only three significant admixture signals were observed in Sui_Duyun when we used Hmong-Mien ancestry (Hmong) as one source and TK-related (Li/Mulam) or AA-related ancestry (Muong) as the other source (*Z*-scores $\leq$ −3.093). Although no more significant negative f_3 values were identified when we used other Sui groups as targets, it is important to note that a nonnegative *admixture-f_3* value does *not* prove that there is no admixture. The lowest *admixture-f_3* values were achieved in each studied Sui group except Yizhou Sui when we used the same pairs of source populations as follows: (1) Hmong-Mien-related Miao_Huanjiang/Hmong as *source1* and Tai-Kadai related Li/Mulam/Gelao as *source2*; and (2) Hmong-Mien-related Miao_Huanjiang/Hmong as *source1* and AN related Ami/Atayal as *source2*. For each newly studied Guizhou Sui group (i.e., Sui_Dushan, Sui_Duyun, Sui_Sandu, and Sui_Libo), when we used Hmong-Mien-related Miao_Huanjiang/Hmong as *source1* and AA-related Muong as *source 2*, low f_3 values can be produced. For Sui_Yizhou, when one of the source populations represented Sino-Tibetan-related ancestry (Sherpa/Tibetan_Shigatse) and the other represented TK-related (Li) or AN-related ancestry (Atayal), top negative f_3 values were generated.

When focused on the form f_3*(ancient ref1, ancient ref2; Sui)*, the negative f_3 values were observed in all Sui groups when one source was from ancient Southeast Asia (such as inland Indonesia_LN_BA_IA, Vietnam_LN_HaLongCulture, and coastal Liangdao) and the other from ancient Northern East Asia (such as Miaozigou_MN and AR_IA), suggesting the north–south admixture pattern for Sui people.

We further applied *qpAdm* to explore the plausible admixture models for our studied Sui populations. We used the Late Neolithic Yellow River Basin farmer-related population (YR_LN) and Early Neolithic Coastal Liangdao2 individuals as proxies for the Northern East Asian-related and Southeast Asian-related source populations in a two-way admixture model (**Figure 8**). We observed that our newly studied Sui individuals were estimated

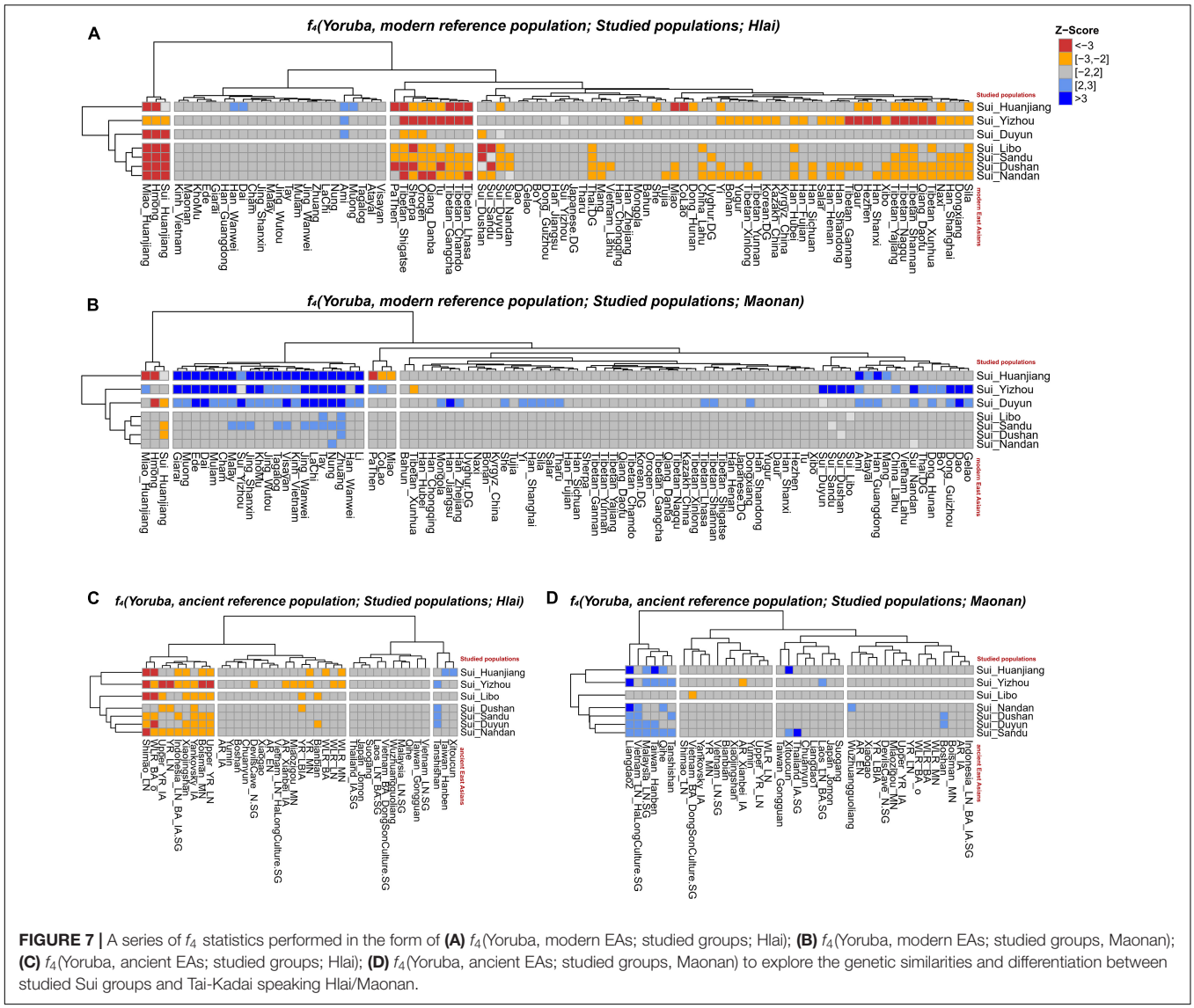

FIGURE 7 | A series of f_4 statistics performed in the form of **(A)** f_4(Yoruba, modern EAs; studied groups; Hlai); **(B)** f_4(Yoruba, modern EAs; studied groups, Maonan); **(C)** f_4(Yoruba, ancient EAs; studied groups; Hlai); **(D)** f_4(Yoruba, ancient EAs; studied groups, Maonan) to explore the genetic similarities and differentiation between studied Sui groups and Tai-Kadai speaking Hlai/Maonan.

to have 45.3–63.8% ancient Fujian Liangdao2 hunter-gatherer-related ancestry and 36.2–54.7% Yellow River farmer-related ancestry. The proportions of Liangdao2-related ancestry in our newly reported Sui_Huanjiang and Sui_Sandu individuals were at the same level, 54.6% (std.error = 7.7%) and 55.1% (std.error = 7.1%), respectively. Tai-Kadai-speaking Dai (69.2%, std.error = 8.2%) and Li (65.1%, std.error = 8.5%) and newly reported Sui_Dushan (63.8%, std.error = 8.3%), Sui_Duyun (59.3%, std.error = 8%), Sui_Libo (59.6%, p-value < 0.05, std.error = 9.4%), and Sui_Nandan (60.5%, std.error = 7.1%) samples harbored a similar proportion of Liangdao2-related ancestry. The Yizhou Sui had the highest proportion of YR_LN-related ancestry (54.7%, std.error = 7.1%) among Sui groups.

As the results of *admixture-f_3-statistics* suggested, our studied Sui might be modeled as an admixture of one ancient inland Southeast Asian group and one ancient North East Asian group. Therefore, we used coastal Southeast Asian (represented by Ami, the indigenous AN-speaking Taiwanese), inland Southeast Asian (represented by Vietnam_N), and Northern East Asian (represented by YR_LN) as three proxies of the possible ancestral sources to infer ancestry mixture coefficients in a three-way admixture model. The studied Sui groups covered similar proportions of coastal SEA ancestry and inland SEA ancestry and clustered with neighboring HM-speaking populations and TK-speaking populations in the ternary diagram (**Figure 9**).

Detecting the Positive Natural Selection Signals of Sui People

We applied haplotype-based iHS and nSL statistical indexes to explore the putative positive selection signals in Sui people (**Figure 10**). We got 317,569 phased SNPs of 58 Sui individuals. A total of 1,829 candidate SNPs with (1) normalized iHS score > top 1% score and (2) normalized nSL score > top 1% score were then used for KEGG pathway analysis. We listed the KEGG enrichment results with a p-value < 0.05 and

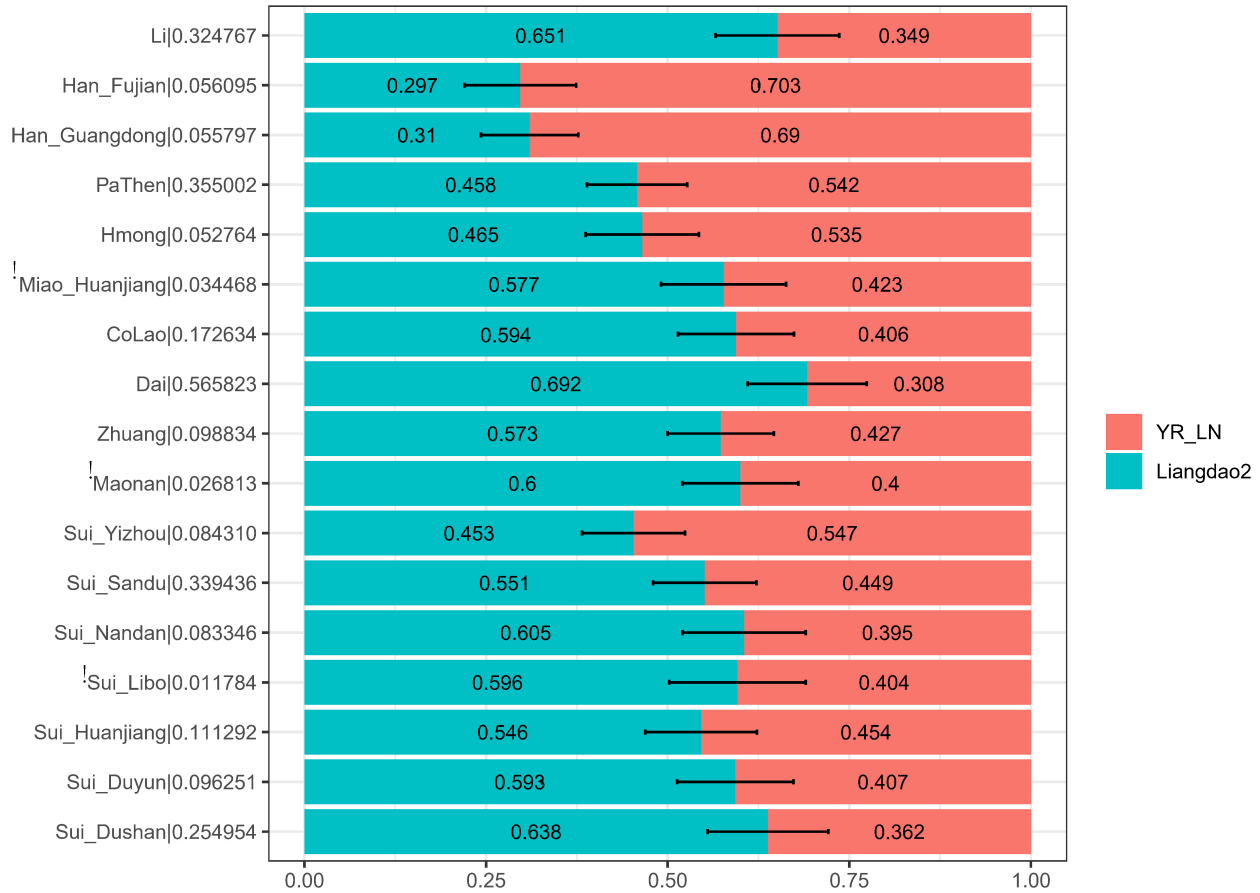

FIGURE 8 | Ancestry mixture coefficients in the two-way admixture model *via qpAdm*. We used Liangdao2, YR_LN as proxies for ancient Southerneast Asian, ancient North East Asian, respectively. Mbuti, Tianyuan, Papuan, Onge, DevilsCave_N, Japan_Jomon, Mongolia_N_East were used as outgroups. For each target population, we reported the *p*-value for rank = 1 behind the corresponding population name in the form of "target name| *p*-value." The population failed in *qpAdm* modeling (i.e., *p*-value < 0.05 for rank = 1) was noted as "!". Error bar denoted the standard error estimated using jackknife.

corrected *p*-value < 0.05 in **Supplementary Figure 5**. Candidate loci were mainly enriched in the pathway with regard to the susceptibility of complex diseases, cellular processes, and so on. Specifically, the SNP which had the highest iHS and nSL scores was SNP rs11599686 (chr10:123863188), located in the *TACC2* gene associated with the susceptibility of complex diseases such as breast cancer (Onodera et al., 2016).

DISCUSSION

The strong correlation between the population structure and linguistic classifications/geographic locations in East Asia has been reported in several genome-wide SNP-based studies (He et al., 2020; Huang et al., 2020; Kutanan et al., 2021; Wang C. C. et al., 2021). The population expansion with the extensive gene flow among populations which belong to the different linguistic classifications also drives the formation of the complex population genetic structure in East Asia (Huang et al., 2020; Liu D. et al., 2020; Kutanan et al., 2021; Wang M. et al., 2021; Wang C. C. et al., 2021; Yang et al., 2020). Yang et al. (2020) recently reconstructed the genetic structure and admixture history of Neolithic ancient NEAs and SEAs, demonstrating that the population structure in East Asia had existed early in the Neolithic; the spread of the NEA-related ancestry led to more genetic homogeneity in present-day EAs than in Neolithic EAs. Wang C. C. et al. (2021) demonstrated that the population expansion of the hunter-gatherers in Mongolia and Amur River Basin, Yellow River Basin farmers, Yangtze River Basin farmers, and Yamnaya Steppe pastoralists during the Holocene contributed to the formation of the population genetic structure in East Asia. Liu D. et al. (2020) analyzed the allele sharing and haplotype sharing profiles in present-day populations from five major language families in Vietnam and found extensive genetic interactions between Hmong-Mien-speaking populations and Tai-Kadai-speaking populations, such as the Tai-Kadai-speaking CoLao, who harbored more Hmong-Mien-related ancestry (represented by Hmong) compared with neighboring Tai-Kadai speakers.

The genetic profile of Tai-Kadai-speaking Sui people and the admixture history and genetic affinity with neighboring populations were largely unknown due to a

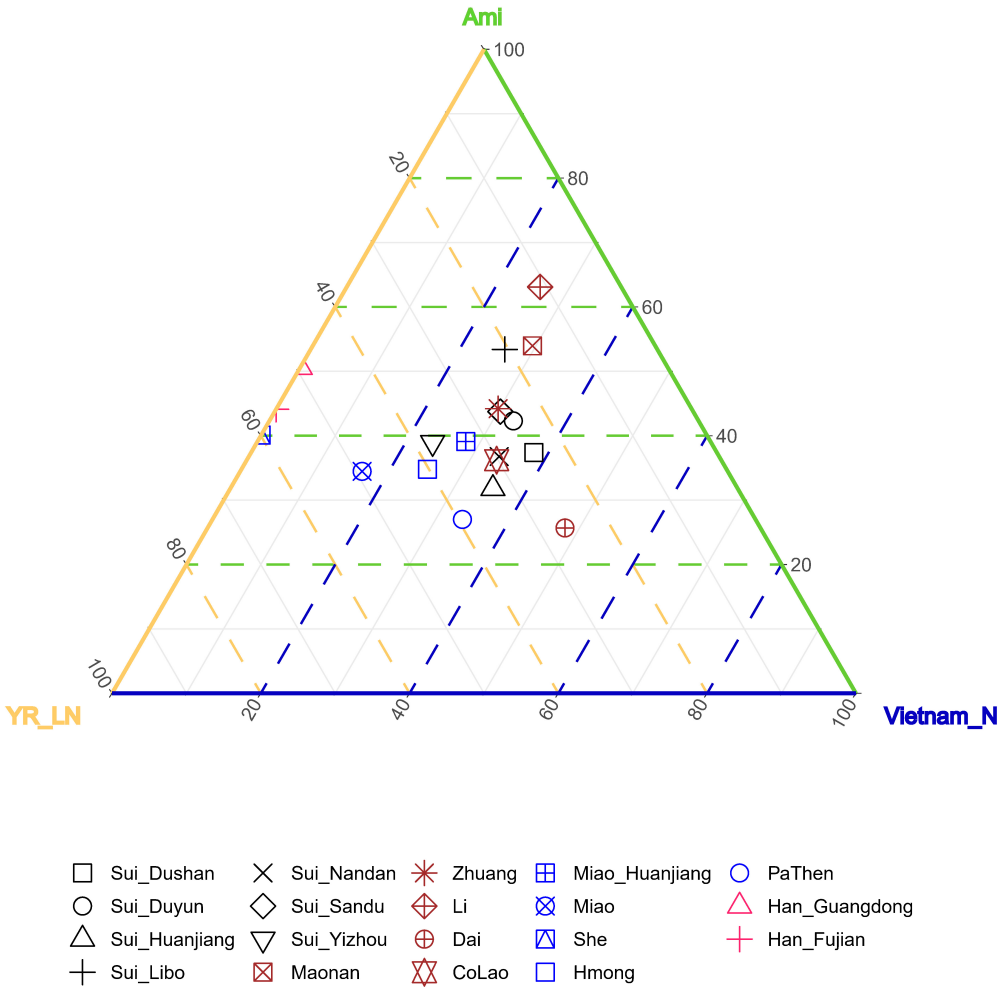

FIGURE 9 | Ancestry mixture coefficients in the three-way admixture model *via qpAdm*. We used Inland Southeast Asian (represented by Vietnam_N), Coastal Southeast Asian (represented by Ami) and North East Asian (represented by YR_LN) as three proxies of the possible ancestral sources of studied Sui. Mbuti, Tianyuan, Papuan, Onge, Liangdao2, DevilsCave_N, Japan_Jomon, Mongolia_N_East, Malaysia_LN were used as outgroups. The Nested p-value > 0.05 when Han_Fujian, Han_Guangdong and She were used as target, suggesting it may be appropriate to drop Vietnam_N source from the model, so we finally showed the admixture proportions in two-way admixture model for these three groups in this figure.

lack of high-density sampling and genome-wide data. In this study, we comprehensively co-analyzed our newly genotyped genome-wide SNP data of 24 Guangxi Sui and 34 Guizhou Sui individuals with published ancient and present-day East Asians to elucidate (1) the genetic relationships between the Sui and reference East Asians; (2) the fine-scale population structure within the Sui people; (3) the admixture history of each Sui population; and (4) the potential positive selection signals of Sui people. The patterns of shared genetic drift measured *via outgroup-f_3-statistics* and the phylogenetic relationships in Fst-based N-J tree supported the finding that the studied Sui had closer genetic relationships with Neolithic-to-modern populations from Southern China and Southeast Asians, especially present-day TK, HM, Sinitic, AA, and AN speakers and ancient Coastal Southeast Asians which were represented by Iron Age Taiwan Hanben and Gongguan individuals compared with most NEAs, supporting the hypotheses from the genomic perspective that Sui people were originally from southern China.

These results were in accordance with the genetic findings based on forensic-related genetic markers (X-STR, Y-STR, and autosomal-STR) indicating that the Sui people displayed a close affinity with geographically close populations, especially Qiandongnan Miao (Chen et al., 2018), AA-speaking Jing (Yang et al., 2012), and Maonan and Guizhou Han (Guo et al., 2019).

Results of PCA and model-based ADMIXTURE analysis revealed the genetic substructure within seven studied Sui populations: (1) the TK-speaking Sui individuals from Guangxi Huanjiang formed a cline which was localized on the intermediate of Hmong-Mien-related genetic cline, rather than clustered with neighboring TK-speaking populations, and (2) a relatively loose cluster which consisted of the rest of the newly reported Sui populations, partially overlapped with TK-speaking populations from Southern China and Southeast Asia.

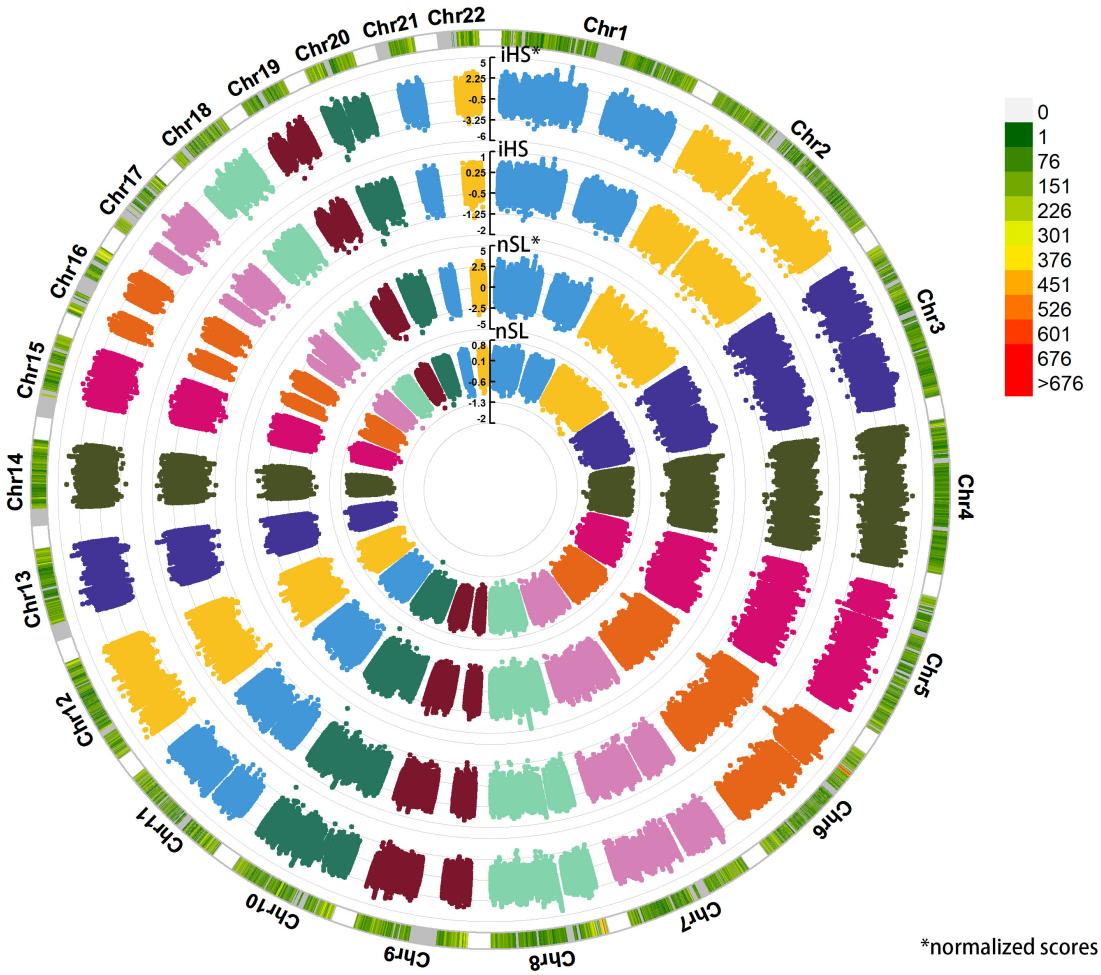

FIGURE 10 | Detecting the positive selection signals. Here, we displayed the raw iHS scores, standardized iHS scores, raw nSL scores, standardized nSL scores *via* circle Manhattan plot.

Symmetric f_4-*statistics* and pairwise *qpWave* consistently showed that the Guizhou Sui people were relatively homogeneous and showed similar genetic profiles with Tai-Kadai-related populations, such as Maonan. While Guangxi Sui groups were relatively heterogeneous, we observed excessive genetic affinity between Tai-Kadai-speaking Huanjiang Sui and the geographically close Hmong-Mien-speaking Miao people although the two groups were ethnically different. Huanjiang Sui did not share the most derived alleles with other Guizhou and Guangxi Sui people, which belonged to the same ethnic group, suggesting that Huanjiang Sui received a significantly geographically close Hmong-Mien-speaking Miao-related ancestry (represented by Huanjiang Miao) after Huanjiang Sui and other Sui people separated from the common ancestor. Sino-Tibetan-related populations contributed to the extra ancestry to Yizhou Sui people compared with other Sui groups, indicating that the Yizhou Sui have been primarily affected genetically by the surrounding Han populations. These results suggested the differentiated demographic history among the studied Sui populations.

The significant negative *Z*-scores of f_4*(Yoruba, East Asians; studied Sui, TK-speaking Li)* revealed that the ancestors of the Sui people might experience excessive admixture events with HM and NAEs after the divergence with the unmixed, indigenous TK-speaking proxy Li islanders. The significant negative values in *admixture-*f_3*(Hmong, Li/Mulam; studied Sui_Duyun)* suggested that the ancestor of HM-related populations might also participate in the formation of the Sui people. Hmong-Mien-speaking populations, such as Miao and Yao, are the dominant ethnic groups in southwest China. While the Sui has a relatively small population, there is only one autonomous country for Sui people in China (i.e., Sandu Autonomous Country of Guizhou). Previous studies suggested that Hmong-Mien-related people might be the direct descendants of Daxi-related people as there was a rare Y-chromosome haplogroup O3d detected in Neolithic Yangtze river Daxi culture-related people and modern Hmong-Mien speakers (Su et al., 1999; Li et al., 2007b). A possible scenario is that Hmong-Mien-speaking populations carried more Neolithic Yangtze River farmer-related ancestry and had a distinct genetic profile when

compared with Sui people; the ancestor of the Sui people migrated southward (according to the historical documents) admixed with the indigenous Hmong-Mien-related populations, transforming the genetic makeup of Sui populations. More Yangtze River-related ancient samples in Neolithic, Bronze Age, Iron Age, and Historical Age are expected to be studied. The weak admixture signals in *admixture-f_3 statistics* shed light on the potential north–south admixture patterns for the studied Sui groups. Furthermore, the *qpAdm*-based admixture model demonstrated that Sui people could be modeled as the admixture of ancient Northern East Asians (ANEAs) and ancient southeast Asians (ASEAs), in which ANEAs were represented by Neolithic Yellow River Basin-related farmer populations and ASEAs were characterized by Neolithic/modern Coastal Southern East Asians. More specifically, in the three-way admixture model, Sui people and neighboring HM- and TK-speaking populations derived their ancestry from more similar (but still different) proportions of ANEA-, coastal SEA-, and inland SEA-related components compared with the Han from Southern China. Conclusively, the formation of the population structure of Tai-Kadai-speaking populations might be plausibly explained by (1) the differential proportions of the primary ancient sources and (2) the various levels of gene flow with surrounding people, such as HM-speaking and Sinitic-speaking populations.

Finally, we detected the potential positive selection signals based on the phased data of Sui people *via* normalized iHS and nSL. Notably, lots of significant SNPs enriched in the KEGG pathway associated with the susceptibility of complex diseases (such as the Type 1 diabetes mellitus pathway and gastric cancer pathway). When focused on the SNPs which were famous for the natural selection in East Asians in previous genetic studies (Sabeti et al., 2007; Peng et al., 2010; Calandra et al., 2011; Kamberov et al., 2013), we observed SNPs rs3827760 (chr2:109513600) on the *EDAR* gene (associated with hair thickness and facial morphology), rs4148211 (chr2:44071742) on the *ABCG8* gene (associated with lipid metabolism), and rs1229984 (chr4:100239318) located in the *ADH1B* gene (associated with alcohol metabolism), which were likely under natural selection in Sui people (rs3827760: normalized iHS score = 2.788630; normalized nSL score = no results; rs4148211: normalized iHS score = 2.536000; normalized nSL score = 1.980440; rs1229984: normalized iHS score = 2.096340; normalized nSL score = 2.303840).

However, more population genetic studies based on the whole-genome sequence (WGS) data *via* next-generation sequencing (NGS) and single-molecule real-time sequencing (SMRTS) of Sui individuals from more geographic locations (such as Yunnan, Sichuan, and Jiangsu) are expected to be conducted to provide genomic insight into the formation of Sui people and to dissect the complex demographic history of populations from Southern China and Southeast Asia comprehensively.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://zenodo.org/record/5483577, doi: 10.5281/zenodo.5483577.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Medical Ethics Committee of Youjiang Medical University for Nationalities and Xiamen University (Approval Number: XDYX2019009). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

C-CW and XH designed the study. RuW and C-CW wrote the manuscript. XB, YH, RoW, and XH collected the samples. KZ, XY, HM, GH, JG, JZ, MY, JC, XZ, LT, and YL conducted the experiment. RuW and C-CW analyzed the data. All authors reviewed the manuscript.

FUNDING

This work was funded by National Natural Science Foundation of China (NSFC 32060208, 31801040), Nanqiang Outstanding Young Talents Program of Xiamen University (X2123302), the Major project of National Social Science Foundation of China (20&ZD248), a European Research Council (ERC) grant to D. Xu (ERC-2019-ADG-883700-TRAM), and Fundamental Research Funds for the Central Universities (ZK1144).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.735084/full#supplementary-material

Supplementary Figure 1 | The Genetic Relationship Matrix (GRM). **(A)** based on the raw data (68 Sui individuals); **(B)** based on the clean data which removed the kinship (58 Sui individuals).

Supplementary Figure 2 | Cross-Validation error for model-based ADMIXTURE analysis. The lowest cross validation error occurred at $K = 4$.

Supplementary Figure 3 | Neighbor-joining tree based on Fst genetic distance among studied Sui and present-day EA and SEA individuals.

Supplementary Figure 4 | *qpWave* Outgroup dropping test, in which we dropped one of the populations in the outgroup set by turn (Mbuti, Mongolia_N_East, DevilsCave_N, Ami, Liangdao2, Vietnam_LN). If we observed $P > 0.05$ for rank = 0, while $P < 0.05$ (rank = 0) in the no-drop qpWave test, suggesting the dropped population might have a unique gene flow with one of the test groups, explained the non-homogeneity between the pairwise test populations.

Supplementary Figure 5 | KEGG pathway analysis. Each row represented an enriched function, and the length of the bar represented the enrich ratio which was calculated as "input gene number"/"background gene number." The color of the bar represents different clusters.

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

Bu, D., Luo, H., Huo, P., Wang, Z., Zhang, S., He, Z., et al. (2021). KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res.* 49, W317–W325. doi: 10.1093/nar/gkab447

Calandra, S., Tarugi, P., Speedy, H., Dean, A., Bertolini, S., and Shoulders, C. (2011). Mechanisms and genetic determinants regulating sterol absorption, circulating LDL levels, and sterol elimination: implications for classification and disease risk. *J. Lipid Res.* 52, 1885–1926. doi: 10.1194/jlr.r017855

Chen, P., He, G., Zou, X., Zhang, X., Li, J., Wang, Z., et al. (2018). Genetic diversities and phylogenetic analyses of three Chinese main ethnic groups in southwest China: a Y-Chromosomal STR study. *Sci. Rep.* 8:15339. doi: 10.1038/s41598-018-33751-x

Delaneau, O., Zagury, J., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* 10, 5–6. doi: 10.1038/nmeth.2307

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., and Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol. Biol. Evol.* 31, 1275–1291. doi: 10.1093/molbev/msu077

Gao, L. (2002). On the origin of the Shui ethnic group. *J. Guangzhou Univ. (Soc. Sci. Ed.)* 3, 11–15.

Guo, J., Ji, J., He, G., Ren, Z., Zhang, H., Wang, Q., et al. (2019). Genetic structure and forensic characterization of 19 X-chromosomal STR loci in Guizhou Sui population. *Ann. Hum. Biol.* 46, 246–253. doi: 10.1080/03014460.2019.1623911

He, G., Wang, Z., Guo, J., Wang, M., Zou, X., Tang, R., et al. (2020). Inferring the population history of Tai-Kadai-speaking people and southernmost Han Chinese on Hainan Island by genome-wide array genotyping. *Eur. J. Hum. Genet.* 28, 1111–1123. doi: 10.1038/s41431-020-0599-7

Huang, X., Xia, Z., Bin, X., He, G., Gui, J., Lin, C., et al. (2020). Genomic insights into the demographic history of Southern Chinese. *bioRxiv* [Preprint] doi: 10.1101/2020.11.08.373225

Huang, X., Zhou, Q., Bin, X., Lai, S., Lin, C., Hu, R., et al. (2018). The genetic assimilation in language borrowing inferred from Jing People. *Am. J. Phys. Anthropol.* 166, 638–648. doi: 10.1002/ajpa.23449

Ji, J., Ren, Z., Zhang, H., Wang, Q., Wang, J., Kong, Z., et al. (2017). Genetic profile of 23 Y chromosomal STR loci in Guizhou Shui population, southwest China. *Forensic Sci. Int. Genet.* 28, e16–e17. doi: 10.1016/j.fsigen.2017.01.010

Kamberov, Y., Wang, S., Tan, J., Gerbault, P., Wark, A., Tan, L., et al. (2013). Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* 152, 691–702. doi: 10.1016/j.cell.2013.01.016

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Kutanan, W., Kampuansai, J., Srikummool, M., Kangwanpong, D., Ghirotto, S., Brunelli, A., et al. (2017). Complete mitochondrial genomes of Thai and Lao populations indicate an ancient origin of Austroasiatic groups and demic diffusion in the spread of Tai-Kadai languages. *Hum. Genet.* 136, 85–98. doi: 10.1007/s00439-016-1742-y

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

Li, H., Cai, X., Winograd-Cort, E. R., Wen, B., Cheng, X., Qin, Z., et al. (2007a). Mitochondrial DNA diversity and population differentiation in southern East Asia. *Am. J. Phys. Anthropol.* 134, 481–488. doi: 10.1002/ajpa.20690

Li, H., Huang, Y., Mustavich, L., Zhang, F., Tan, J., Wang, L., et al. (2007b). Y chromosomes of prehistoric people along the Yangtze River. *Hum. Genet.* 122, 383–388. doi: 10.1007/s00439-007-0407-2

Li, H., Wen, B., Chen, S. J., Su, B., Pramoonjago, P., Liu, Y., et al. (2008). Paternal genetic affinity between western Austronesians and Daic populations. *BMC Evol. Biol.* 8:146. doi: 10.1186/1471-2148-8-146

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

Liu, C., Han, C., Min, Y., Liu, H., Xu, Q., and Yang, X. (2018). Genetic polymorphism analysis of 40 Y-chromosomal STR loci in seven populations from South China. *Forensic Sci. Int.* 291, 109–114. doi: 10.1016/j.forsciint.2018.08.003

Liu, D., Duong, N., Ton, N., Van, P., Pakendorf, B., Van, H. 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, J., Du, W., Wang, M., Liu, C., Wang, S., He, G., et al. (2020). Forensic features, genetic diversity and structure analysis of three Chinese populations using 47 autosomal InDels. *Forensic Sci. Int. Genet.* 45:102227. doi: 10.1016/j.fsigen.2019.102227

Liu, Y., Zhang, H., He, G., Ren, Z., Zhang, H., Wang, Q., et al. (2020). Forensic features and population genetic structure of dong, Yi, Han, and Chuanqing human populations in Southwest China inferred from insertion/deletion markers. *Front. Genet.* 11:360. doi: 10.3389/fgene.2020.00360

Lu, Y., Quan, C., Chen, H., Bo, X., and Zhang, C. (2016). 3DSNP: a database for linking human noncoding SNPs to their three-dimensional interacting genes. *Nucleic Acids Res.* 45, D643–D649. doi: 10.1093/nar/gkw1022

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

Onodera, Y., Takagi, K., Miki, Y., Takayama, K., Shibahara, Y., Watanabe, M., et al. (2016). TACC2 (transforming acidic coiled-coil protein 2) in breast carcinoma as a potent prognostic predictor associated with cell proliferation. *Cancer Med.* 5, 1973–1982. doi: 10.1002/cam4.736

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

Peng, Y., Shi, H., Qi, X., Xiao, C., Zhong, H., Ma, R., et al. (2010). The ADH1B Arg47His polymorphism in east Asian populations and expansion of rice domestication in history. *BMC Evol. Biol.* 20:15. doi: 10.1186/1471-2148-10-15

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

Sabeti, P., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913–918. doi: 10.1038/nature06250

Su, B., Xiao, J., Underhill, P., Deka, R., Zhang, W., Akey, J., et al. (1999). Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. *Am. J. Hum. Genet.* 65, 1718–1724. doi: 10.1086/302680

Szpiech, Z., and Hernandez, R. (2014). Selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* 31, 2824–2827. doi: 10.1093/molbev/msu211

Voight, B., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72. doi: 10.1371/journal.pbio.0040072

Wang, C. C., Yeh, H., Popov, A., Zhang, H., Matsumura, H., Sirak, K., et al. (2021). 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., and Tang, R. (2021). New insights from the combined discrimination of modern/ancient genome-wide shared alleles and haplotypes: differentiated demographic history reconstruction of Tai-Kadai and Sinitic people in South China. *bioRxiv* [Preprint] doi: 10.1101/2021.06.19.449013

Wang, T., Wang, W., Xie, G., Li, Z., Fan, X., and Yang, Q. (2021). Human population history at the crossroads of

East and Southeast Asia since 11,000years ago. *Cell* 184, 3829–3841.e21.

Weir, B., and Cockerham, C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358–1370. doi: 10.1111/j.1558-5646.1984.tb05657.x

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

Yang, J., Lee, S., Goddard, M., and Visscher, P. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011

Yang, L., Zhao, Y., Liu, C., Chan, D., Chan, M., and He, M. (2012). Allele frequencies of 15 STRs in five ethnic groups (Han, Gelao, Jing, Shui and Zhuang) in South China. *Forensic Sci. Int. Genet.* 7, e9–e14. doi: 10.1016/j.fsigen.2012.10.009

Yang, M., Fan, X., Sun, B., Chen, C., Lang, J., Ko, Y., et al. (2020). Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* 17, 282–288. doi: 10.1126/science.aba0909

Zhang, X., and Zhang, J. (2018). A new study on the origin of Shui nationality. *J. Guangxi Normal Univ. Natl.* 35, 44–48.

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.

The reviewer SW declared a past co-authorship with the authors XY, GH, JG, JZ, and C-CW to the handling editor.

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 11 October 2021
doi: 10.3389/fgene.2021.756802

Population Genetic Polymorphism of Skeletal Muscle Strength Related Genes in Five Ethnic Minorities in North China

Bonan Dong[1,2†], *Qiuyan Li*[1,2,3†], *Tingting Zhang*[1,2], *Xiao Liang*[1,2], *Mansha Jia*[4], *Yansong Fu*[1,2], *Jing Bai*[1,2] *and Songbin Fu*[1,2*]

[1]*Laboratory of Medical Genetics, Harbin Medical University, Harbin, China,* [2]*Key Laboratory of Preservation of Human Genetic Resources and Disease Control in China (Harbin Medical University), Ministry of Education, Harbin, China,* [3]*Editorial Department of International Journal of Genetics, Harbin Medical University, Harbin, China,* [4]*Scientific Research Centre, The Second Affiliated Hospital of Harbin Medical University, Harbin, China*

OPEN ACCESS

Edited by:
Chuan-Chao Wang, Xiamen University, China

Reviewed by:
Seyed Mehdi Talebi, Arak University, Iran
Omid Jafari, Gorgan University of Agricultural Sciences and Natural Resources, Iran
Ayman Hassan Abd El-Aziz, Damanhour University, Egypt

***Correspondence:**
Songbin Fu
fusb@ems.hrbmu.edu.cn

†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: *11 August 2021*
Accepted: *21 September 2021*
Published: *11 October 2021*

Citation:
Dong B, Li Q, Zhang T, Liang X, Jia M, Fu Y, Bai J and Fu S (2021) Population Genetic Polymorphism of Skeletal Muscle Strength Related Genes in Five Ethnic Minorities in North China. Front. Genet. 12:756802. doi: 10.3389/fgene.2021.756802

Musculoskeletal performance is a complex trait influenced by environmental and genetic factors, and it has different manifestations in different populations. Heilongjiang province, located in northern China, is a multi-ethnic region with human cultures dating back to the Paleolithic Age. The Daur, Hezhen, Ewenki, Mongolian and Manchu ethnic groups in Heilongjiang province may have strong physical fitness to a certain extent. Based on the genetic characteristics of significant correlation between some important genes and skeletal muscle function, this study selected 23 SNPs of skeletal muscle strength-related genes and analyzed the distribution of these loci and genetic diversity in the five ethnic groups. Use Haploview (version 4.1) software to calculate the chi-square and the Hardy-Weinberg equilibrium to assess the difference between the two ethnic groups. Use R (version 4.0.2) software to perform principal component analysis of different ethnic groups. Use MEGA (version 7.0) software to construct the phylogenetic tree of different ethnic groups. Use POPGENE (version 1.32) software to calculate the heterozygosity and the F_{ST} values of 23 SNPs. Use Arlequin (version 3.5.2.2) software to analyze molecular variance (AMOVA) among 31 populations. The results showed that there was haplotype diversity of *VDR*, *angiotensin-converting enzyme*, *ACTN3*, *EPO* and *IGF1* genes in the five ethnic groups, and there were genetic differences in the distribution of these genes in the five ethnic groups. Among them, the average gene heterozygosity (AVE_HET) of the 23 SNPs in the five populations was 0.398. The F_{ST} values of the 23 SNPs among the five ethnic groups varied from 0.0011 to 0.0137. According to the principal component analysis, the genetic distance of Daur, Mongolian and Ewenki is relatively close. According to the phylogenetic tree, the five ethnic groups are clustered together with the Asian population. These data will enrich existing genetic information of ethnic minorities.

Keywords: skeletal muscle strength-related genes, SNP, ethnic groups, phylogenetic relationship, population genetics

INTRODUCTION

Skeletal muscle is one of the most dynamic and plastic tissues of the human body, and it is an important part of the human body. The skeletal muscles are involved in various functions of human life. From a mechanical perspective, the main function of skeletal muscle is to convert the body's chemical energy into mechanical energy, so that the body can generate force and strength, and then generate movement to maintain or benefit human health. From a metabolic perspective, the roles of skeletal muscle include promoting basal energy metabolism, storing important substrates such as amino acids and carbohydrates, and providing most of the oxygen and energy for human movement (Frontera and Ochala, 2015).

With the development of exercise physiology, studies have found that acquired physical training has an important and positive effect on the improvement of human muscle mass, strength and function (Phu et al., 2015). In addition, genetic differences can influence the ability of the body's skeletal muscles to produce and use energy during exercise (Yan et al., 2016). Studies have highlighted a significant correlation between potentially important genes and musculoskeletal function. For example, the *VDR* gene may have a positive effect on skeletal muscle (Książek et al., 2019), and the *IGF1* gene increases muscle mass and improves skeletal muscle regeneration (Vassilakos and Barton, 2018), and the *EPO* gene promotes differentiation and survival of myoblasts (Lamon and Russell, 2013). In addition, other genes involved in skeletal muscle strength include the endurance gene *ACE* (Ahmetov and Fedotovskaya, 2015), and the strength-related genes, such as *ACTN3* (Ahmetov and Fedotovskaya, 2015; Seto et al., 2021), *AGT* (Pickering et al., 2019), *PPARG* (Ahmetov and Fedotovskaya, 2015; Norouzi et al., 2019) and *IL6* (Pickering et al., 2019). Due to environmental and genetic factors, there are different manifestations in different ethnic groups (Pitsiladis et al., 2016). For instance, the frequencies of the three *ACE* genotypes (II, ID, DD) were 25, 50, and 25%, respectively, in Caucasian populations (Jones et al., 2002), which were not significantly different from those of Asian populations in Korea (23, 66, and 11%, respectively) (Oh, 2007). Other studies have found that the ID genotype is significantly associated with outstanding endurance quality in both European and African American populations (Weyerstraß et al., 2018). The A allele of rs699 locus of *AGT* gene was significantly correlated with Brazilian endurance quality (Guilherme et al., 2018). CT genotype of *ACTN3* gene was markedly correlated with explosive power of Caucasian. CC genotype was substantially correlated with Asian explosive power. The T allele or TT genotype was significantly correlated with the explosive power of both Caucasian and Asian male populations, and the TT genotype also significantly affected the explosive power of Russian athletes (Weyerstraß et al., 2018). CC genotype of *AGT* gene has a high performance in Polish power athletes, with a genotype frequency of 40% (Zarębska et al., 2013). The C allele of *IL6* was positively associated with athletic ability in Israelis of Ethiopian descent, which not only improved speed but also improved training recovery (Ben-Zaken et al., 2021). China is a multi-ethnic country, consisting of the Han nationality and 55 ethnic minorities, of which the population of 55 ethnic minorities accounts for about 8% of the total population. To a certain extent, it provides abundant genetic resources for the study of genes related to skeletal muscle strength. Heilongjiang province, located in northern China, is a multi-ethnic region with human culture since the Paleolithic Age. To some extent, the Daur, Mongolian, Ewenki, Manchu and Hezhen belong to the Altaic language family in Heilongjiang may have stronger physical fitness. According to reports, the grip strength of Mongolian, Daur and Ewenki adults is significantly higher than the national level (Dong et al., 2004). In addition, some scholars believed that some indexes of physical characteristics in Hezhen people are slightly higher than those of Han people due to engaged in fishing and hunting activities for a long time (Chen et al., 1999; Wang et al., 2014). Some scholars sorted out and counted the relevant materials of 263 Manchu college students aged 19 to 22, and found that the physical fitness of Manchu college students was significantly better than that of Han (Bi, 1993).

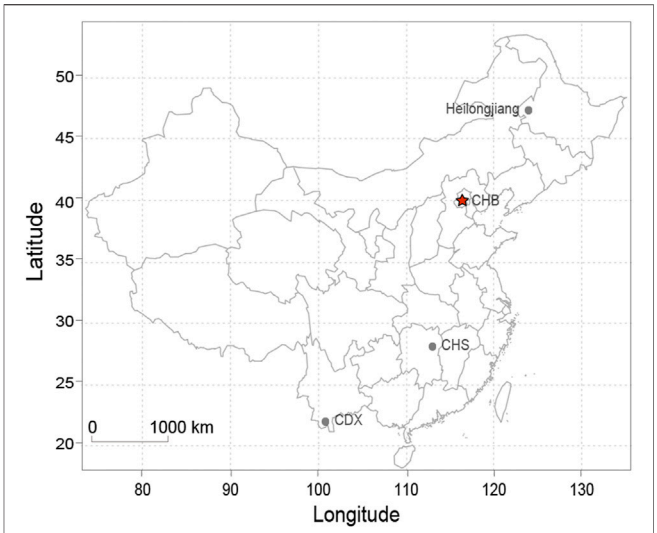

FIGURE 1 | The geographical distribution of eight ethnic groups in China. Note: CHB: Han Chinese in Beijing; CHS: Southern Han Chinese; CDX: Chinese Dai in Xishuangbanna.

Single nucleotide polymorphisms (SNPs) refer to DNA sequence polymorphisms caused by a single nucleotide variation at the genome level, with a frequency generally greater than 1% in the population. SNP is closely related to the genetic traits of populations and can be used as genetic markers for the genetic structure of different populations (Galinsky et al., 2019). Based on the genetic characteristics of significant correlation between some important genes and skeletal muscle function, this study intends to select 23 SNPs in *AGT* (rs699, rs4762, rs5051, rs5050), *PPARG* rs3856806, *IL6* rs2066992, *ACE* (rs4309, rs4331, rs4341, rs4343, rs4362), *ACTN3* (rs1815739, rs540874), *EPO* (rs1617640, rs551238), *IGF1* (rs5742714, rs1520220, rs5742612, rs972936), *VDR* (rs7975232, rs757343, rs2228570, rs11568820) genes. We analyzed the allele frequency of these loci in Daur, Hezhen, Ewenki, Mongolian and Manchu, and compared with the 26 populations from 1,000

TABLE 1 | The genotype distribution and Hardy-Weinberg equilibrium test for the 23 SNPs in five ethnic populations from China.

Gene	Loci	A/B	AA[a]	AB[a]	BB[a]	HWPval
AGT	rs699	G/A	567	271	36	>0.05
	rs4762	G/A	749	113	2	>0.05
	rs5051	T/C	552	265	39	>0.05
	rs5050	T/G	605	229	30	>0.05
PPARG	rs3856806	C/T	590	262	21	>0.05
IL6	rs2066992	T/G	396	378	108	>0.05
EPO	rs1617640	A/C	449	359	72	>0.05
	rs551238	T/G	431	374	77	>0.05
ACTN3	rs1815739	C/T	460	253	167	>0.05
	rs540874	G/A	442	233	169	>0.05
IGF1	rs5742714	C/G	651	204	21	>0.05
	rs1520220	C/G	415	303	163	>0.05
	rs5742612	A/G	477	345	60	>0.05
	rs972936	C/T	416	332	133	>0.05
VDR	rs7975232	C/A	481	345	56	>0.05
	rs757343	C/T	546	240	24	>0.05
	rs2228570	G/A	393	311	164	>0.05
	rs11568820	C/T	410	311	143	>0.05
ACE	rs4309	T/C	402	346	132	>0.05
	rs4331	G/A	403	356	123	>0.05
	rs4341	C/G	402	355	124	>0.05
	rs4343	A/G	404	355	122	>0.05
	rs4362	C/T	422	325	132	>0.05

[a]*AA wild homozygote, AB heterozygote, BB mutant homozygote.*

TABLE 2 | The minimum allele frequencies of 23 SNPs in five populations.

Loci	Daur	Mongolian	Ewenki	Manchu	Hezhen
rs699	0.238	0.197	0.205	0.147	0.197
rs4762	0.056	0.069	0.041	0.06	0.093
rs5051	0.239	0.208	0.225	0.156	0.192
rs5050	0.144	0.154	0.16	0.151	0.211
rs3856806	0.178	0.199	0.178	0.172	0.161
rs2066992	0.33	0.429	0.425	0.311	0.3
rs1617640	0.282	0.324	0.342	0.282	0.26
rs551238	0.307	0.321	0.336	0.293	0.278
rs1815739	0.429	0.443	0.452	0.45	0.476
rs540874	0.425	0.441	0.459	0.5	0.478
rs5742714	0.139	0.105	0.116	0.166	0.141
rs1520220	0.367	0.376	0.404	0.425	0.358
rs5742612	0.238	0.259	0.281	0.284	0.266
rs972936	0.367	0.381	0.404	0.436	0.36
rs7975232	0.236	0.33	0.336	0.277	0.212
rs757343	0.133	0.2	0.229	0.215	0.169
rs2228570	0.337	0.44	0.4178	0.435	0.413
rs11568820	0.425	0.368	0.37	0.481	0.338
rs4309	0.38	0.365	0.37	0.336	0.422
rs4331	0.382	0.349	0.384	0.307	0.412
rs4341	0.382	0.348	0.384	0.311	0.412
rs4343	0.382	0.348	0.384	0.311	0.408
rs4362	0.393	0.365	0.39	0.35	0.434

genome project, to investigate the genetic polymorphism of skeletal muscle strength related genes in the five ethnic groups and to provide theoretical support for explaining the genetic polymorphism of skeletal muscle strength related genes between different populations.

MATERIALS AND METHODS

Study Populations

Blood samples were collected from 882 unrelated individuals (413 males, 469 females, 45 average age) belonging to five Chinese ethnic minorities in Heilongjiang province at least three generations. These individuals include 233 Daur individuals, 106 Mongolian individuals, 73 Ewenki individuals, 220 Manchu individuals, and 250 Hezhen individuals. The geographical distribution on the map is shown in **Figure 1**. The study was carried out in strict accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Harbin Medical University. All the participants signed a written informed consent form.

DNA Extraction and Genotyping

Genomic DNA was extracted from 200 μl blood using the QIAamp DNA Blood Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Genotyping was performed using the SNPscan™ Kit (Genesky Biotechnologies Inc., Shanghai, China) according to the manufacturer's instructions.

Database Data

The genotype and allele frequency data of individuals from the 26 populations in the world were downloaded from the ensemble database at http://grch37.ensembl.org/Homo_sapiens/Tools/DataSlicer The abbreviations and full names of the 26 populations in the world were downloaded from the https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes.

Statistical Analysis

Chi-square and Hardy-Weinberg equilibrium were calculated to assess the differences between two populations using the Haploview software, the linkage disequilibrium and the haplotype analysis of SNPs also were performed by it (Barrett et al., 2005). In the haplotype analysis the r^2 threshold was 0.8. Phylogenetic tree was generated using the UPGMA dendrogram method in MEGA7 (Kumar et al., 2016). The parameter such as AVE_HET, FST, Nm and the Nei's genetic distance based on UPGMA of the five ethnic groups were calculated using the POPGENE software (Yeh et al., 1997). Principal component analysis (PCA) were carried out in the R packages "factoextra" and "ggplot2" (Luu et al., 2017; Singh and Soman, 2019). Analysis of molecular variance (AMOVA) was calculated by Arlequin (Excoffier et al., 2007).

RESULTS

Genotyping Data and Hardy-Weinberg Test

The genotype distribution in the study is summarized in **Table 1**. The 23 SNPs included in the study were all in line with Hardy-Weinberg equilibrium ($p > 0.05$). The minimum allele frequencies and genotype frequencies of 23 SNPs in five populations are summarized in **Table 2** and **Supplementary Table S1** respectively.

TABLE 3 | Summary statistical different SNPs after Pairwise comparison of five populations.

Populations	Gene	Loci	Assoc allele	Chi square	*p* Value
Daur vs Ewenki	*IL6*	rs2066992	G	4.318	0.0377
	VDR	rs7975232	A	5.731	0.0167
		rs757343	T	7.547	0.006
Daur vs Hezhen	*VDR*	rs2228570	A	5.91	0.0151
		rs11568820	C	7.601	0.0058
	AGT	rs4762	A	4.737	0.0295
		rs5050	G	7.459	0.0063
Daur vs Manchu	*ACE*	rs4343	A	4.975	0.0257
	VDR	rs757343	T	9.932	0.0016
		rs2228570	A	9.076	0.0026
	ACE	rs4331	G	5.653	0.0174
		rs4341	C	4.975	0.0257
	IGF1	rs972936	T	4.541	0.0331
	AGT	rs699	G	12.014	0.0005
		rs5051	T	9.64	0.0019
Daur vs Mongolian	*IL6*	rs2066992	G	6.16	0.0131
	VDR	rs7975232	A	6.622	0.0101
		rs757343	T	4.569	0.0326
		rs2228570	A	6.396	0.0114
Ewenki vs Hezhen	*IL6*	rs2066992	T	7.965	0.0048
	VDR	rs7975232	C	9.469	0.0021
	AGT	rs4762	A	4.042	0.0444
Ewenki vs Manchu	*IL6*	rs2066992	T	6.274	0.0123
	VDR	rs11568820	T	5.256	0.0219
Manchu vs Hezhen	*ACE*	rs4343	G	9.455	0.0021
	VDR	rs7975232	C	5.427	0.0198
		rs11568820	C	19.535	9.88E-06
	ACE	rs4309	C	7.271	0.007
		rs4331	A	11.201	0.0008
		rs4341	G	10.228	0.0014
		rs4362	T	6.857	0.0088
	IGF1	rs1520220	C	4.419	0.0355
		rs972936	C	5.707	0.0169
	AGT	rs699	A	4.068	0.0437
		rs5050	G	5.551	0.0185
Mongolian vs Hezhen	*IL6*	rs2066992	T	11.107	0.0009
	VDR	rs7975232	C	11.175	0.0008
Mongolian vs Manchu	*IL6*	rs2066992	T	8.742	0.0031
	VDR	rs11568820	T	7.228	0.0072
	IGF1	rs5742714	G	4.243	0.0394

The Frequencies of the Polymorphisms Among Different Populations

The SNPs with statistical differences in the comparison between the two ethnic groups are summarized in **Table 3**. In the comparison between Daur and Ewenki, Daur and Hezhen, Daur and Manchu, Daur and Monngolin, there were three, four, eight and four SNPs with statistical difference, respectively. In the comparison between Ewenki and Hezhen, Ewenki and Manchu, there were three and two SNPs with statistical difference, respectively. In the comparison between Manchu and Hezhen, Mongolin and Hezhen, Mongolin and Manchu, there were eleven, two and three SNPs with statistical difference, respectively ($p < 0.05$).

The average gene heterozygosity (AVE_HET) of the 23 SNPs in the five populations was 0.398 (**Table 4**). The average observed heterozygosity (OBS_HET) was 0.3957. The observed heterozygosity of rs1815739 and rs540874 in five populations was relatively large. The observed heterozygosity of rs4762 was the lowest. The F_{ST} values of the 23 SNPs among the five populations varied from 0.0009 to 0.0137, with an average of 0.0049, that is, 0.49% genetic variation existed between populations and 99.51% genetic variation existed within populations (**Table 5**). The gene flow of rs3856806 and rs1815739 was relatively large, and the mean value of Nm was 50.6913.

Haplotype Analysis

There were five blocks in 23 SNPs, the r^2 threshold of haplotype blocks were 0.8 Five blocks were distributed in *VDR*, *ACE*, *ACTN3*, *EPO* and *IGF1* genes (**Table 6**; **Figure 2**). The five blocks with statistical differences were mainly concentrated in *VDR* and *ACE* genes. The results showed that there were differences in haplotype distribution among the five ethnic groups. A block1 containing two SNPs was constructed in the *VDR* gene. The most common haplotype was CC, followed by AT and AC. The frequency distribution of CC was statistically significant between Daur

TABLE 4 | summary of heterozygosity statistics for 23 SNPs.

Loci	Sample size	Obs_Hom	Obs_Het	Exp_Hom[a]	Exp_Het[a]	Nei[b]	Ave_Het
rs699	1748	0.6899	0.3101	0.6844	0.3156	0.3154	0.3144
rs4762	1728	0.8692	0.1308	0.8737	0.1263	0.1262	0.1187
rs5051	1712	0.6904	0.3096	0.6794	0.3206	0.3204	0.3231
rs5050	1728	0.735	0.265	0.7213	0.2787	0.2785	0.2729
rs3856806	1746	0.6999	0.3001	0.7122	0.2878	0.2876	0.2918
rs2066992	1764	0.5714	0.4286	0.5531	0.4469	0.4467	0.454
rs1617640	1760	0.592	0.408	0.5915	0.4085	0.4082	0.4166
rs551238	1764	0.576	0.424	0.5803	0.4197	0.4195	0.4246
rs1815739	1760	0.4773	0.5227	0.5045	0.4955	0.4952	0.4945
rs540874	1728	0.4884	0.5116	0.5044	0.4956	0.4953	0.4945
rs5742714	1752	0.7671	0.2329	0.7585	0.2415	0.2414	0.2305
rs1520220	1762	0.5289	0.4711	0.5271	0.4729	0.4727	0.4728
rs5742612	1764	0.6088	0.3912	0.6115	0.3885	0.3882	0.3897
rs972936	1762	0.5278	0.4722	0.5252	0.4748	0.4745	0.4741
rs7975232	1764	0.6088	0.3912	0.6159	0.3841	0.3839	0.3968
rs757343	1,620	0.7037	0.2963	0.7075	0.2925	0.2923	0.3042
rs2228570	1736	0.5472	0.4528	0.5191	0.4809	0.4806	0.4805
rs1156882	1728	0.5255	0.4745	0.5186	0.4814	0.4811	0.4733
rs4309	1760	0.5432	0.4568	0.5293	0.4707	0.4704	0.4671
rs4331	1764	0.5431	0.4569	0.5346	0.4654	0.4651	0.4619
rs4341	1762	0.5437	0.4563	0.5341	0.4659	0.4656	0.4624
rs4343	1762	0.5414	0.4586	0.5347	0.4653	0.465	0.4621
rs4362	1758	0.5199	0.4801	0.5238	0.4762	0.4759	0.4726
Mean	1745	0.6043	0.3957	0.6019	0.3981	0.3978	0.398
St. Dev		0.1005	0.1005	0.1003	0.1003	0.1002	0.1009

[a]*Expected homozygosty and heterozygosity were computed using Levene (1949).*
[b]*Nei's (1973) expected heterozygosit.*

TABLE 5 | Summary of the F-Statistics and gene flow for all the SNPs in five populations.

Loci	Sample size	Fis	Fit	Fst	Nm[a]
rs699	1748	0.0069	0.0123	0.0055	45.5111
rs4762	1728	–0.0257	-0.0207	0.0049	50.9974
rs5051	1712	0.0228	0.0279	0.0052	48.2872
rs5050	1728	0.0338	0.038	0.0043	57.7779
rs3856806	1746	–0.024	–0.0229	0.0011	233.5086
rs2066992	1764	0.04	0.0532	0.0137	17.9482
rs1617640	1760	–0.0027	0.0017	0.0044	56.6706
rs551238	1764	–0.0338	–0.0318	0.0019	130.0856
rs1815739	1760	–0.0617	–0.0607	0.0009	264.183
rs540874	1728	–0.0356	–0.0343	0.0013	195.3216
rs5742714	1752	0.0327	0.0365	0.0039	63.7296
rs1520220	1762	0.0164	0.019	0.0026	95.5691
rs5742612	1764	–0.0205	–0.0191	0.0014	178.7291
rs972936	1762	0.0147	0.0179	0.0032	76.7436
rs7975232	1764	–0.0614	–0.0486	0.0121	20.3944
rs757343	1620	–0.05	–0.0418	0.0078	31.9331
rs2228570	1736	0.0887	0.0939	0.0057	43.3971
rs11568820	1728	0.0301	0.0406	0.0108	22.8355
rs4309	1760	0.0405	0.0436	0.0033	75.9503
rs4331	1764	0.0286	0.0341	0.0056	44.638
rs4341	1762	0.0321	0.0371	0.0052	48.2127
rs4343	1762	0.0281	0.0328	0.0049	51.1232
rs4362	1758	0.006	0.0094	0.0034	72.6963
Mean	1745	0.0065	0.0114	0.0049	50.6913
St. Dev		0.0371	0.0376	0.0033	67.5013

[a]*Nm = Gene flow estimated from F_{ST} = 0.25 (1 - F_{ST})/F_{ST}.*

and Ewenki (P = 0.0167) and between Daur and Mongolian (P = 0.0101). The frequency distribution of AT in Daur and Ewenki (P = 0.0045), Daur and Manchu (P = 1.00E-04), and Daur and Mongolian (P = 0.0071) were statistically significant. The frequency distribution of AC in Daur and Hezhen (P = 1.00E-04), Daur and Manchu (P = 0.0012), Ewenki and

TABLE 6 | Haplotype frequencies in five ethnic populations.

Gene	Block	Haplotype	Daur	Hezhen	Ewenki	Manchu	Mongolian
VDR	Block 1	CC	0.764[b,d]	0.788	0.664	0.723	0.67
		AT	0.134[b,c,d]	0.173	0.236	0.232	0.215
		AC	0.102[a,c]	0.039	0.100[c]	0.046	0.115
ACE	Block 2	GCC	0.605	0.554	0.609	0.638	0.613
		AGT	0.38[c]	0.398	0.383[c]	0.295	0.33
		GCT	0.013[a,c,d]	0.034	0.007[c]	0.05	0.038
ACTN3	Block 3	CG	0.571	0.522	0.541	0.548	0.557
		TA	0.425	0.476	0.452	0.452	0.443
EPO	Block 4	AT	0.691	0.72	0.644	0.704	0.66
		CG	0.281	0.258	0.322	0.279	0.302
		AG	0.026	0.02	0.014	0.014	0.019
IGF1	Block 5	CC	0.633	0.642	0.596	0.575	0.625
		CG	0.227	0.216	0.288	0.258	0.271
		GG	0.139	0.142	0.116	0.167	0.104

[a]*Compared with Hezhen, p < 0.05.*
[b]*Compared with Ewenki, p < 0.05.*
[c]*Compared with Manchu, p < 0.05.*
[d]*Compared with Mongolian, p < 0.05.*

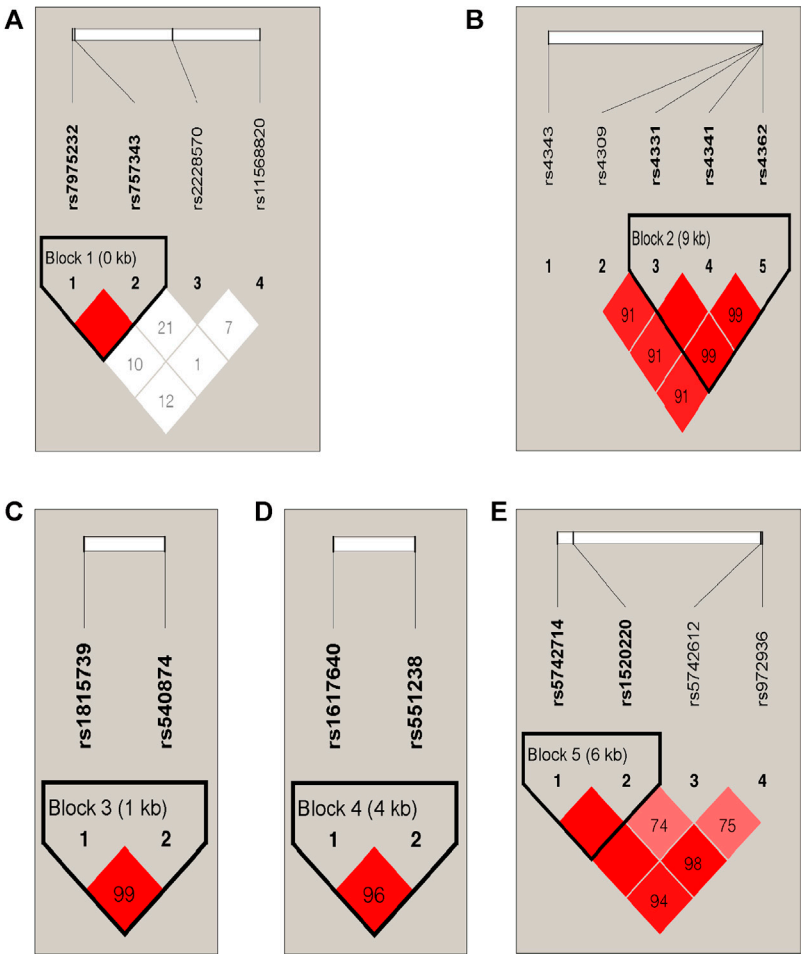

FIGURE 2 | LD Plot. **(A)** The block1 containing two SNPs was constructed in the *VDR* gene. **(B)** The block2 containing three SNPs was constructed in the *ACE* gene. **(C)** The block3 containing two SNPs was constructed in the *ACTN3* gene. **(D)** The block4 containing two SNPs was constructed in the *EPO* gene. **(E)** The block5 containing two SNPs was constructed in the *IGF1* gene.

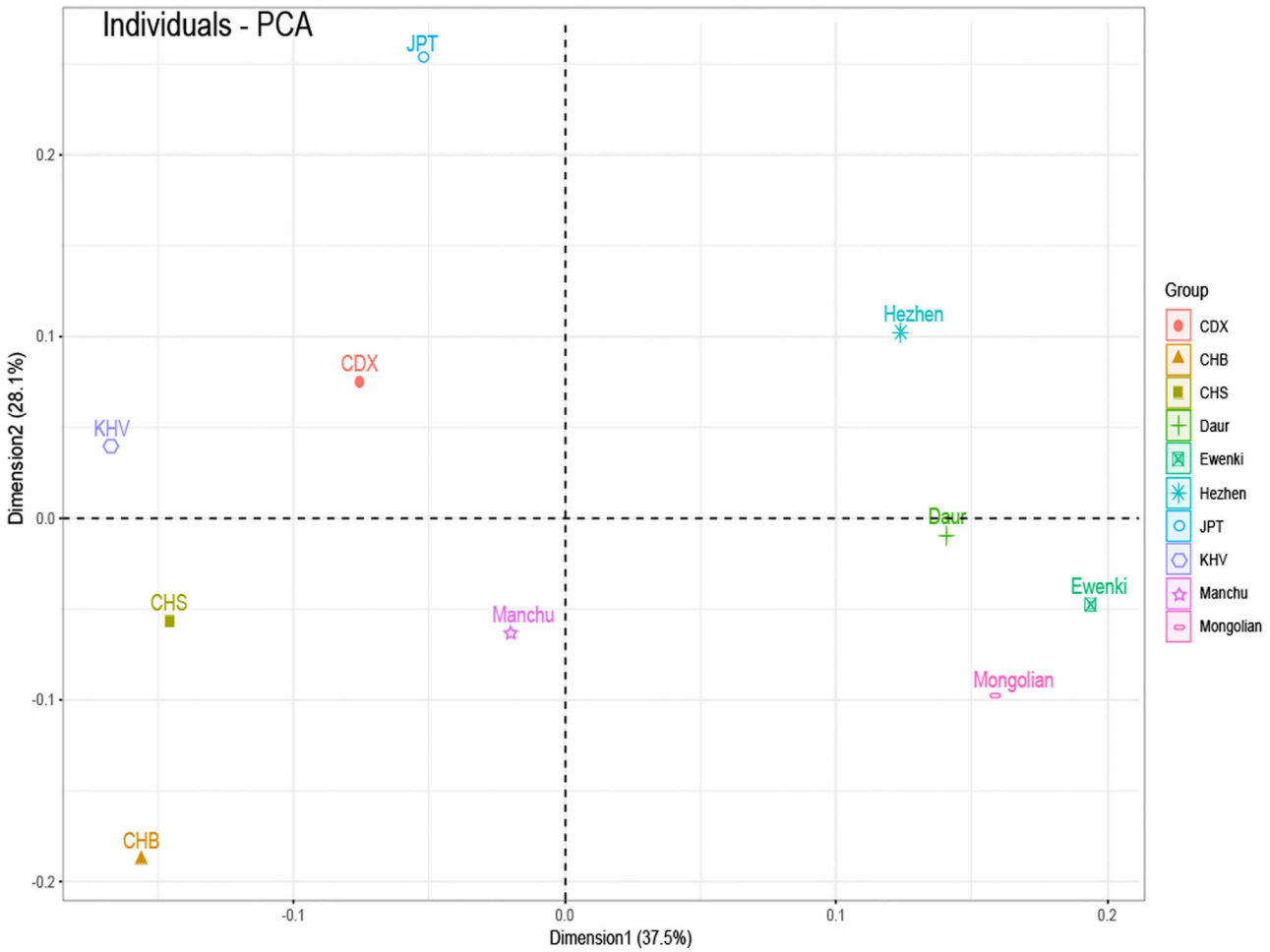

FIGURE 3 | PCA based on the minimum allele frequency of the Asia populations. Note: CDX: Chinese Dai in Xishuangbanna; CHB: Han Chinese in Beijing; CHS: Southern Han Chinese; JPT: Japanese in Tokyo; KHV: Kinh in Ho Chi Minh City, Vietnam.

Manchu (P = 0.004) has statistical significance. The block2 containing three SNPs was constructed in the ACE gene. The most common haplotype was GCC, followed by AGT and GCT. The frequency distribution of AGT in Daur and Manchu (P = 0.0073), Ewenki and Manchu (P = 0.0475) is statistically significant, GCT in Daur and Hezhen (P = 0.0311), Daur and Manchu (P = 0.0013), Daur and Mongolian (P = 0.0339). Ewenki and Manchu (P = 0.0206) is statistically significant.

Inter Population Genetic Distances

F_{ST} value between five populations based on 23 SNPs indicated that the F_{ST} values of Daur and Mongolian (0.0026), Daur and Ewenki (0.0027), Mongolian and Ewenki (0.0006) are relatively small (**Supplementary Table S2**). According to the Nei's genetic distance of the five ethnic groups based on UPGMA method (**Supplementary Table S3**). The genetic distance between Daur and Mongolian was relatively close (0.0035); the genetic distance between Mongolian and Ewenki was the closest (0.0007); The genetic distance between Daur and Ewenki was relatively close (0.0036). According to the PCA plot of the Asia populations (**Figure 3**), PC1 and PC2 accounted for 37.5 and 28.1% of the total genetic variation, respectively. The genetic distance between Daur, Ewenki and Mongolian were relatively close, which was consistent with the result of the FST value and the Nei's Genetic Distance between the five ethnic groups (**Supplementary TableS2,3**). According to the PCA plot of the world populations (**Figure 4A**), PC1 and PC2 accounted for 51.7 and 32.7% of the total genetic variation, respectively. PCA plot divided the 31 world populations into five groups, namely AFR, AMR, EAS, EUR, and SAS, which named according to their geographic location of African, American, East Asian, European and South Asian. Population belonging to the same large group are generally clustered together, which are consistent with results from the phylogenetic tree of the world populations (**Figure 4B**). We found that the five ethnic groups included in the study were clustered in one cluster with the Asian population. In addition, the mean F_{ST} values and the mean Nm values of the 23 SNPs among the 31 populations was 0.098, 2.3017, respectively (**Supplementary Table S4**). According to the analysis of

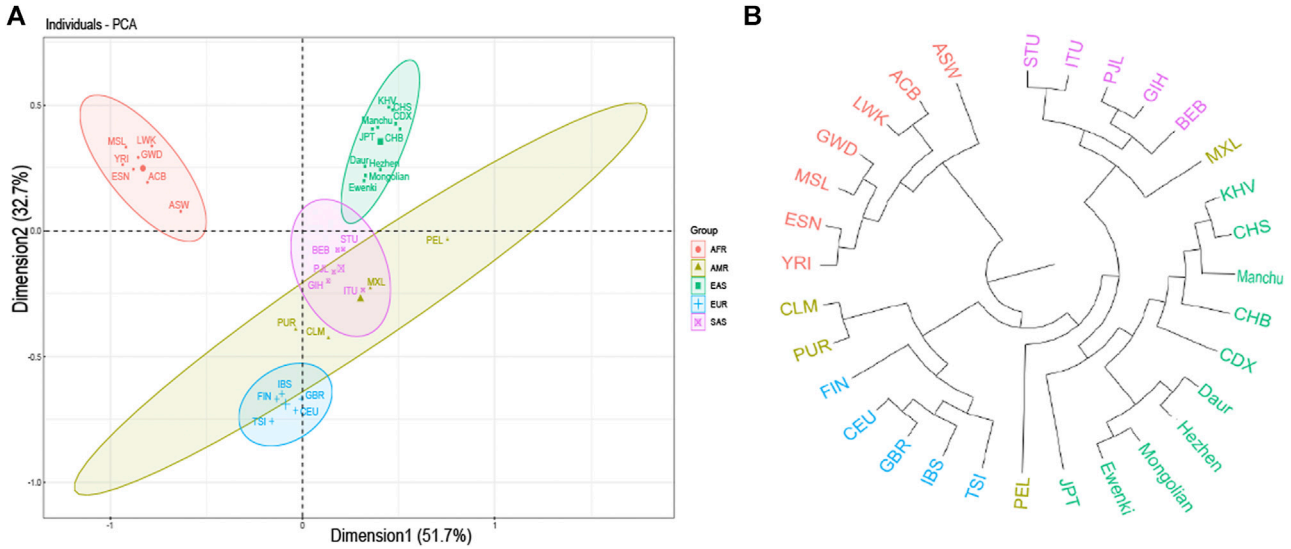

FIGURE 4 | PCA and Neighbor-joining tree results. **(A)** PCA based on the minimum allele frequency of the world populations. **(B)** Neighbor-joining tree based on genetic distance from the world population genotype data. Note: CHB: Han Chinese in Beijing; JPT: Japanese in Tokyo; CHS: Southern Han Chinese; CDX: Chinese Dai in Xishuangbanna; KHV: Kinh in Ho Chi Minh City, Vietnam; CEU: Utah residents (CEPH) with Northern and Western European Ancestry; TSI: Toscani in Italia; FIN: Finnish in Finland; GBR: British in England and Scotland; IBS: Iberian Population in Spain; YRI: Yoruba in Ibadan, Nigeria; LWK: Luhya in Webuye, Kenya; GWD: Gambian in Western Divisions in the Gambia; MSL: Mende in Sierra Leone; ESN: Esan in Nigeria; ASW: Americans of African Ancestry in SW United States; ACB: African Caribbeans in Barbados; MXL: Mexican Ancestry from Los Angeles United States; PUR: Puerto ricans from Puerto rico; CLM: Colombians from Medellin, Colombia; PEL: Peruvians from Lima, Peru; GIH: Gujarati Indian from Houston, Texas; PJL: Punjabi from Lahore, Pakistan; BEB: Bengali from Bangladesh; STU: Sri Lankan Tamil from the United Kingdom; ITU: Indian Telugu from the United Kingdom.

molecular variance (AMOVA) among the 31 populations, the percentage of variation among groups was 0.83%, while the percentage of variation within populations was 99.15% (**Supplementary Table S5**).

DISCUSSION

Different nations have formed specific genetic structures of different cultures, phenotypes and languages under the natural selection of different environments, material conditions and various pathogens. In East Asia, China has the largest Han population in the world, with 55 officially recognized ethnic groups making up their specific cultural backgrounds. They speak more than nine language families in China (Chen et al., 2019). Among them, five ethnic groups belonging to the Altaic language family in Heilongjiang province in northern China may have stronger physical fitness (Bi, 1993), the performance of the basic ability of human muscle activity. Some studies have found that there is a significant association between genotype and skeletal muscle phenotype. For example, the presence of SNPs is associated with better skeletal muscle strength performance (Khanal et al., 2020). We selected the 23 SNPs included in this study were all focused on genes related to skeletal strength, to further study the genetic composition and phylogeny of the five ethnic groups. 23 SNPs are consistent with the Hardy Weinberg equilibrium. In addition, in the pair comparison of the five populations studied, the genetic differences were mainly found on genes *IL6*, *VDR*, *AGT*, *ACE* and *IGF1*. for example, AGT encodes angiotensinogen, a protein involved in the renin-angiotensin-aldosterone system (RAAS) and is related to muscle growth (Ben-Zaken et al., 2019). *IGF1* is an important regulator not only of muscle mass and function, but also of bone. This is true not only during development, but throughout the human life cycle (Moriwaki et al., 2019). Vitamin D levels are closely related to the presence of vitamin D receptors in most human exoskeletal cells, and exposure to vitamin D in skeletal muscle leads to the expression of multiple myogenic transcription factors that promote the proliferation and differentiation of muscle cells (Wiciński et al., 2019). The angiotensin-converting enzyme (*ACE*) gene is associated with superior muscle metabolic performance and muscle endurance (Vaughan et al., 2013). Erythropoietin plays an important role in regulating metabolic homeostasis and bone remodeling (Suresh et al., 2019). Interleukin-6 (*IL-6*), the prototype of muscle factor, was identified as a muscle-derived cytokine 15 years ago (Karstoft and Pedersen, 2016). F_{ST} plays a core role in population and evolutionary genetics, it can reflect the degree of genetic differentiation between populations (Meirmans and Hedrick, 2011). The F_{ST} values of the 23 SNPs among the five ethnic groups varied from 0.0009 to 0.0137. There is almost no genetic differentiation in each population. According to the mean value of Nm, indicating that genetic differentiation did not occur between populations, but was mainly caused by genetic differentiation within populations, this is consistent with the population genetic differentiation results shown by the F_{ST} value of this study. We found that there was little difference in genetic distance between the five populations studied on the whole, this may because the five ethnic groups are all located in Heilongjiang province. which is consistent with the

geographical location of the population (Tian et al., 2008). In addition, a total of five blocks exist in 23 SNPs (**Figure 2**). We concluded that rs7975232 and rs1815739 were statistically different in the five ethnic groups based on the F_{ST} values (**Table 5**). The same gene can perform different functions in the body, we found that the rs7975232 of *VDR* gene was related to the obesity and diabetes, it is also as the genetic makers of them. rs7975232 polymorphism of *VDR* gene was found to be positively correlated with obesity according to skin fold thickness and body fat rate in Chinese Han population (Shen et al., 2019). Another recent study also found that rs797523 polymorphism appears to be associated with overweight/obesity (Wang et al., 2021). The five ethnic groups in Heilongjiang province may be at higher risk of obesity or overweight due to environmental、eating habits and genetic factors because they are located in the extremely cold area of northern China. As we known, obesity is an important risk factor for diabetes. Meanwhile, it may reveal that the ethnic groups in the extremely cold area of northern China may be susceptible to diabetes to a certain extent (Li et al., 2020). Interestingly, another locus of significant genetic variation explains exactly how extreme cold affects skeletal muscle in humans. The positive selection of the allele of rs1815739 in cold climates provides the mechanism, that is, the slower type of I MyHC in the α-actinin-3 muscle, combined with a shift in neuronal muscle activation to increase muscle tone rather than obvious tremor, supporting the key thermogenesis of human skeletal muscle during cold exposure (Wyckelsma et al., 2021). Therefore, we believe that the genetic difference of rs1815739 and rs7975232 among the five ethnic groups may be caused by the fact that the five ethnic groups are located in Heilongjiang province, a high-latitude and severe cold region in China. The largest component of genotypic variation is the reduction of high-order data (all genotypes) to low-order variation. According to the PCA results of the Asian population, the genetic distances of Daur, Ewenki, and Mongolian were relatively close, indicates that in the mixing process of history and modern times. There may be gene exchange between Daur and Mongolian and Ewenki to some extent (Liu et al., 2007; Gao et al., 2006), which was consistent with the result of the Nei's genetic distance and F_{ST} values between the five ethnic groups. There are also studies showing that from the perspective of linguistic kinship, immigration history and origin, the kinship between the Mongolian and the Daur is very close, which indicates that the two groups in each pair may be of the same origin (Hou et al., 2007). According to the world population phylogeny tree and PCA, the genetic distance between the five populations and the Asian population is relatively close, and they are clustered with the Asian population. The genetic variation of 31 populations occurred mainly within the population (**Supplementary Table S4,5**).

In Conclusion, geographical and linguistic divisions have shaped the genetic structure of modern populations. Cluster analysis shows that the five ethnic groups in Heilongjiang province are clustered together with the East Asian ethnic groups. The genetic distance between Daur, Mongolian and Ewenki is closer, in order to better study the genetic characteristics of skeletal muscle strength related genes in different population, in addition to the more national, cultural, geographical and linguistic diversity group, also need more genome data combining with archaeological data and population history for further analysis and validation.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The study involving human participants were reviewed and approved by Harbin Medical University. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SF, QL and BD came up with the idea for this study. JB and SF perform or supervise laboratory work. BD, QL, TZ and MJ conducted experiments. BD, QL, TZ, MJ, XL and YF analysis data. BD wrote the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.756802/full#supplementary-material

REFERENCES

Ahmetov, I. I., and Fedotovskaya, O. N. (2015). Current Progress in Sports Genomics. *Adv. Clin. Chem.* 70, 247–314. doi:10.1016/bs.acc.2015.1003.1003

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: Analysis and Visualization of LD and Haplotype Maps. *Bioinformatics* 21 (2), 263–265. doi:10.1093/bioinformatics/bth457

Ben-Zaken, S., Eliakim, A., Nemet, D., and Meckel, Y. (2019). Genetic Variability Among Power Athletes: The Stronger vs. The Faster. *J. Strength Cond Res.* 33 (6), 1505–1511. doi:10.1519/JSC.0000000000001356

Ben-Zaken, S., Meckel, Y., Nemet, D., Kassem, E., and Eliakim, A. (2019). Genetic Basis for the Dominance of Israeli Long-Distance Runners of Ethiopian Origin. *J. Strength Cond Res.* 35 (7), 1885–1896. Publish Ahead of Print. doi:10.1519/jsc.0000000000002989

Bi, L. (1993). Analysis of Physical Constitution of Manchu College Students. *Chin. J. Sch. Doctor.* 3, 13–16.

Chen, B., Wang, Q., Xue, Y., and Li, P. (1999). Research on Physical Anthropology of the Hezhe Nationality in China. *J. Yunnan Univ. Nat. Sci. Edition* 1999 (S3), 267. doi:10.1023/a:1018753013035

Chen, P., Wu, J., Luo, L., Gao, H., Wang, M., Zou, X., et al. (2019). Population Genetic Analysis of Modern and Ancient DNA Variations Yields New Insights into the Formation, Genetic Structure, and Phylogenetic Relationship of Northern Han Chinese. *Front. Genet.* 10, 1045. doi:10.3389/fgene.2019.01045

Dong, X., Ding, F., and Hou, B. (2004). "Investigation and Research on the Physical Fitness of the Adults of Mongolian, Daur, Ewenki and Oroqen Minority

Nationalities in Inner Mongolia Autonomous Region", in: National Sports Science Conference, Beijing, China.

Excoffier, L., Laval, G., and Schneider, S. (2007). Arlequin (Version 3.0): an Integrated Software Package for Population Genetics Data Analysis. *Evol. Bioinform Online* 1, 47–50. doi:10.1143/JJAP.34.L418

Frontera, W. R., and Ochala, J. (2015). Skeletal Muscle: a Brief Review of Structure and Function. *Calcif Tissue Int.* 96 (3), 183–195. doi:10.1007/s00223-00014-09915-y

Galinsky, K. J., Reshef, Y. A., Finucane, H. K., Loh, P.-R., Zaitlen, N., Patterson, N. J., et al. (2019). Estimating Cross-Population Genetic Correlations of Causal Effect Sizes. *Genet. Epidemiol.* 43 (2), 180–188. doi:10.1002/gepi.22173

Gao, Y., Jin, T. B., and Li, S. B. (2006). Genetic Relationships between Ewenki Minority in Inner Mongolia and Other 14 Groups. *Zhong Nan Da Xue Xue Bao Yi Xue Ban* 31 (4), 475–478.

Guilherme, J. P. L. F., Bertuzzi, R., Lima-Silva, A. E., Pereira, A. d. C., and Lancha Junior, A. H. (2018). Analysis of Sports-relevant Polymorphisms in a Large Brazilian Cohort of Top-level Athletes. *Ann. Hum. Genet.* 82 (5), 254–264. doi:10.1111/ahg.12248

Hou, Q.-F., Yu, B., and Li, S.-B. (2007). Genetic Polymorphisms of Nine X-STR Loci in Four Population Groups from Inner Mongolia, China. *Genomics, Proteomics. Bioinformatics* 5 (1), 59–65. doi:10.1016/S1672-0229(1007)60015-60011

Jones, A., Montgomery, H. E., and Woods, D. R. (2002). Human Performance: a Role for the ACE Genotype? *Exerc. Sport Sci. Rev.* 30 (4), 184–190. doi:10.1097/00003677-200210000-00008

Karstoft, K., and Pedersen, B. K. (2016). Skeletal Muscle as a Gene Regulatory Endocrine Organ. *Curr. Opin. Clin. Nutr. Metab. Care* 19 (4), 270–275. doi:10.1097/MCO.0000000000000283

Khanal, P., He, L., Herbert, A. J., Stebbings, G. K., Onambele-Pearson, G. L., Degens, H., et al. (2020). The Association of Multiple Gene Variants with Ageing Skeletal Muscle Phenotypes in Elderly Women. *Genes (Basel)* 11 (12), 1459. doi:10.3390/genes11121459

Książek, A., Zagrodna, A., and Słowińska-Lisowska, M. (2019). Vitamin D, Skeletal Muscle Function and Athletic Performance in Athletes-A Narrative Review. *Nutrients* 11 (8), 1800. doi:10.3390/nu11081800

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33 (7), 1870–1874. doi:10.1093/molbev/msw1054

Lamon, S., and Russell, A. P. (2013). The Role and Regulation of Erythropoietin (EPO) and its Receptor in Skeletal Muscle: How Much Do We Really Know? *Front. Physiol.* 4, 176. doi:10.3389/fphys.2013.00176

Li, Y., Teng, D., Shi, X., Qin, G., Qin, Y., Quan, H., et al. (2020). Prevalence of Diabetes Recorded in mainland China Using 2018 Diagnostic Criteria from the American Diabetes Association: National Cross Sectional Study. *Bmj* 369, m997. doi:10.1136/bmj.m997

Liu, Y., Huang, Y., Li, X., Wang, Z., Shen, G., and Feng, J. (2007) "Study on Y-SNP of Hezhen, Ewenki, Daur and Mongolian Nationalities in Northeast China," in *Genetics Progress and Population Health Summit Forum*. Kunming, Yunnan, China.

Luu, K., Bazin, E., and Blum, M. G. B. (2017). Pcadapt: anRpackage to Perform Genome Scans for Selection Based on Principal Component Analysis. *Mol. Ecol. Resour.* 17 (1), 67–77. doi:10.1111/1755-0998.12592

Meirmans, P. G., and Hedrick, P. W. (2011). Assessing Population Structure: F ST and Related Measures. *Mol. Ecol. Resour.* 11 (1), 5–18. doi:10.1111/j.1755-0998.2010.02927.x

Moriwaki, K., Matsumoto, H., Tanishima, S., Tanimura, C., Osaki, M., Nagashima, H., et al. (2019). Association of Serum Bone- and Muscle-Derived Factors with Age, Sex, Body Composition, and Physical Function in Community-Dwelling Middle-Aged and Elderly Adults: a Cross-Sectional Study. *BMC Musculoskelet. Disord.* 20 (1), 276. doi:10.1186/s12891-019-2650-9

Norouzi, S., Adulcikas, J., Henstridge, D., Sonda, S., Sohal, S., and Myers, S. (2019). The Zinc Transporter Zip7 Is Downregulated in Skeletal Muscle of Insulin-Resistant Cells and in Mice Fed a High-Fat Diet. *Cells* 8 (7), 663. doi:10.3390/cells8070663

Oh, S. D. (2007). The Distribution of I/D Polymorphism in the ACE Gene Among Korean Male Elite Athletes. *J. Sports Med. Phys. Fitness* 47 (2), 250–254.

Phu, S., Boersma, D., and Duque, G. (2015). Exercise and Sarcopenia. *J. Clin. Densitom.* 18 (4), 488–492. doi:10.1016/j.jocd.2015.04.011

Pickering, C., Suraci, B., Semenova, E. A., Boulygina, E. A., Kostryukova, E. S., Kulemin, N. A., et al. (2019). A Genome-wide Association Study of Sprint Performance in Elite Youth Football Players. *J. Strength Cond Res.* 33 (9), 2344–2351. doi:10.1519/jsc.0000000000003259

Pitsiladis, Y. P., Tanaka, M., Eynon, N., Bouchard, C., North, K. N., Williams, A. G., et al. (2016). Athlome Project Consortium: a Concerted Effort to Discover Genomic and Other "omic" Markers of Athletic Performance. *Physiol. Genomics* 48 (3), 183–190. doi:10.1152/physiolgenomics.00105.2015

Seto, J. T., Roeszler, K. N., Meehan, L. R., Wood, H. D., Tiong, C., Bek, L., et al. (2021). ACTN3 Genotype Influences Skeletal Muscle Mass Regulation and Response to Dexamethasone. *Sci. Adv.* 7 (27), eabg0088. doi:10.1126/sciadv.abg0088

Shen, F., Wang, Y., Sun, H., Zhang, D., Yu, F., Yu, S., et al. (2019). Vitamin D Receptor Gene Polymorphisms Are Associated with Triceps Skin Fold Thickness and Body Fat Percentage but Not with Body Mass index or Waist Circumference in Han Chinese. *Lipids Health Dis.* 18 (1), 97. doi:10.1186/s12944-019-1027-2

Singh, G., and Soman, B. (2019). *Data Visualisation Using Ggplot2 Package in R.*

Suresh, S., Rajvanshi, P. K., and Noguchi, C. T. (2019). The Many Facets of Erythropoietin Physiologic and Metabolic Response. *Front. Physiol.* 10, 1534. doi:10.3389/fphys.2019.01534

Tian, C., Kosoy, R., Lee, A., Ransom, M., Belmont, J. W., Gregersen, P. K., et al. (2008). Analysis of East Asia Genetic Substructure Using Genome-wide SNP Arrays. *PLoS One* 3 (12), e3862. doi:10.1371/journal.pone.0003862

Vassilakos, G., and Barton, E. R. (2018). Insulin-Like Growth Factor I Regulation and its Actions in Skeletal Muscle. *Compr. Physiol.* 9 (1), 413–438. doi:10.1002/cphy.c180010

Vaughan, D., Huber-Abel, F. A., Graber, F., Hoppeler, H., and Flück, M. (2013). The Angiotensin Converting Enzyme Insertion/deletion Polymorphism Alters the Response of Muscle Energy Supply Lines to Exercise. *Eur. J. Appl. Physiol.* 113 (7), 1719–1729. doi:10.1007/s00421-012-2583-6

Wang, D., Su, K., Ding, Z., Zhang, Z., and Wang, C. (2021). Association of Vitamin D Receptor Gene Polymorphisms with Metabolic Syndrome in Chinese Children. *Ijgm* 14, 57–66. doi:10.2147/ijgm.S287205

Wang, Z., Yi, G., Meng, F., Pan, H., Cui, X., Cui, X., et al. (2014). *Investigation and Research on the Current Situation of Hezhe Nationality's Population Physique. Sports World*. Academic Edition 5, 129–131. doi:10.16730/j.cnki.61-1019/g8.2014.05.056

Weyerstraß, J., Stewart, K., Wesselius, A., and Zeegers, M. (2018). Nine Genetic Polymorphisms Associated with Power Athlete Status - A Meta-Analysis. *J. Sci. Med. Sport.* 21 (2), 213–220. doi:10.1016/j.jsams.2017.06.012

Wiciński, M., Adamkiewicz, D., Adamkiewicz, M., Śniegocki, M., Podhorecka, M., Szychta, P., et al. (2019). Impact of Vitamin D on Physical Efficiency and Exercise Performance-A Review. *Nutrients* 11 (11), 2826. doi:10.3390/nu11112826

Wyckelsma, V. L., Venckunas, T., Houweling, P. J., Schlittler, M., Lauschke, V. M., Tiong, C. F., et al. (2021). Loss of α-actinin-3 during Human Evolution Provides superior Cold Resilience and Muscle Heat Generation. *Am. J. Hum. Genet.* 108 (3), 446–457. doi:10.1016/j.ajhg.2021.01.013

Yan, X., Papadimitriou, I., Lidor, R., and Eynon, N. (2016). Nature versus Nurture in Determining Athletic Ability. *Med. Sport Sci.* 61, 15–28. doi:10.1159/000445238

Yeh, F. C., Yang, R. C., Boyle, T., Ye, Z. H., and Mao, J. X. (1997). *POPGENE, the User-Friendly Shareware for Population Genetic Analysis*. Molecular Biology and Biotechnology Centre. University of Alberta.

Zarębska, A., Sawczyn, S., Kaczmarczyk, M., Ficek, K., Maciejewska-Karłowska, A., Sawczuk, M., et al. (2013). Association of Rs699 (M235T) Polymorphism in the AGT Gene with Power but Not Endurance Athlete Status. *J. Strength Cond Res.* 27 (10), 2898–2903. doi:10.1519/JSC.0b013e31828155b5

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 October 2021
doi: 10.3389/fgene.2021.736876

ACE and *ACTN3* Gene Polymorphisms and Genetic Traits of Rowing Athletes in the Northern Han Chinese Population

Qi Wei [1,2]*

[1]Key Laboratory of General Administration of Sport of China, Wuhan, China, [2]Hubei Institute of Sports Science, Wuhan, China

This investigation aimed to explore the effects of *ACE* I/D and *ACTN3* R577X gene polymorphisms on specific quantitative variables, including height, weight, arm span, biacromial breadth, forced vital capacity (FVC), FVC/weight, maximal oxygen uptake (VO_2max), prone bench pull (PBP), loaded barbell squat (LBS), and 3,000-m run, in 243 Chinese rowing athletes. The *ACE* and *ACTN3* genotypes were obtained for each athlete *via* polymerase chain reaction on saliva samples, and the genotype frequency was analyzed. The *ACE* genotype frequency of rowing athletes were 45.8% II, 42.2% ID, and 12% DD for males and 33.6% II, 48% ID, and 18.4% DD for females. There were significant differences in weight in male athletes, PBP in female athletes, and *ACE* genotypes. A linear regression analysis using PBP and LBS as different dependent variables and *ACE* genotypes as independent variables based on the *ACE* I allele additive genetic effect showed a statistical significance in female athletes ($p < 0.05$). There was a significant difference in the distribution of the three genotypes among male athletes (36.7% XX, 38.5% RX, and 24.8% RR, $\chi 2 = 5.191$, df = 2, $p = 0.022 < 0.05$). There were no significant differences in the distribution of the three genotypes among female athletes (23.8% XX, 47.8% RX, 28.4% RR, $\chi^2 = 0.24$, df = 2, $p = 0.619 > 0.05$). The *ACTN3* gene polymorphism of male rowing athletes was dominated by the *ACTN3* 577X allele. There were significant differences in the χ^2 test between groups of male athletes. The *ACTN3* R577 allele was dominant in female athletes. There were significant differences between PBP and FVC/body weight and *ACTN3* genotypes in male athletes by ANOVA, respectively ($p < 0.05$). A linear regression analysis using FVC and FVC/body weight as dependent variables and *ACTN3* genotypes as independent variables based on the *ACTN3* 577X allele recessive genetic effect showed statistical significance in male athletes ($p < 0.05$). These results suggested that *ACE* and *ACTN3* gene polymorphisms may be used as biomarkers of genetic traits in Chinese rowing athletes.

Keywords: angiotensin converting enzyme, α-actin-3, single nucleotide polymorphism, trait variables, genetic model

OPEN ACCESS

Edited by:
Wibhu Kutanan,
Khon Kaen University, Thailand

Reviewed by:
Wanna Thongnoppakhun,
Wanna Thongnoppakhun, Thailand
Guanglin He,
Nanyang Technological University, Singapore

***Correspondence:**
Qi Wei
48750028@qq.com

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

Received: *06 July 2021*
Accepted: *20 September 2021*
Published: *14 October 2021*

Citation:
Wei Q (2021) ACE *and* ACTN3 *Gene Polymorphisms and Genetic Traits of Rowing Athletes in the Northern Han Chinese Population. Front. Genet. 12:736876. doi: 10.3389/fgene.2021.736876*

Abbreviations: FVC, forced vital capacity; VO2max, maximal oxygen uptake; PBP, prone bench pull; LBS, loaded barbell squat; PCR, polymerase chain reaction; ANOVA, analysis of variance; SNP, single-nucleotide polymorphism; ACE, angiotensin-converting enzyme; ACTN3, α-actinin-3; STREGA, Strengthening the Reporting of Genetic Association Studies; HWE, Hardy–Weinberg equilibrium; ALFA, NCBI Allele Frequency Aggregator; gnomAD, Genome Aggregation database; 1KGP, 1,000 Genomes Project; ChinaMAP, China Metabolic Analytics Project; CHB, Han Chinese in Beijing; CHS, Southern Han Chinese; Y DNA, Y chromosome; mtDNA, mitochondrial deoxyribonucleic acid.

INTRODUCTION

Human physical performance is widely accepted as individual characteristics dependent on interactions between genes and environment (Bouchard, 2011). Sports phenotypes, such as endurance, explosive power, muscle fiber type and proportion, and flexibility, are highly influenced by genetic and epigenetic factors (Zempo et al., 2017; Miyamoto-Mikami et al., 2018). These reviews showed the heritability of body height, body mass index, VO_2max adjusted for body weight, isometric grip strength, other isometric strength, isotonic strength, isokinetic strength, jumping ability, and other power measurements by meta-analysis 0.87–0.93 (Silventoinen et al., 2003), 0.47–0.9 (Elks et al., 2012), 0.56 (Miyamoto-Mikami et al., 2018), 0.56, 0.49, 0.49, 0.49, 0.55, and 0.51, respectively (Zempo et al., 2017). Environmental effects begin as early as preconception *via* gametic imprinting and continue after conception during the growth and development period (Bouchard, 2011).

Studies have suggested that genes are partially responsible for determining the anthropometric, physical, and physiological traits needed to achieve athletic performance (Pickering, 2019). Single-nucleotide polymorphisms (SNPs) are genetic sequence variations related to the expression variation of key genes regulating the physiological process of exercise (Boulygina, 2020). With the continuous development of genetic technology, it has become easier to further explore the genetic basis of excellent sports performance and find that SNPs and other genetic variations may directly or indirectly influence the athletic performance of aerobic exercise ability and other physiological characteristics, such as muscle strength and speed. Many SNPs associated with high heritability of phenotypes related to athlete status have been identified in the last 2 decades (Maciejewska-Skrendo et al., 2019; Semenova et al., 2019; Valeeva et al., 2019). The *ACE* I/D and *ACTN3* R577X polymorphisms have been intensely investigated with athletic performance in endurance- and power-oriented events (Jacob et al., 2018).

The *ACE* gene is a 21-kb single-copy gene located on 17q23 that encodes angiotensin-converting enzyme (ACE), which regulates human circulatory homeostasis, skeletal muscle growth, and cardiovascular functions; its I/D polymorphism denotes either an insertion or deletion of a 287-bp Alu repeat sequence at intron 16 (Pescatello et al., 2019). In 1998, the *ACE* I allele was found to be significant in elite British mountaineers (Montgomery et al., 1998). The proportion of Australian elite rowers carrying the *ACE* I allele was higher than that of the control group. The II genotype tends to reduce cardiac afterload during exercise and can effectively couple the ventricle and blood vessels to improve exercise endurance (Gayagay et al., 1998). Studies have speculated that the *ACE* I allele may decrease ACE enzyme activity to enhance human endurance performance (Ma et al., 2013; Pescatello et al., 2019). The D allele is associated with increased muscle volume-related baseline and rapid-twitch muscle fiber proportions with greater strength performance (Eider et al., 2013).

The *ACTN3* gene is located on chromosome 11, and its polymorphism is caused by the C-T polymorphism mutation at the R577X site, which produces a stop codon, resulting in the change of the amino acid at 577 from arginine (577 R) to a stop code (577X). Muscle contraction strength and speed are required at high levels of activity, and subjects carrying the 577X allele encode an early termination of actin3, resulting in muscle loss that affects muscle performance. Approximately 18% of the world population (approximately 1.5 billion people) has the XX genotype and deficiency of α-actinin-3 (ACTN3) without causing any significant muscle disease (Yang et al., 2003). However, some research found that XX homozygosity leading to ACTN3 deficiency can adversely impact sports performance through muscle type and energy metabolism (Papadimitriou et al., 2008; Cięszczyk et al., 2011; Mikami et al., 2014).

Rowing is competed over a 2,000-m track, requiring the rowers to exhibit extreme physiological power and endurance, technical proficiency, and environmental characteristics (Keenan et al., 2018). Studies have found that the performance of the rowers is related to individual anthropometric variables, such as height, weight, length of legs and body span, and muscular strength in the trunk and upper and lower limbs (Majumdar et al., 2017; Keenan et al., 2018; Maciejewski et al., 2019; Penichet-Tomas et al., 2019). Previous studies are limited to simulated case–control evaluations of *ACE* and *ACTN3* polymorphism designs based on exercise state without quantitative traits of athlete performance. This study is the first to explore the contributions of *ACE* and *ACTN3* gene polymorphisms and the effects of different alleles on sport performance-related phenotypic indicators of Asian athletes, including anthropometric, physical, and strength trait variables of Chinese elite rowing athletes, using one-way ANOVA and linear regression analysis based on two genetic models. We hope to determine the genetic effects and contributions of *ACE* and *ACTN3* gene polymorphisms to the phenotypic indexes related to Chinese rowing athletic ability, which may be conducive to cultivating outstanding athletes.

METHODS

Participants

We recruited 243 open-category elite rowers: 109 male athletes with an average age of 21.73 ± 2.32 years and training duration of 7.9 ± 1.8 years and 134 female athletes with an average age of 20.58 ± 1.24 years and training duration of 7.5 ± 1.4 years. All subjects were Han Chinese from five provinces (Henan, Shangdong, Hubei, Liaoning, and Jiangsu) who participated in the 2020 National Rowing Championship, including 36 international-level athletes (21 male and 15 female), 54 national-level athletes (24 male and 30 female), and 153 national second-level athletes (64 male and 89 female). The Sports Medicine Committee of the Hubei Sports Science Society Review Board approved the project, and written informed consent was obtained from all participants before testing. The guidelines for Strengthening the Reporting of Genetic Association Studies (Little et al., 2009), an extension of the Strengthening the Reporting of Observational Studies in Epidemiology statement, were followed to report the results of this study.

TABLE 1 | Primers and PCR conditions for polymorphism of *ACE* and *ACTN3* genes.

		***ACE* I/D, rs1799752**		***ACTN3* R577X, rs1815739**	
SNP, primers sequence(5' -3')		F:5' CTG GAG ACC ACT CCC ATC CTT TCT 3' R:5' GAT GTG GCC ATC ACA TTC GTC AGA 3'		F: 5′-CTG TTG CCT GTG GTA AGT GGG-3′ R:5′-TGG TCA CAG TAT GCA GGA GGG-3′	
Protocol(30 cycles)	Predenaturation	95°C	3 min	95°C	3 min
	Denaturation	95°C	30 s	95°C	30 s
	Annealing	60°C	30 s	68°C	30 s
	Extension	72°C	90 s	68°C	90 s
	extension(The final cycle)	72°C	8 min	68°C	8 min

Genotyping

Human genomic DNA was isolated from 2 ml of saliva sample collected in Oragene DNA OG-500 collection tubes (DNA Genotek, Canada) and stored at room temperature *via* a DNA extraction and purification kit (DNA Genotek prepIT-L2P, Canada). The primer was synthesized by Wuhan Gene Create Biological Engineering Co., and the amplification primer was also a sequencing primer. A 50-μl reaction system was used for polymerase chain reaction (PCR) amplification, which consisted of 50–80 μg/ml template DNA 2 μl, 10× KoD Buffer 5 μl (TOYOBO, Tokyo), 10 mmol/L forward primer and reverse primer 1.5 μl each, 2 mmol/L dNTPs 4 μl, 25 mM MgSO4 1 μl, KoD-plus amplification enzyme 1 μl (TOYOBO, Tokyo), and was supplemented with deionized water to 50 μl. PCR amplifications were performed by Applied Biosystems PCR (Thermo Fisher Scientific, United States). The primers and PCR conditions for the polymorphism analysis of the *ACE* and *ACTN3* genes are shown in **Table 1**.

The PCR products of the *ACE* gene were separated *via* 6% polyacrylamide gel electrophoresis and confirmed as follows: 190-bp fragment for DD genotype, 490-bp fragment for II genotype, and 490- and 190-bp fragments for ID genotype. The PCR products digested by the DdeI restriction enzyme (Promega) of each sample were detected as follows: 108-, 97-, and 86-bp fragments for the 577X allele; 205 and 86 bp for the 577 R allele; and sequenced by an ABI 3730 DNA Analyzer (Thermo Fisher Scientific, United States) to identify the *ACTN3* genotypes as reported by Eynon *et al.* (2009).

Measurement of Anthropometric Trait Variables

The athletes were required to wear sports attire, and measurements of their height, arm span, and biacromial breadth were performed using a large anthropometer (Anthroscan 3D VITUS, Human Solution, Germany). After removing their shoes at 7:00 in the morning, the height of the athletes was measured in an upright position, from the vertex to the floor and from the akropodion to dactylion, as morphological and technical parameters. Bisacromial breadth is usually considered the distance between the two acromial processes. Arm span was measured with the athletes standing straight while the arms were maximally stretched horizontally (making a 90° angle with the trunk). The distance between the tips of the middle fingers was then measured by an anthropometer.

Then, the athletes were measured by a body composition analyzer (X-Scan Plus II, Jawon, Korea) to obtain their weight. All of the measurements were repeated three times by the same observer, and the average of the measurements was taken as the final value.

Measurement of Physical and Strength Trait Variables

Spirometry was performed using the Chest Microspiro HI-501 vital capacity instrument (CHEST M.I. Inc, Tokyo, Japan). After maximal inhalation, the athletes sealed their lips around the mouthpiece and exhaled as hard and as fast as possible, and the FVC values were displayed digitally. Each athlete completed two tests, with an interval of 30 s, and the highest value was recorded. Finally, the test data were printed.

The VO_2max test was conducted according to previously reported procedures (Izquierdo-gabarren et al., 2010). The experimental equipment used was the German lung function tester MAX II connected to the German H/P/PULS platform.

Prone bench pull (PBP) and loaded barbell squat (LBS) were performed with the use of standard free-weight equipment (Salter, Madrid, Spain) according to the test methods (Riboli et al., 2021).

Statistical Analysis

The data in this study were statistically analyzed using SPSS 21.0 for Windows software. The frequency distribution of the *ACE* and *ACTN3* genotypes was verified by the Hardy–Weinberg equilibrium (HWE) law. We converted the 3,000-m run times to seconds and compared the height, weight, arm span, biacromial breadth, VO_2max, FVC, FVC/body weight, PBP, LBS, and 3,000-m run between the *ACTN3* R577X or *ACE* I/D genotypes using one-way ANOVA. Considering that rowing performance is mainly based on aerobic stamina, we selected two kinds of genetic models to calculate the genetic effects of the advantage of the *ACE* I allele and the 577X allele. One was the additive model (*i.e.*, the R/R = 0, R/X = 1, X/X = 2, or D/D = 0, I/D = 1, I/I = 2); the other was the recessive model (R/R = R/X = 0, X/X = 1, Or D/D = I/D = 0, I/I = 1) (Papadimitriou et al., 2018). A simple linear regression using the anthropometric, physical, and strength traits as the dependent variables and the genotypes of the two genetic models as independent variables were then applied, and the significance level was set at $p < 0.05$. The male and female athletes were analyzed separately.

Considering the effect of genetic structure in athletes with different geographical locations, we searched public databases

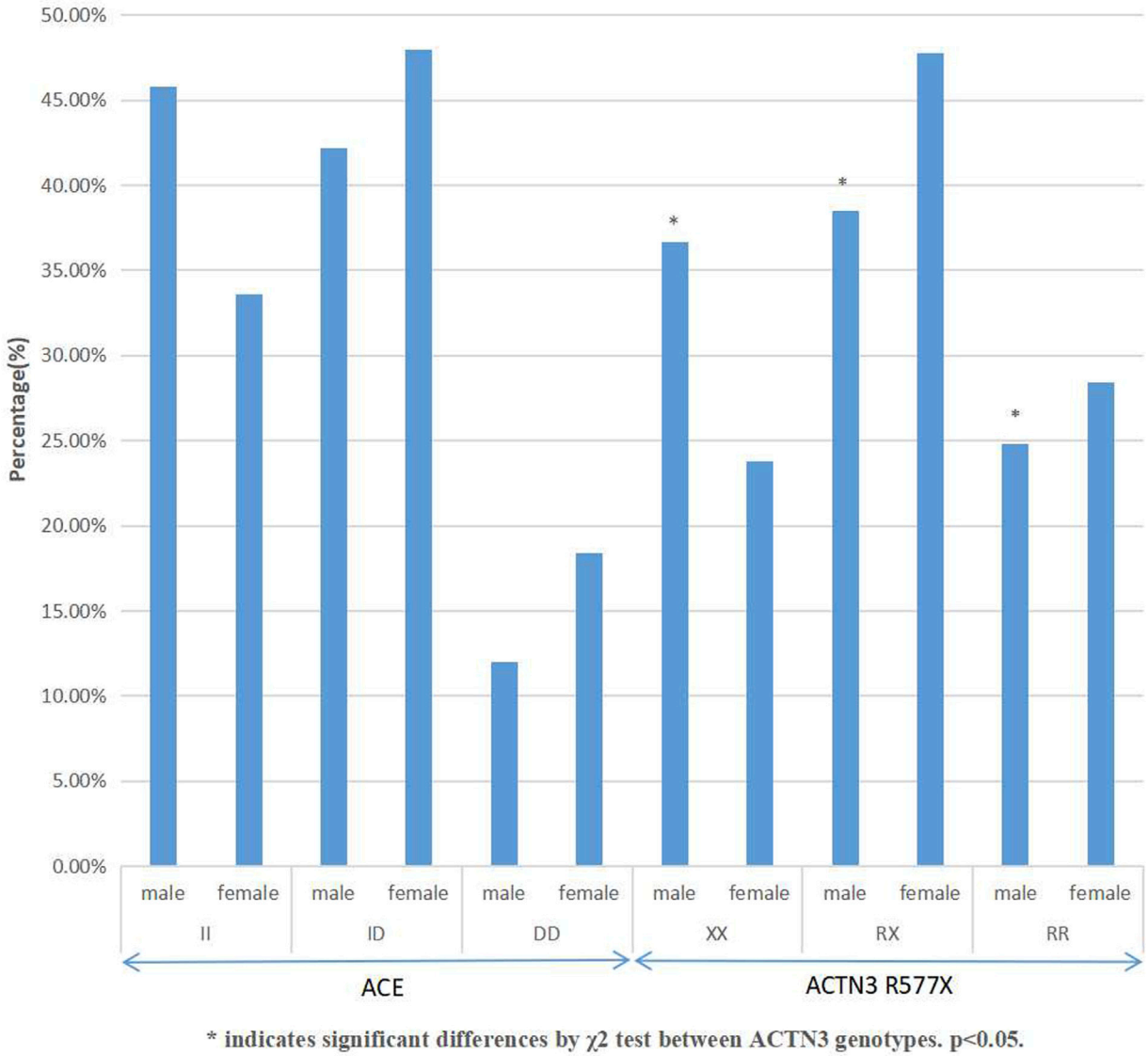

FIGURE 1 | Distribution of *ACE* and *ACTN3* gene polymorphisms in elite Chinese rowers.

from the NCBI Allele Frequency Aggregator (ALFA), Genome Aggregation database (gnomAD), 1,000 Genomes Project (1KGP), and China Metabolic Analytics Project (ChinaMAP) (Cao et al., 2020) for the different population gene allele frequency data for *ACE* I/D (rs1799752) and *ACTN3* R577X (rs1815739) loci.

RESULTS

Distributions of the *ACE* I/D Polymorphism and the *ACTN3* R577X Polymorphism in Rowing Athletes

The genotype and allele frequency of the *ACE* I/D and the *ACTN3* R577X polymorphisms of rowing athletes in this study were tested by the HWE and χ^2 test, indicating that the subjects selected in this study were representative of the population ($p > 0.05$). In **Figure 1**, the frequency of the *ACE* I allele of male athletes was 66.9%, while the frequency of the D allele was 33.1%. The genotypes of male athletes were 45.8% II, 42.2% ID, and 12% DD. There was no significant difference in the distribution of the three genotypes among male athletes ($\chi^2 = 0.23$, df = 2, $p = 0.63 > 0.05$). The *ACE* I allele frequency of female athletes was 57.5%, and the D allele frequency of female athletes was 42.5%. The genotypes of female athletes were 33.6% II, 48% ID, and 18.4% DD, with no significant differences among the three genotypes ($\chi^2 = 0.07$, df = 2, $p = 0.79 > 0.05$).

The frequency of the *ACTN3* 577X allele of male athletes was 55.9%, while the frequency of the *ACTN3* R577 allele was 44.1%. The genotypes of male athletes were 36.7% XX, 38.5% RX, and 24.8% RR. There was a significant difference in

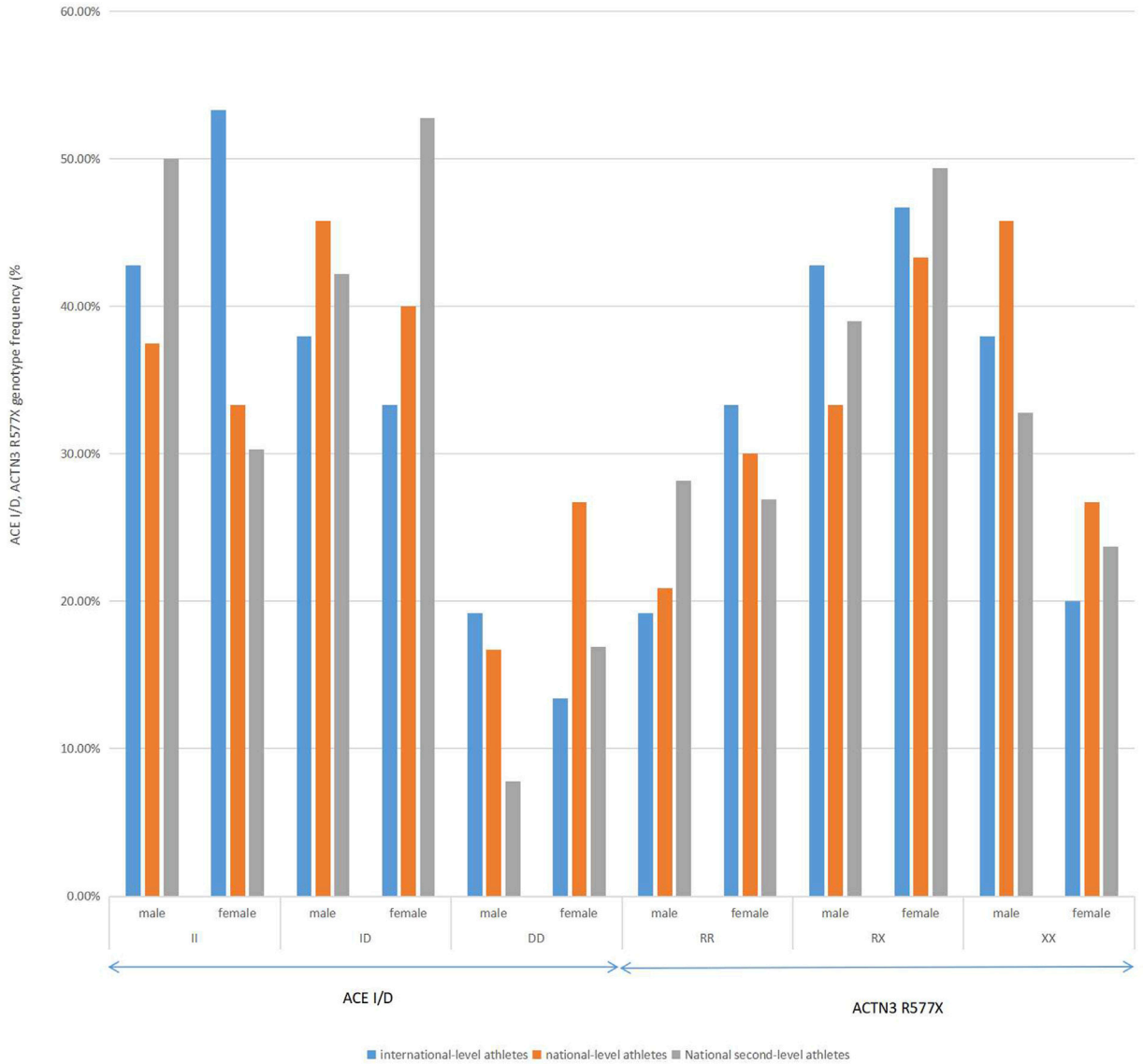

FIGURE 2 | Distribution of *ACTN3* gene polymorphism in male international-level athletes, national-level athletes, and national second-level athletes of rowing in China.

the distribution of the three genotypes among male athletes (χ^2 = 5.191, df = 2, p = 0.022 < 0.05). The 577X allele frequency of female athletes was 47.8%, and the 577 R allele frequency of female athletes was 52.2%. The genotypes of female athletes were 23.8% XX, 47.8% RX, and 28.4% RR, with insignificant differences among the three genotypes (χ^2 = 0.24, df = 2, p = 0.619 > 0.05).

Subgroup comparisons among different levels of rowing athletes. including international-level athletes, national-level athletes, and national second-level athletes, were performed by gender in **Figure 2**. There were no significant differences in the distributions of the *ACE* genotype among the subgroups of male and female athletes (χ^2 = 0.532, df = 2, p = 0.323 > 0.05; χ^2 = 0.583, df = 2, p = 0.317 > 0.05). There was a significant difference among the three genotypes of *ACTN3* R577X in the subgroup of male athletes (χ^2 = 2, df = 2, p = 0.007 < 0.01). There were no significant differences among the three genotypes of *ACTN3* R577X in the subgroup of female athletes (χ^2 = 0.471, df = 2, p = 0.411 > 0.05).

In **Figure 3**, the frequency of the *ACTN3* 577X allele among all athletes in this study was 51.44%, which was higher than the allele frequency among the total Chinese population in ChinaMAP, the East Asian, Han Chinese in Beijing (CHB), and Southern Han Chinese (CHS) populations from 1KGP, and the East Asian populations from gnomAD and NCBI ALFA. The frequency of the *ACTN3* R577 allele among all athletes in this study was 48.56%.

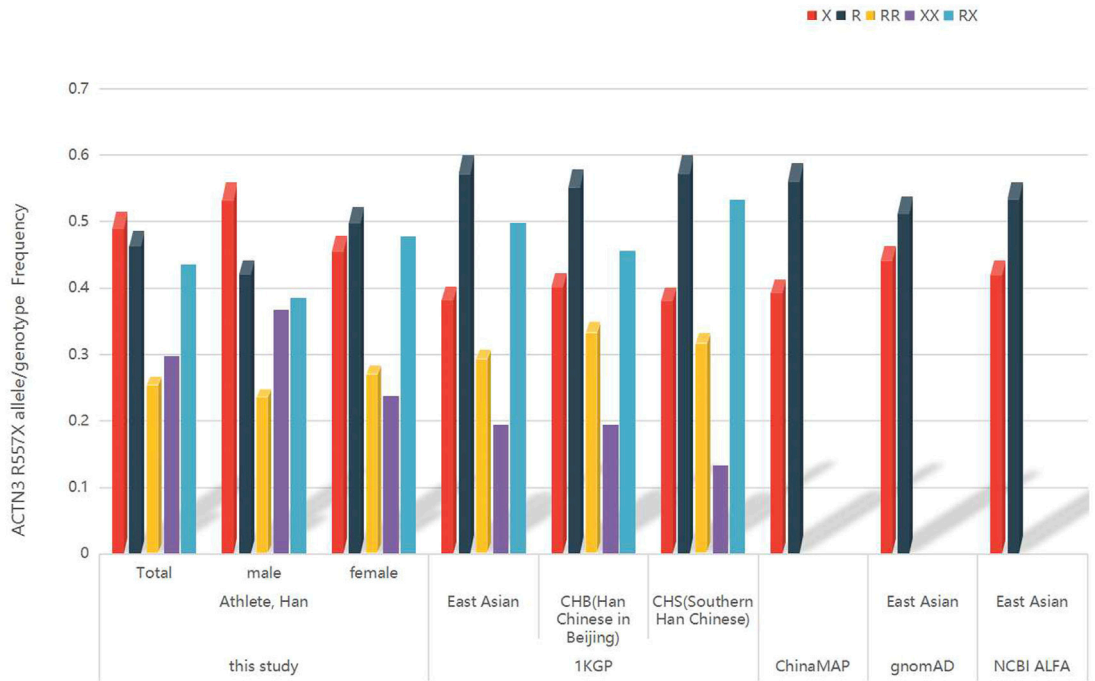

FIGURE 3 | The frequency of *ACTN3* R577X allele/genotype in rowing athletes in China, Chinese population from the ChinaMAP, East Asian, Han Chinese in Beijing, and Southern Han Chinese populations from 1KGP, East Asian population from gnomAD, and NCBI ALFA.

Analysis of *ACE*/*ACTN3* Polymorphisms and Trait Variables of Rowers

The anthropometric, physical, and strength trait variables related to rowing sports, according to the *ACE* and *ACTN3* genotype and distribution, are shown in **Table 2** and **Table 3**. There was a significant difference between weight and *ACE* genotypes, FVC/body weight, and *ACTN3* genotypes in male athletes by ANOVA, respectively ($p < 0.05$). The linear regression analysis of FVC and FVC/body weight as different dependent variables and *ACTN3* genotypes as independent variables based on the 577X allele recessive genetic effect was statistically significant in male athletes ($p < 0.05$).

ANOVA revealed a significant difference between PBP and *ACE* genotypes in female athletes ($p < 0.01$). The linear regression analysis of the PBP and the LBS as different dependent variables and *ACE* genotypes as independent variables based on the *ACE* I allele additive genetic effect showed a statistical significance in female athletes ($p < 0.05$). The trait variables of female athletes and *ACTN3* genotypes had no statistical significance in the linear regression analysis ($p > 0.05$).

DISCUSSION

Olympic rowing is a typical strength/power endurance sport in which physical fitness, strength, rowing techniques, and tactics influence rower success. The anthropometric length or breadth of the human body is almost entirely genetically decided and can hardly be altered within the range of training periodization (Almeida-Neto et al., 2020). Numerous studies found that 2,000-m rowing ergometer performance was predicted by body mass, VO_2max (Almeida-Neto et al., 2020), age, height, weight, and body fat percentage (Majumdar et al., 2017).

Through twin studies, genetic factors have been found to have an important impact on muscle strength, flexibility, and balance. The heritability of standing long jump is 62%, grip strength 63%, balance 35%, and flexibility 50% (Schutte et al., 2016). Many SNPs have been found to be related to the physiological characteristics of elite athletes, such as muscle power, speed, and aerobic capacity. Height and body mass are highly heritable and contribute to performance.

Studies have consistently provided more associations between genotype II and endurance capability (Ma et al., 2013). The *ACE* I allele is highly expressed in British climbers (Montgomery et al., 1998), Australian athletes (Gayagay et al., 1998), Polish male rowers (Jastrzębski et al., 2014), Russian rowing athletes (Ahmetov et al., 2008), Chinese female soccer athletes (Qi, 2021), and Tunisian athletes (Znazen et al., 2015). Another study found that the proportion of endurance athletes carrying the *ACE* I allele was not significant (Papadimitriou et al., 2018). Aerobic endurance refers to the ability of the body to maintain aerobic exercise. VO_2max is a high heritability index for quantifying aerobic endurance. The heritability of VO_2max can reach 80–90%, and it can be increased by 20–25% through exercise training. The cardiovascular system of the body delivers oxygen to the muscles during exercise so that the muscles can use oxygen to exercise. ACE promotes the synthesis of aldosterone and the degradation of

TABLE 2 | The ACE I/D genotype and distribution in rowing athletes showing the related anthropometric, physical and strength-trait variables.

	ACE (male)						ACE (female)					
	Genotypes			ANOVA	The linear regression		Genotypes			ANOVA	The linear regression	
	II	ID	DD	*p*	Additive	Recessive	II	ID	DD	*p*	Additive	Recessive
	45.8% (N=50)	42.2% (N=46)	12% (N=13)		(DD= 0, ID=1, II =2)	(DD= ID=0, II =1)	33.6% (N=45)	48% (N=64)	18.4% (N=25)		(DD= 0, ID=1, II =2)	(DD= ID=0, II =1)
Height, cm	191.61±3.29	189.08±3.75	193±1.22	0.059	0.311	0.992	177.58±6.88	176.16±6.46	173.2±3.23	0.671	0.37	0.509
Weight, kg	84.81±10.87	76.74±11.21	90.84±5.94	0.040*	0.206	0.842	65.52±7.43	69.22±10.91	62.46±9.28	0.58	0.8	0.755
Arm span, cm	195.6±2.12	189.67±3.21	191.82±1.58	0.067	0.359	0.888	180.33±6.74	177.78±5.01	173.2±3.95	0.404	0.173	0.334
Biacromial breadth, cm	42.67±1.21	42±1	42.8±2.04	0.766	0.911	0.841	39.16±1.89	39.75±1.54	38.5±1.32	0.551	0.617	0.879
VO_2max, L min^{-1} kg^{-1}	4.81±0.58	4.21±0.1	5.17±0.34	0.428	0.639	0.438	3.05±0.14	3.68±0.15	3.22±0.8	0.215	0.692	0.214
FVC, ml	6012±983	5265.5±1117	5728±751	0.609	0.872	0.745	3733±394	4083±643	3673±342	0.577	0.973	0.549
FVC/body weight, ml kg^{-1}	68.83±9.72	60.25±24.39	63.85±5.67	0.598	0.523	0.351	59.6±8.62	61.05±5.65	63.6±8.33	0.777	0.474	0.619
PBP, kg	84±9.69	86±3.6	84±5.47	0.92	0.986	0.85	76.6±14.43	69±12.75	50±18	0.115	0.043#	0.222
LBS, kg	118±11.04	128.33±2.88	124±8.94	0.296	0.293	0.145	113.3±11.54	111.16±14.4	68.3±24.6	0.0092&	0.018#	0.345
3,000-m run time,s	637.17±34.43	576.67±5.77	526.67±59.35	0.068	0.745	0.232	742.3±53.46	739.16±34.98	741.67±73.22	0.995	0.987	0.965

*Notes: * indicates that a significant difference between the trait variables and ACE three genotypes in male athletes (p < 0.05); & indicates that a significant difference between the trait variables and ACE three genotypes in female athletes (p < 0.05); # indicates that the linear regression analysis of trait variables and ACE genotypes based on ACE I dominant genetic effect has very significant statistical significance in female athletes (p < 0.05).*

TABLE 3 | The ACTN3 R577X genotype and distribution in rowing athletes showing the related anthropometric, physical and strength-trait variables.

	ACTN3 (male)						ACTN3 (female)					
	Genotypes			ANOVA	The linear regression		Genotypes			ANOVA	The linear regression	
	RR 24.8% (N=27)	RX 38.5% (N=42)	XX 36.7% (N=40)	*p*	Additive (RR= 0, RX=1,XX=2)	Recessive (RR=RX=0, XX =1)	RR 28.4% (N=38)	RX 47.8% (N=64)	XX 23.8% (N=32)	*p*	Additive (RR= 0, RX=1,XX=2)	Recessive (RR=RX=0, XX =1)
Height, cm	189.63±4.89	192±2.89	191.83±2.59	0.494	0.425	0.232	172.46±2.92	177.56±6.17	175.22±6.15	0.41	0.698	0.716
Weight, kg	75.8±12.27	86.9±9.39	87±10.7	0.226	0.188	0.081	61.06±6.8	70.28±9.76	63.52±8.45	0.269	0.884	0.455
Arm span, cm	191.55±0.49	193.1±2.7	190.3±3.84	0.436	0.566	0.803	176±6.4	177.1±5.85	178.7 ±6.4	0.892	0.622	0.642
Biacromial breadth, cm	42.5±0.7	42.57±1.71	42±1.17	0.7787	0.658	0.944	39.12±1.54	39.75±1.54	38.75±2.47	0.805	0.915	0.804
VO_2max, L min^{-1} kg^{-1}	4.21±0.58	5.02±0.47	4.86±0.6	0.372	0.906	0.234	3.05±0.14	3.43±0.63	3.63±0.21	0.436	0.193	0.214
FVC, ml	4585.5±232	5786±691	5538±1015	0.149	0.092	0.049%	3726±327	4083±643	3099±342	0.135	0.819	0.54
FVC/body weight, ml kg^{-1}	44.4±27.7	68.4±9.56	65.3±6.17	0.047%	0.113	0.012€	61.5±7.61	61.05±5.65	58.4±1.13	0.776	0.778	0.919
PBP, kg	86.5±4.94	87.71±6.44	79±5.47	0.079	0.071	0.666	71.5±15.67	60.5±12.75	72.5±17.67	0.543	0.839	0.459
LBS, kg	127.5±3.53	126.14±7.58	115±10	0.087	0.045ˆ	0.432	108±14.23	92.5±32.2	112.5±10.6	0.53	0.937	0.517
3,000-m run, s	575±7.07	637±52.2	516.5±25.5	0.095	0.669	0.08	746±43.84	746.67±37.67	728.75±65.13	0.843	0.614	0.548

Notes: % indicates that a significant difference between the trait variables and ACTN3 three genotypes in male athletes (p < 0.05);ˆindicates that the linear regression analysis of trait variables and ACTN3 genotypes based on ACTN3 X dominant genetic effect has statistical significance in male athletes (p < 0.05); € indicates that the linear regression analysis of trait variables and ACTN3 genotypes based on ACTN3 R dominant genetic effect has statistical significance in male athletes (p < 0.05).

angiodilators through the synthesis of angiotensin II and plays a tonic regulatory role in circulatory homeostasis. In this study, the athletes were all of Han nationality, and there were no differences between *ACE* genotypes. The proportion of individuals carrying the *ACE* I allele was higher than that of individuals carrying the D allele. The majority of male athletes had the II genotype, and weight was significantly different among the three *ACE* genotypes. We constructed two genetic models to calculate the *ACE* I allele genetic effects on trait variables of the linear regression analysis. The weight and other trait variables of the male athletes had no relationship with the *ACE* genotype ($p > 0.05$) in the linear regression, which suggested that the *ACE* I allele may have no genetic effect on weight in either homozygotes or heterozygotes.

Research suggests that mononuclear cells and ACE enzyme activity of the heart of the II genotype, compared with the DD genotype, can strengthen the myocardial contraction and cardiac output and by adjusting the level of bradykinin substrates and affect the growth of skeletal muscle energy metabolism. A high expression of the *ACE* I allele can enhance muscle oxygen absorption (Tsianos et al., 2004). Excessive body fat may reduce human body oxygen consumption and affect aerobic capacity. The *ACE* gene was found to be associated with slow-twitch type I muscle fiber (Papadimitriou et al., 2016). Higher oxygen availability and nutrient delivery for muscle fibers in contraction decrease the ACE serum levels and activity (Gunel et al., 2014). The majority of female athletes in this study had the ID genotype, with LBS revealing a very significant difference between *ACE* genotypes by ANOVA ($p < 0.01$). With the *ACE* I allele additive genetic model, it was further found that the PBP and LBS of female athletes showed significant linear regression with the *ACE* genotype ($p < 0.05$ and $p < 0.01$, respectively). PBP and LBS were both associated with muscle strength. The Caucasus power projects excellent athletic sports ability associated with the *ACE* D allele (Kim et al., 2010; Scott et al., 2010). The *ACE* D allele may improve blood ACE activity and the content of angiotensin II to transfer higher muscle strength (Charbonneau et al., 2008). Based on the representation of the *ACE* D allele of female athletes in this study, it is speculated that the effect of DD homozygotes on the muscle type of female athletes is more obvious.

The highest prevalence of the *ACTN3* gene polymorphism was the heterozygous RX genotype in male and female rowing athletes. Male rowing athletes in this study mainly carried the 577X allele, and the proportion of male athletes was 38.5% RX, which was significantly different from 36.7% XX and 24.8% RR among the groups ($p < 0.05$). The same finding was found in Australian endurance athletes (Yang et al., 2003) and Israeli top-level long-distance runners (Jastrzebski et al., 2014). The 577X allele with a high proportion of slow muscle fibers is associated with endurance events (Papadimitriou et al., 2018). In particular, studies have reported that the 577X allele is underrepresented in Russian male rowers (Ahmetov et al., 2008; Cięszczyk et al., 2011), Chinese male endurance athletes (Shang et al., 2010), and Polish rowers (Cieszczyka et al., 2012; Jastrzebski et al., 2014). There was a significant difference among the three genotypes of *ACTN3* R577X in the subgroup of male athletes. The *ACTN3* XX genotype frequency in Chinese female endurance athletes is significantly linearly increasing among average, sub-elite, elite, and highly elite athletes (Shang et al., 2010). On the contrary (Cieszczyka Pawel et al., 2012), analyzed that the genotype distribution and allele frequency between the elite and non-elite were not significantly different ($p = 0.82$ and $p = 0.56$, respectively).

The athletes in this study were recruited from the Han Chinese population in the northern and central provinces of China. The Han Chinese, an ethnic group native to China, is the largest ethnic group in the world, distributed in East Asia, Southeast Asia, and other parts of the world (Chen et al., 2019). Paleolithic ancient Tianyuan DNA (Mao et al., 2021) showed the paleolithic–modern genomic continuity in East Asia. In recent years, the Simons Genome Project, 1KGP, Human Genome Diversity Project, and HapMap Project have been conducted to study the genetic diversity and population structure of East Asians (Li et al., 2008; Consortium et al., 2009; Sudmant et al., 2015; Mallick et al., 2016). Genetic admixture with local ethnic groups and substantial genetic diversity within Han Chinese have been reported in previous studies. He *et al.* genotyped 36 Tai-Kadai-speaking Qiongzhong Hlai and 48 Haikou Han individuals at 497,637 SNPs, which revealed that East Asian populations are characterized by a north–south genetic cline (He et al., 2020). Tujia people and central Han Chinese suggested a genetic admixture under language borrowing (He G. L. et al., 2021). Genetic studies based on higher-resolution, genome-wide autosomal and uniparental Y-chromosome and mitochondrial deoxyribonucleic acid SNP data from 599 Northwest Han (Gansu Province) individuals showed increased genetic homogeneity in northwest Han individuals relative to the Mongolian/Turkic/Tungus and Tibetan–Burmese populations in the north (Yao et al., 2021). The 986 previous genome-wide analyses from southernmost, central, and northern modern Han Chinese are consistent with the primary ancestry of modern southeastern coastal Han Chinese originating from northern China (He G. et al., 2021; He G. G. et al., 2021). There was a genetic substructure in Shaanxi Han in terms of north–south-related ancestry corresponding well to latitudes (He G. L. et al., 2021) and great genetic differentiation compared to Guizhou Han (Wang et al., 2021). The ChinaMAP has established a large-scale resource of 10,588 individual deep whole-genome sequencing data and the genetic bases of metabolic traits for the genetic study of East Asians.

Considering the effect of genetic structure with different geographical locations, we searched the *ACTN3* R577X frequency of the East Asian population from public databases. The *ACTN3* 577X allele frequency of all athletes was higher than that of other populations. Furthermore, the frequency of XX homozygote among all athletes was higher than that of East Asian, CHB, and CHS populations from the 1KGP and revealed the endurance ability cline of the athlete population. We speculated that the differences between the athlete population and those described above are because the athlete population was selected based on their athletic performance and influenced by the geographic/genetic diversity of northern and central China. The fine genetic structure of the athletes will be considered in the design of further studies.

Through ANOVA, we found that FVC/body weight was significantly different among the *ACTN3* R577X genotypes ($p < 0.05$), and FVC and FVC/body weight had a statistically significant linear regression with the *ACTN3* genotype based on

the 577X allele recessive genetic model ($p < 0.01$). It was speculated that, for male athletes, RR and XX homozygosity had a genetic influence on FVC and FVC/body weight, respectively. α-Actinin-3 is encoded by the *ACTN3* gene, and its expression in glycolytic skeletal muscle contributes to enhancing muscle function and coordinating fast-twitch muscles (Yang et al., 2003). Ortiz et al. (2022) reported that Colombian athletes carrying the RX genotype might express more ACTN3 protein that is involved in the optimization of muscle contraction in fast-twitch fiber. The frequency of the RR genotype was revealed to be higher in male skiers than in the control group, but male skiers with the XX genotype evidently had an increased VO_2 peak within 5 years (Magi et al., 2016). Pimenta *et al.* (2013) found that VO_2max in males with the XX genotype was higher than that in males with the RR genotype. In mice with an *ACTN3* gene knockout model, the loss of ACTN3 protein induces the transformation of skeletal muscle fast-twitch fibers to oxidative metabolism, and the increase in actin-2 levels can work as a compensatory mechanism for the loss of ACTN3 protein, which is conducive to endurance performance (Seto et al., 2013).

There were no significant differences among the female groups, with genotype frequencies of 47.8% RX, 23.8% XX, and 28.4% RR ($p > 0.05$). On the other hand, the female rowing athletes in this study mainly carried the 577 R allele, and previous research indicated that the frequency of the 577 R allele on top-level Polish rowers was higher (Jastrzębski et al., 2014). Shang *et al.* (2010) reported that the 577X allele of Chinese endurance female athletes was overrepresented and significantly different compared with that of the controls (51.3 *vs* 41.1%, $p = 0.019$). The 577 R allele was significantly higher in elite speed and power athletes than in the control group (Cięszczyk et al., 2011; Papadimitriou et al., 2016; Weyerstraß et al., 2018). The reason may be that rowing is not strictly an endurance sport with a mixed character of endurance, isokinetic strength, and power (Ahmetov et al., 2010). The 577 R allele has been demonstrated to be related to muscle mass (Galeandro et al., 2017). A meta-analysis has clearly summarized the associations between the RX and RR genotypes and the 577 R allele with power-oriented performance (Stucky-Byler et al., 2018; Lammi et al., 2019). Furthermore, the RR genotype might contribute to generating powerful and forceful muscle contractions (Del Coso et al., 2019). However, in our study, the trait variables of female athletes with *ACTN3* genotypes had no significant linear regression ($p > 0.05$). A previous study found that there were significant differences in the peak heart rate of the RR genotype and RX genotype in the male group but no significant differences in the female group, and high values of all endurance indexes appeared in the female group with the XX homozygous genotype (Deschamps et al., 2015). Similarly, in this study, the 577X allele was not significantly associated with any traits of female athletes, consistent with the results of excellent endurance-oriented athletes such as Italian rowers (Paparini et al., 2007), Russian rowers (Weyerstraß et al., 2018), and Hungarian rowers (Bosnyák et al., 2015). This could be explained by the fact that rowing is a complex discipline with a vigorous power-start demanding starting anaerobic capacity, an immediately high aerobic steady state, and an extremely exhausting spurt.

CONCLUSION

As conserved genes, *ACE* and *ACTN3* polymorphisms have been reported in endurance, explosive power, sensitivity, sports injury training, and other related studies, and the results are consistent or inconsistent with each other, whether in outstanding athletes or in the normal population. Athletes are people with outstanding performance phenotypes. The sports ability phenotype is a complex phenotypic trait that is not regulated by a single gene polymorphism; therefore, good rowing athletes also do not have a single *ACE* or *ACTN3* genotype, which may be advantageous. These results suggest that the *ACE* I allele and XX genotype might genetically affect the endurance traits of male athletes, with the RR genotype being a power trait. The *ACE* D allele might genetically affect the strength traits of female athletes. Therefore, the influence of genes on the performance of movement is extensive, multifactorial, and universal. However, this is the first study to compare the gene contribution with the quantitative traits of rowing athletes. This study is also limited by the fine genetic structure of the Chinese Han population, the sample size of the athletes, the discipline of rowing, regional differences, sports performance-related traits, and other factors. These limitations will be gradually addressed in our future research.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Human Ethics Committee of Hubei Institute of Sport and Science. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

QW collected and analyzed the data during the experiment, performed the statistical analysis, and wrote the manuscript.

FUNDING

This paper was funded by the Key Laboratory Project of the General Administration of Sport of China (2015KYA004).

ACKNOWLEDGMENTS

Wuhan Gene Create Biological Engineering Co. was responsible for the genotyping of the volunteers and athletes.

REFERENCES

Ahmetov, I. I., Druzhevskaya, A. M., Astratenkova, I. V., Popov, D. V., Vinogradova, O. L., and Rogozkin, V. A. (2010). The *ACTN3* R577X Polymorphism in Russian Endurance Athletes. *Br. J. Sports Med.* 44 (9), 649–652. doi:10.1136/bjsm.2008.051540

Ahmetov, I. I., Popov, D. V., Astratenkova, I. V., Druzhevskaya, A. M., Missina, S. S., Vinogradova, O. L., et al. (2008). The Use of Molecular Genetic Methods for Prognosis of Aerobic and Anaerobic Performance in Athletes. *Hum. Physiol.* 34, 338–342. doi:10.1134/s0362119708030110

Almeida-Neto, P. F. d., Silva, L. F. d., Matos, D. G. D., Jeffreys, I., Cesário, T. d. M., Neto, R. B., et al. (2020). Equation for Analyzing the Peak Power in Aquatic Environment: An Alternative for Olympic Rowing Athletes. *PLoS ONE* 15 (12), e0243157. doi:10.1371/journal.pone.0243157

Bosnyák, E., Trájer, E., Udvardy, A., Komka, Z., Protzner, A., Kováts, T., et al. (2015). *ACE* and *ACTN3* Genes Polymorphisms Among Female Hungarian Athletes in the Aspect of Sport Disciplines. *Acta Physiol. Hungarica* 102 (4), 451–458. doi:10.1556/036.102.2015.4.12

Bouchard, C. (2011). Overcoming Barriers to Progress in Exercise Genomics. *Exerc. Sports Rev.* 39 (4), 212–217. doi:10.1097/jes.0b013e31822643f6

Boulygina, E. A., Borisov, O. V., Valeeva, E. V., Semenova, E. A., Kostryukova, E. S., Kulemin, N. A., et al. (2020). Whole Genome Sequencing of Elite Athletes. *bs* 37 (3), 295–304. doi:10.5114/biolsport.2020.96272

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

Charbonneau, D. E., Hanson, E. D., Ludlow, A. T., Delmonico, M. J., and B., F.S. M. (2008). *ACE* Genotype and the Muscle Hypertrophic and Strength Responses to Strength Training. *Med. Sci. Sports Exerc.* 40 (4), 677–683. doi:10.1249/MSS.0b013e318161eab9

Chen, P., Wu, J., Luo, L., Gao, H., Wang, M., Zou, X., et al. (2019). Population Genetic Analysis of Modern and Ancient DNA Variations Yields New Insights into the Formation, Genetic Structure, and Phylogenetic Relationship of Northern Han Chinese. *Front. Genet.* 10:1045. doi:10.3389/fgene.2019.01045

Cięszczyk, P., Eider, J., Ostanek, M., A., A.S., et al. (2011). Association of the *ACTN3* R577X Polymorphism in Polish Power-Orientated Athletes. *J. Hum. Kinet* 28, 55–61. doi:10.2478/v10078-011-0022-0

Cieszczyk, P., Sawczuk, M., Maciejewska-Karlowska, A., and Ficek, K. (2012). *ACTN3* R577X Polymorphism in Top-Level Polish Rowers. *J. Exerc. Sci. Fitness.* 10, 12–15, 2012.doi:10.1016/j.jesf.2012.04.003

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

Del Coso, J., Hiam, D., Houweling, P., Pérez, L. M., Eynon, N., and Lucía, A. (2019). More Than a 'speed Gene': ACTN3 R577X Genotype, Trainability, Muscle Damage, and the Risk for Injuries. *Eur. J. Appl. Physiol.* 119, 49–60. doi:10.1007/s00421-018-4010-0

Deschamps, C. L., Connors, K. E., Klein, M. S., Johnsen, V. L., Shearer, J., Vogel, H. J., et al. (2015). The *ACTN3* R577X Polymorphism Is Associated with Cardiometabolic Fitness in Healthy Young Adults. *PLoS ONE* 10 (6), e0130644. doi:10.1371/journal.pone.0130644

Eider, J., Cieszczyk, P., Ficek, K., Leonska-Duniec, A., Sawczuk, M., Maciejewska-Karlowska, A., et al. (2013). The Association between D Allele of the ACE Gene and Power Performance in Polish Elite Athletes. *Sci. Sports* 28 (6), 325–330. doi:10.1016/j.scispo.2012.11.005

Elks, C. E., Den, Hoed, M., Zhao, J. H., Sharp, S. J., Wareham, N. J., Loos, R. J. F., et al. (2012). Variability in the Heritability of Body Mass index: A Systematic Review and Meta-Regression. *Front. Endocrin.* 3, 29. doi:10.3389/fendo.2012.00029

Eynon, N., Duarte, J. A., Oliveira, J., Sagiv, M., Yamin, C., Meckel, Y., et al. (2009). ACTN3 R577X Polymorphism and Israeli Top-Level Athletes. *Int. J. Sports Med.* 30 (09), 695–698. doi:10.1055/s-0029-1220731

Galeandro, V., Notarnicola, A., Bianco, A., Tafuri, S., Russo, L., Pesce, V., et al. (2017). ACTN3/ACE Genotypes and Mitochondrial Genome in Professional Soccer Players' Performance. *J. Biol. Regul. Homeost Agents* 31 (1), 207–213.

Gayagay, G., Yu, B., Hambly, B., Boston, T., Hahn, A., Celermajer, D. S., et al. (1998). Elite Endurance Athletes and the ACE I Allele - the Role of Genes in Athletic Performance. *Hum. Genet.* 103, 48–50. doi:10.1007/s004390050781

Gunel, T., Gumusoglu, E., Hosseini, M. K., Yilmazyildirim, E., Dolekcap, I., and Aydinli, K. (2014). Effect of Angiotensin I-Converting Enzyme and α-actinin-3 Gene Polymorphisms on Sport Performance. *Mol. Med. Rep.* 9 (4), 1422–1426. doi:10.3892/mmr.2014.1974

He, G. G., Zhang, Y., Wei, L.-H., Wang, M., Yang, X., Guo, J., et al. (2021). The Genomic Formation of Tanka People, an Isolated 'Gypsies in Water' in the Coastal Region of Southeast China. *bioRxiv* 07 (18), 452867. doi:10.1101/2021.07.18.452867

He, G. L., Li, Y. X., Wang, M. G., Zou, X., Yeh, H. Y., Yang, X. M., et al. (2021). 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., Wang, M. G., Li, Y. X., Zou, X., Yeh, H. Y., Tang, R. K., et al. (2021). Fine-scale north-to-south Genetic Admixture Profile in Shaanxi Han Chinese Revealed by Genome-wide Demographic History Reconstruction. *J. Syst. Evol.* 00 (0), 1–18. doi:10.1111/jse.12715

He, G., Wang, M., Zou, X., Tang, R., Yeh, H.-Y., Wang, Z., et al. (2021). Genomic Insights into the Differentiated Population Admixture Structure and Demographic History of North East Asians. *bioRxiv* 07 (19), 452943. doi:10.1101/2021.07.19.452943

He, G., Wang, Z., Guo, J., Wang, M., Zou, X., Tang, R., et al. (2020). Inferring the Population History of Tai-Kadai-Speaking People and Southernmost Han Chinese on Hainan Island by Genome-wide Array Genotyping. *Eur. J. Hum. Genet.* 28, 1111–1123. doi:10.1038/s41431-020-0599-7

Izquierdo-gabarren, M., González De Txabarri Expósito, R., García-pallarés, J., Sánchez-medina, L., De Villarreal, E. S. S., and Izquierdo, M. (2010). Concurrent Endurance and Strength Training Not to Failure Optimizes Performance Gains. *Med. Sci. Sports Exerc.* 42 (6), 1191–1199. doi:10.1249/mss.0b013e3181c67eec

Jacob, Y., Spiteri, T., Hart, N., and Anderton, R. (2018). The Potential Role of Genetic Markers in Talent Identification and Athlete Assessment in Elite Sport. *Sports* 6, 88. doi:10.3390/sports6030088

Jastrzebski, Z., Leonska-Duniec, A., Kolbowicz, M., and Tomiak, T. (2014). Association of the ACTN3 R577X Polymorphism in Polish Rowers. *Baltic J. Health Phys. Activity* 6 (3), 205–210. doi:10.2478/bjha-2014-0019

Jastrzębski, Z., Leońska-Duniec, A., Kolbowicz, M., and Tomiak, T. (2014). The Angiotensin Converting Enzyme Gene I/D Polymorphism in Polish Rowers. *Cent. Eur. J. Sport Sci. Med.* 5 (1), 77–82.

Keenan, K. G., Senefeld, J. W., and Hunter, S. K. (2018). Girls in the Boat: Sex Differences in Rowing Performance and Participation. *PLoS One* 13 (1), e0191504. doi:10.1371/journal.pone.0191504

Kim, C. H., Cho, J. Y., Jeon, J. Y., Koh, Y. G., Kim, Y. M., Kim, H. J., et al. (2010). *ACE* DD Genotype Is Unfavorable to Korean Short-Term Muscle Power Athletes. *Int. J. Sports Med.* 31 (1), 65–71. doi:10.1055/s-0029-1239523

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., et al. (2008). Worldwide Human Relationships Inferred from Genome-wide Patterns of Variation. *Science* 319, 1100–1104. doi:10.1126/science.1153717

Little, J., Higgins, J. P. T., Ioannidis, J. P. A., Moher, D., Gagnon, F., von Elm, E., et al. (2009). Strengthening the Reporting of Genetic Association Studies (STREGA)-an Extension of the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement. *J. Clin. Epidemiol.* 62 (6), 597–608. doi:10.1016/j.jclinepi.2008.12.004

Ma, F., Yang, Y., Li, X., Zhou, F., Gao, C., Li, M., et al. (2013). The Association of Sport Performance with ACE and ACTN3 Genetic Polymorphisms: A Systematic Review and Meta-Analysis. *PLoS One* 8, e54685. doi:10.1371/journal.pone.0054685

Maciejewska-Skrendo, A., Sawczuk, M., Cięszczyk, P., and Ahmetov, I. I. (2019). "Genes and Power Athlete Status," in *Sports, Exercise, and Nutritional Genomics: Current Status and Future Directions*. Editors D Barh and I Ahmetov (London, United Kingdom: Academic Press), 41–72. doi:10.1016/b978-0-12-816193-7.00003-8

Maciejewski, H., Rahmani, A., Chorin, F., Lardy, J., Samozino, P., and Ratel, S. (2019). Methodological Considerations on the Relationship between the 1,500-m Rowing Ergometer Performance and Vertical Jump in National-Level Adolescent Rowers. *J. Strength Conditioning Res.* 33 (11), 3000–3007. doi:10.1519/JSC.0000000000002406

Mägi, A., Unt, E., Prans, E., Raus, L., Eha, J., Veraksitš, A., et al. (2016). The Association Analysis between ACE and ACTN3 Genes Polymorphisms and Endurance Capacity in Young Cross-Country Skiers: Longitudinal Study. *J. Sports Sci. Med.* 15, 287–294. eCollection 2016

Majumdar, P., Das, A., and Mandal, M. (2017). Physical and Strength Variable Sasapredictor of 2000m Rowing Ergometer Performance in Elite Rowers. *J. Phys. Education Sport* 17 (4), 2502–2507. doi:10.7752/jpes.2017.02106

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., et al. (2016). The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations. *Nature* 538, 201–206. doi:10.1038/nature18964

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

Mikami, E., Fuku, N., Murakami, H., Tsuchie, H., Takahashi, H., Ohiwa, N., et al. (2014). *Actn3* R577X Genotype Is Associated with Sprinting in Elite Japanese Athletes. *Int. J. Sports Med.* 35, 172–177. doi:10.1055/s-0033-1347171

Miyamoto-Mikami, E., Zempo, H., Fuku, N., Kikuchi, N., Miyachi, M., and Murakami, H. (2018). Heritability Estimates of Endurance-Related Phenotypes: A Systematic Review and Meta-Analysis. *Scand. J. Med. Sci. Sports* 28, 834–845. doi:10.1111/sms.12958

Montgomery, H. E., Marshall, R., Hemingway, H., Myerson, S., Clarkson, P., Dollery, C., et al. (1998). Human Gene for Physical Performance. *Nature* 393, 221–222. doi:10.1038/30374

Ortiz, M., Ayala, A., Petro, J. L., Argothy, R., Garzón, J., and Bonilla, D. A. (2022). Evaluation of ACTN3 R577X and ACE I/D Polymorphisms in Young Colombian Athletes: An Exploratory Research. *J. Human Sport Exerc.* 17 (3). In Press. doi:10.14198/jhse.2022.173.14

Papadimitriou, I. D., Lockey, S. J., Voisin, S., Herbert, A. J., Garton, F., Houweling, P. J., et al. (2018). No Association between *ACTN3* R577X and *ACE* I/D Polymorphisms and Endurance Running Times in 698 Caucasian Athletes. *BMC Genomics* 19, 13. doi:10.1186/s12864-017-4412-0

Papadimitriou, I. D., Lucia, A., Pitsiladis, Y. P., Pushkarev, V. P., Dyatlov, D. A., Orekhov, E. F., et al. (2016). *ACTN3* R577X and *ACE* I/D Gene Variants Influence Performance in Elite Sprinters: a Multi-Cohort Study. *BMC Genomics* 17 (1), 285. doi:10.1186/s12864-016-2462-3

Papadimitriou, I., Papadopoulos, C., Kouvatsi, A., and Triantaphyllidis, C. (2008). The *ACTN3* Gene in Elite Greek Track and Field Athletes. *Int. J. Sports Med.* 29, 352–355. doi:10.1055/s-2007-965339

Paparini, A., Ripani, M., Giordano, G. D., Santoni, D., Pigozzi, F., and Romanospica, V. (2007). ACTN3 Genotyping by Real-Time PCR in the Italian Population and Athletes. *Med. Sci. Sports Exerc.* 39, 810–815. doi:10.1097/mss.0b013e3180317491

Penichet-Tomás, A., Pueo, B., and Jiménez-Olmedo, J. M. (2019). Physical Performance Indicators in Traditional Rowing Championships. *J. Sports Med. Phys. Fitness* 59 (5), 767–773. doi:10.23736/S0022-4707.18.08524-9PMID:309364177

Pescatello, L. S., Corso, L. M. L., Santos, L. P., Livingston, J., and Taylor, B. A. (2019). Angiotensin-converting Enzyme and the Genomics of Endurance Performance. *Routledge Handbook Sport Exerc. Syst. Genet.*, 216–250. doi:10.4324/9781315146287-21

Pickering, C., Kiely, J., Grgic, J., Lucia, A., and Del Coso, J. (2019). Can Genetic Testing Identify Talent for Sport. *Genes* 10, 972. doi:10.3390/genes10120972

Pimenta, E. M., Coelho, D. B., Veneroso, C. E., Barros Coelho, E. J., Cruz, I. R., Morandi, R. F., et al. (2013). Effect of *ACTN*3 Gene on Strength and Endurance in Soccer Players. *J. Strength Cond. Res.* 27, 3286–3292. doi:10.1519/jsc.0b013e3182915e66

Qi, W. (2021). The *ACE* and *ACTN3* Polymorphisms in Female Soccer Athletes. *Genes Environ.* 43, 5. doi:10.1186/s41021-021-00177-3

Riboli, A., Coratella, G., Rampichini, S., Limonta, E., and Esposito, F. (2021). Testing Protocol Affects the Velocity at VO2max in Semi-professional Soccer Players. *Res. Sports Med.* 29, 1–11. doi:10.1080/15438627.2021.1878460

Schutte, N. M., Nederend, I., Hudziak, J. J., Bartels, M., and de Geus, E. J. C. (2016). Twin-sibling Study and Meta-Analysis on the Heritability of Maximal Oxygen Consumption. *Physiol. Genomics* 48, 210–219. doi:10.1152/physiolgenomics.00117.2015

Scott, R. A., Irving, R., Irwin, L., Morrison, E., Charlton, V., Austin, K., et al. (2010). *ACTN3* and *ACE* Genotypes in Elite Jamaican and US Sprinters. *Med. Sci. Sports Exerc.* 42 (1), 107–112. doi:10.1249/mss.0b013e3181ae2bc0

Semenova, E. A., Fuku, N., and Ahmetov, I. I. (2019). "Genetic Profile of Elite Endurance Athletes," in *Sports, Exercise, and Nutritional Genomics: Current Status and Future Directions*. Editors D. Barh and I. Ahmetov (London, United Kingdom: Academic Press), 73–104. doi:10.1016/b978-0-12-816193-7.00004-x

Seto, J. T., Quinlan, K. G. R., Lek, M., Zheng, X. F., Garton, F., MacArthur, D. G., et al. (2013). *ACTN3* Genotype Influences Muscle Performance through the Regulation of Calcineurin Signaling. *J. Clin. Invest.* 123 (10), 4255–4263. doi:10.1172/JCI67691

Shang, X., Huang, C., Chang, Q., Zhang, L., and Huang, T. (2010). Association between the *ACTN3* R577X Polymorphism and Female Endurance Athletes in China. *Int. J. Sports Med.* 31 (12), 913–916. doi:10.1055/s-0030-1265176

Silventoinen, K., Sammalisto, S., Perola, M., Boomsma, D. I., Cornes, B. K., Davis, C., et al. (2003). Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight Countries. *Twin Res.* 6, 399–408. doi:10.1375/136905203770326402

Stucky-Byler, L., and Rodríguez-Buitrago, A. (2018). Análisis exploratorio de la relación entre el polimorfismo *ACTN3* R577X y el rendimiento deportivo en levantadores de pesas colombianos. Master's thesis. Bogotá, Colombia: Universidad de Ciencias Aplicadas y Ambientales U.D.C.A. Available at: https://repository.udca.edu.co/handle/11158/1056.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015). An Integrated Map of Structural Variation in 2,504 Human Genomes. *Nature* 526, 75–81. doi:10.1038/nature15394

Tharabenjasin, P., Pabalan, N., Jarjanazi, H., and Jarjanazi, H. (2019). Association of the ACTN3 R577X (Rs1815739) Polymorphism with Elite Power Sports: A Meta-Analysis. *PLoS One* 14 (5), e0217390. doi:10.1371/journal.pone.0217390

Tsianos, G., Sanders, J., Dhamrait, S., Humphries, S., Grant, S., and Montgomery, H. (2004). The *ACE* Gene Insertion/deletion Polymorphism and Elite Endurance Swimming. *Eur. J. Appl. Physiol.* 92 (3), 360–362. doi:10.1007/s00421-004-1120-7

Valeeva, E. V., Ahmetov, I. I., and Rees, T. (2019). "Psychogenetics and Sport," in *Sports, Exercise, and Nutritional Genomics: Current Status and Future Directions*. Editors D. Barh and I. I. Ahmetov (Academic Press), 147–165. doi:10.1016/b978-0-12-816193-7.00007-5

Wang, M., Yuan, D., Zou, X., Wang, Z., Yeh, H.-Y., Liu, J., et al. (2021). 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. doi:10.3389/fgene.2021.727821

Weyerstraß, J., Stewart, K., Wesselius, A., and Zeegers, M. (2018). Nine Genetic Polymorphisms Associated with Power Athlete Status - A Meta-Analysis. *J. Sci. Med. Sport* 21(2), 213–220,. doi:10.1016/j.jsams.2017.06.012

Yang, N., MacArthur, D. G., Gulbin, J. P., Hahn, A. G., Beggs, A. H., Easteal, S., et al. (2003). *ACTN3* Genotype Is Associated with Human Elite Athletic Performance. *Am. J. Hum. Genet.* 73, 627–631. doi:10.1086/377590

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

Zempo, H., Miyamoto-Mikami, E., Kikuchi, N., Fuku, N., Miyachi, M., and Murakami, H. (2017). Heritability Estimates of Muscle Strength-Related Phenotypes: a Systematic Review and Meta-Analysis. *Scand. J. Med. Sci. Sports* 27, 1537–1546. doi:10.1111/sms.12804

Znazen, H., Mejri, A., Touhami, I., Chtara, M., Siala, H., LE Gallais, D., et al. (2015). Genetic Advantageous Predisposition of Angiotensin Converting Enzyme Id Polymorphism in Tunisian Athletes. *J. Sports Med. Phys. Fitness* 56 (6), 724–730.

Conflict of Interest: The author declares 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 December 2021
doi: 10.3389/fgene.2021.735786

Genomic Insights Into the Genetic Structure and Natural Selection of Mongolians

Xiaomin Yang[1,2,3†], Sarengaowa[1†], Guanglin He[1,2,3], Jianxin Guo[1,2,3], Kongyang Zhu[1,2,3], Hao Ma[1,2,3], Jing Zhao[1,2,3], Meiqing Yang[4], Jing Chen[4], Xianpeng Zhang[5], Le Tao[1,2,3], Yilan Liu[1,2,3], Xiu-Fang Zhang[6*] and Chuan-Chao Wang[1,2,3*]*

[1]Department of Anthropology and Ethnology, Institute of Anthropology, National Institute for Data Science in Health and Medicine, School of Sociology and Anthropology, Xiamen University, Xiamen, China, [2]State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen, China, [3]State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen, China, [4]Department of Forensic Medicine, Guizhou Medical University, Guiyang, China, [5]Institute of Biological Anthropology, Jinzhou Medical University, Liaoning, China, [6]Department of Pediatrics, Xiang'an Hospital of Xiamen University, Xiamen, China

Mongolians dwell at the Eastern Eurasian Steppe, where is the agriculture and pasture interlaced area, practice pastoral subsistence strategies for generations, and have their own complex genetic formation history. There is evidence that the eastward expansion of Western Steppe herders transformed the lifestyle of post-Bronze Age Mongolia Plateau populations and brought gene flow into the gene pool of Eastern Eurasians. Here, we reported genome-wide data for 42 individuals from the Inner Mongolia Autonomous Region of North China. We observed that our studied Mongolians were structured into three distinct genetic clusters possessing different genetic affinity with previous studied Inner Mongolians and Mongols and various Eastern and Western Eurasian ancestries: two subgroups harbored dominant Eastern Eurasian ancestry from Neolithic millet farmers of Yellow River Basin; another subgroup derived Eastern Eurasian ancestry primarily from Neolithic hunter-gatherers of North Asia. Besides, three-way/four-way qpAdm admixture models revealed that both north and southern Western Eurasian ancestry related to the Western Steppe herders and Iranian farmers contributed to the genetic materials into modern Mongolians. ALDER-based admixture coefficient and haplotype-based GLOBETROTTER demonstrated that the former western ancestry detected in modern Mongolian could be recently traced back to a historic period in accordance with the historical record about the westward expansion of the Mongol empire. Furthermore, the natural selection analysis of Mongolians showed that the Major Histocompatibility Complex (MHC) region underwent significantly positive selective sweeps. The functional genes, alcohol dehydrogenase (*ADH*) and lactase persistence (*LCT*), were not identified, while the higher/lower frequencies of derived mutations were strongly correlated with the genetic affinity to East Asian/Western Eurasian populations. Our attested complex population movement and admixture in the agriculture and pasture interlaced area played an important role in the formation of modern Mongolians.

Keywords: Mongolian, genetic heterogeneity, admixture history, natural selection, functional genes

OPEN ACCESS

Edited by:
Gyaneshwer Chaubey, Banaras Hindu University, India

Reviewed by:
Min-Sheng Peng, Kunming Institute of Zoology (CAS), China
Eaaswarkhanth Muthukrishnan, New York University Abu Dhabi, United Arab Emirates

***Correspondence:**
Xiaomin Yang
xmyang36@163.com
Xiu-Fang Zhang
13287787056@wo.com
Chuan-Chao Wang
wang@xmu.edu.cn

†*These authors have contributed equally to this work*

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

Received: *03 July 2021*
Accepted: *02 November 2021*
Published: *08 December 2021*

Citation:
Yang X, Sarengaowa, He G, Guo J, Zhu K, Ma H, Zhao J, Yang M, Chen J, Zhang X, Tao L, Liu Y, Zhang X-F and Wang C-C (2021) Genomic Insights Into the Genetic Structure and Natural Selection of Mongolians. Front. Genet. 12:735786. doi: 10.3389/fgene.2021.735786

INTRODUCTION

The vast Eurasian steppe zone stretching from Hungary in the west to Mongolia and northeastern China in the east has witnessed a dynamic demographic history. Ancient DNA findings from Western Eurasian Steppe showed the massive continental-scale steppe population migrations, admixture, and turnover since the Early Bronze Age (Allentoft et al., 2015; Mathieson et al., 2015; Damgaard et al., 2018; Wang et al., 2019). Both archaeologically and genetically attested evidence also showed the Western Steppe populations migrated to the Eastern Steppe zone and had influenced the genetic makeup of the Eastern Eurasians (Damgaard et al., 2018; de Barros Damgaard et al., 2018; Narasimhan et al., 2019; Ning et al., 2019; Jeong et al., 2020; Wang et al., 2021), whose genetic structure with a west-east admixture cline of the ancestry of Ancient North Eurasian (ANE) and Ancient Northeast Asian (ANA) stretching from Botai in Central Asia to Lake Baikal, Mongolia, and Devil's Gate Cave of Eastern Eurasian has existed during the Pre-Bronze Age periods (Siska et al., 2017; de Barros Damgaard et al., 2018; Jeong et al., 2020). The Eastern Steppe has served as a crossroad for human population movements and plays a pivotal role in achieving cultural exchanges. The eastward expansions of Western Steppe populations associated with the Yamnaya (ca. 3300–2700 BCE) and Afanasievo (ca. 3300–2500 BCE) cultures in the Early and Middle Bronze Age and later ones associated with Andronovo (ca. 1800–1300 BCE) and Sintashta (ca. 2200–1700 BCE) in the Late Bronze Age not only brought related culture into the Eastern Steppe but also substantially contributed to the gene pool of the Eastern Steppe, forming the genetic heterogeneity with west-east admixture cline of Western Steppe-related ancestry. An additional genetic influx related to Central/Southern Asia populations was detected in the Early Iron Age western Mongolia ancient populations, which still exists in modern Mongolic and Turkic speaking groups (Jeong et al., 2019; Jeong et al., 2020). Subsequently, Xiongnu (209 BCE–98 CE), the first historically documented empire founded by pastoralists, received more complex gene flows in accordance with the historical records and showed highly heterogeneous populations structure, harboring different Han-related ancestry and more recent Western Steppe-related ancestry (Damgaard et al., 2018; Jeong et al., 2020; Wang et al., 2021). The Mongol empire emerged and established the largest continental empire across Asia and eastern Europe in the 13th century, controlling vast territories and trade routes, and diverse populations flowed into the steppe heartland. However, the genetic heterogeneity of the Eastern Steppe during this period was lower than that of previous nomadic regimes, with more Eastern Eurasian-related ancestry, marking the beginning of the formation of the modern Mongolians' gene pool (Jeong et al., 2020; Wang et al., 2021). Even though the Western Steppe-related ancestry fluctuated in ancient Mongolia populations, modern Mongolian groups still show some extent of affinity with Western Eurasian-related populations and show genetic structure with different proportions of the Western Eurasian-related ancestry (Bai et al., 2018; He et al., 2019; Jeong et al., 2019; Zhao et al., 2020).

Across the Eurasian Steppe, dairy is a staple food and traditional diet. At the beginning of the Bronze Age, the multi-phased introduction of pastoralism drastically changed lifeways and subsistence on the Eastern Steppe (Jeong et al., 2018; Wilkin et al., 2020). Milk consumption in Mongolia before 2500 BCE by individuals affiliated with the Afanasievo, Chemurchek (2750–1900 BCE) and the Deer Stone-Khirigsuur Complex (DSKC) cultures in Khövsgöl was confirmed by large-scale paleogenomics studies. In contrast, the whole genome analysis of ancient populations in Mongolia revealed that despite the pastoralist lifestyle with evidence of milk consumption, the absence of positive selection of lactase persistence-related gene (*LCT/MCM6*) leading to the negligibly low frequency of derived mutations conferring lactase persistence indicates that animal husbandry for livelihood was adopted in the Eastern Steppe by local hunter-gatherers instead of causing by massive populations movements and turnover in Mongolian (Jeong et al., 2018; Jeong et al., 2020).

Inner Mongolia Autonomous Region, located in northern China, adjacent to the Central Plain and the West Liao River in northeastern East Asia, and some parts of it belong to the Yellow River Basin, which is the cradle of millet farming of China, the Middle Neolithic Miaozigou Culture in Inner Mongolia showed the characteristic of northward expansion of millet farmers in the Yellow River Basin (Ning et al., 2020). Moreover, Inner Mongolia Autonomous Region has been a farming-pastoral transitional zone in East Asia since the development of agriculture in the Neolithic Age and served as a key communication point between the nomadic culture of the northern grassland and the farming culture of the Central Plain. In addition, south-north bidirectional migration and coastal route of population movement between East Asia and Siberia have impacted the observed genetic variations among modern East Asians (Yang et al., 2020; Wang et al., 2021). Here, we obtained high-density SNP data of 42 Mongolian individuals from the boundary between Inner Mongolia Autonomous of China and Mongolia to provide a dense portrait of the genetic structure of Mongolians. We aimed to address the following three questions: 1) the extent of genetic heterogeneity or homogeneity among geographically different Mongolians; 2) the admixture sources and timing of Mongolians; 3) the signals of natural selection and the environmentally adapted gene in Mongolians.

MATERIALS AND METHODS

Sample Collections

We collected saliva samples from 42 Mongolian individuals from Baotou city of Inner Mongolia Autonomous Region. Each included individual followed the criteria of sampling collection that require people to have long-term resident history and do not have recorded intermarriages with other surrounding populations for at least three generations. Our work was approved by the Medical Ethics Committee of Xiamen University (Approval Number: XDYX2019009). Informed consent was obtained from all participants included in the study.

Genotyping and Data Merging

Genotyping was performed on the Illumina arrays covering genome-wide 600,000 SNPs designed to identify all known paternal Y chromosome and maternal mtDNA lineages. We first analyzed the relatedness of individuals measured by IBD (identified by descent) segments using KING software (Manichaikul et al., 2010); unrelated individuals were identified using the value of kinship < 0.0442. A total of 39 unrelated participants without family relationships were retained for subsequent analysis. We conducted quality control using PLINK (Chang et al., 2015) with --geno 0.2, --hwe 10e-10, filtering 670,269 SNPs. Then, the whole genome data of Mongolian was merged with the availably published dataset, including the Genome-Wide Human Origins Array genotype dataset and ancient/modern DNA of China and ancient Eastern Eurasian samples from 1240K capture dataset from David Reich Lab (Damgaard et al., 2018; de Barros Damgaard et al., 2018; Narasimhan et al., 2019; Ning et al., 2019; Ning et al., 2020; Yang et al., 2020; Wang et al., 2021), generating a combined Human Origins (HO) dataset covering 72,037 SNPs for subsequent analysis. Apart from this, the 1240K capture dataset, just combining the 1240K dataset, covered 186,187 SNPs.

Analysis of Population Structure and Relationships

We performed principal component analysis (PCA) on the merged dataset using the smartpca built-in EIGENSOFT package (Patterson et al., 2006). Modern individuals were used to calculate PCs, and ancient individuals were projected onto the pre-calculated components using the "lsqproject: YES" option. To characterize population structure further, we calculated f_3 in the form of f_3 (*population1*, *population2*; *Mbuti*) and f_4 statistics using qp3Pop and qpDstat in the ADMIXTOOLS package (Patterson et al., 2012). We added the "f4mode: YES" option to the parameter file for calculating f_4 statistics. We also estimated pairwise genetic distance by Fst using the smartpca program of EIGENSOFT (Patterson et al., 2006) with fstonly: YES and inbreed: YES parameter. We estimated relative genetic drifts and inferred a rooted maximum likelihood tree by TreeMix software (Pickrell and Pritchard, 2012). We conducted the best qpGraph-based models with population split and admixture events via the ADMIXTOOLS package.

Analysis of Population Admixture History Based on Sharing Allele Frequency

To investigate ancestry components in our Mongolian sample compared with other published Mongolian studies in different regions, an unsupervised clustering approach implemented in ADMIXTURE (Alexander et al., 2009) was firstly conducted, after filtering linkage disequilibrium using PLINK (Chang et al., 2015) with "--indep-pairwise 200 25 0.4" option, which retained a total 61,866 SNPs. Ancestry components and cluster memberships of 2084 individuals from 189 ancient and modern populations were calculated using the ADMIXTURE software. Clustering was performed for K = 2 to K = 20 in 100 bootstraps with different random seeds; we calculated the cross-validation errors to choose the best-fitted model. We also conducted admixture-f_3-statistics in the form f_3(*Source1*, *Source2*; *Mongolian_sub*) using the qp3pop program with default parameters in ADMIXTOOLS to explore the potential admixture surrogates showing significantly negative f_3 value. For modeling f_4 statistics-based admixture and estimating ancestral proportions in Mongolian, we applied *qpWave* (Patterson et al., 2012; Haak et al., 2015; Agranat-Tamir et al., 2020) to test for variation in ancestry proportions among the Mongolian and other modern Mongolian-related populations and detect the minimum number of ancestral sources; *qpWave* tests whether each possible pair of groups (Test i, Test j) is consistent with being a clade—since separation from the ancestors of a set of outgroup populations. qpAdm (Patterson et al., 2012) was used to calculate target populations as a combination of ancestry proportions from putatively selected source populations (references). To evaluate potential sex bias, we applied qpAdm to both the autosomes (default setting) and the X chromosome (adding "chrom:23" to the parameter file) for comparing the difference in the estimated ancestry proportions. For a certain ancestry, we calculated sex bias Z-score using the proportion difference between P_A and P_X divided by their standard errors ($Z=(P_A-P_X)/\sqrt{\dot{\alpha}_A^2+\dot{\alpha}_X^2}$, where $\dot{\alpha}_A$ and $\dot{\alpha}_X$ are the corresponding jackknife standard errors) (Mathieson et al., 2018). Therefore, a positive Z-score suggests that autosomes harbor a certain ancestry more than X chromosomes, indicating male-driven admixture, whereas a negative Z-score suggests female-driven admixture (Jeong et al., 2020). To understand the time scale of population mixture events in the Mongolian population, we used ALDER based on weight linkage disequilibrium statistics to date the admixtures with 28 years as one generation (Loh et al., 2013).

Fine-Scale Genetic Structure Based on FineSTRUCTURE

Bayesian clustering implemented in FineSTRUCTURE was used to reconstruct polygenetic relationships and further identify population structure. To reduce the computational burden, we randomly sampled 10 to 20 individuals in a large reference group. We first phased genome-wide dense SNP data using the SHAPEIT2 version (Delaneau et al., 2013) and then conducted FineSTRUCTURE (Lawson et al., 2012) analysis. FineSTRUCTURE R scripts based on the coancestry matrix inferred from ChromoPainter were conducted to construct the finer-scale population structure *via* heatmap, clustering dendrogram, and PCA.

ChromoPainterv2 and GLOBETROTTER Admixture Modeling

We performed a GLOBETROTTER (Hellenthal et al., 2014) analysis for Mongolian subgroups to obtain haplotype-sharing-based evidence of admixture. Using these haplotypes from SHAPEIT2, the "chunk length" output was obtained by running ChromoPainterv2 across all chromosomes. Using the chunk length output and painting samples, we ran GLOBETROTTER to estimate admixture date

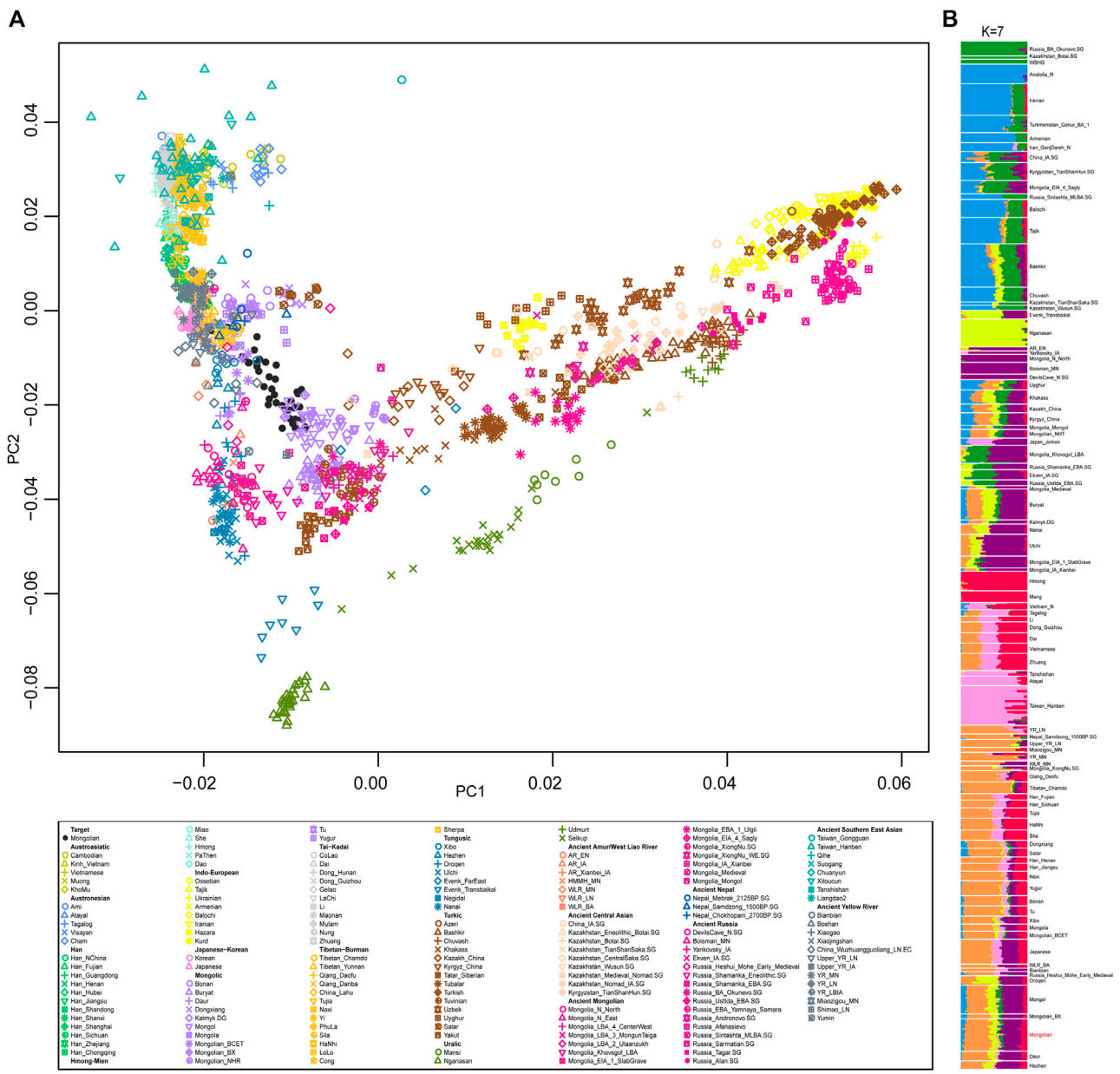

FIGURE 1 | The population structure of modern and ancient populations in Eurasia based on genome-wide data. **(A)**. PCA result showed an overview population relationship between modern populations and ancient populations. **(B)**. ADMIXTURE results (the lowest CV errors K = 7): ancestral components among Mongolian and modern and ancient populations in Eurasia.

by running 100 bootstrap replicates, assuming that there is detectable admixture using the "pro.ind:1" and "bootstap.date.ind:1" options.

Signals of Recent Positive Selection

The integrated haplotype score (iHS) and XP-EHH analysis were conducted to identify recent natural signatures of positive selective sweeps in the Mongolian population using the R packaged rehh2 (Gautier et al., 2017). The SNPs used in calculated iHS and XP-EHH were filtered by minor allele frequency (--maf 0.01) and snp missing (--geno 0.05). XP-EHH requires the definition of a reference population, and we chose the southern Altaic-speaking population in Guizhou and southern Tibetan-Burman population as references to explore whether there were differences in natural selection between different geographical Altaic populations and between northern and southern populations. The SNPs with maximum negative logical p value($-\log(p) > 4$) of iHS and XP-EHH were regarded as candidate sites under natural selection and used as test statistics. We performed the gene annotation by 3DSNP (Lu et al., 2017) and chose genes under the natural selection of the Mongolian population to conduct Gene Ontology (GO) enrichment analysis *via* DAVID Bioinformatics

Resources (Huang da et al., 2009a; Huang da et al., 2009b) and searched for related PheWAS traits and gene expression information from the global databases GeneATLAS (http://geneatlas.roslin.ed.ac.uk/) and GTEx (https://www.gtexportal.org/home/index.html), respectively.

RESULTS

Population Genetic Substructure Showing the West-East Admixture Cline

We generated and filtered 39 unrelated Mongolian individuals from Inner Mongolia Autonomous Region and merged the data with that published on modern and ancient populations in Eurasia to obtain a comprehensive population profile. In a principal component analysis (PCA) of Eurasian individuals, modern and ancient Eastern and Western Eurasian populations were separated into PC1 and PC2 split Eastern Eurasians along a north-south cline with Tungusic and Mongolic speakers who also connecting with the west-east Eurasian cline (**Figure 1A**). Mongolian individuals were scattered between Mongolic-speaking groups in China and ancient Mongolians, and a clear substructure was observed. To obtain a more focused Eastern Eurasian genetic profile, we removed Western Eurasian populations and the Mongolian population was stratified more obviously.

A model-based populations clustering analysis using ADMIXTURE showed a similar pattern (**Figure 1B**). Overall, the proportions of ancestry components associated with Eastern or Western Eurasians were well concordant with the results of PCA. The Mongolians derived most of their Eastern Eurasian ancestry from two components: one was most enriched in Sino-Tibetan speakers and the other was most represented by Mongolia_N_North that is Neolithic hunter-gatherers in Mongolia. The level of southern Eastern Eurasian-related ancestry represented by Hmong and Taiwan_Hanben in Mongolians was roughly higher than that of Mongols and Buryat. In addition, a small proportion of Western Eurasian-related ancestral component was detected in all Mongolians and Tungusic speakers. The level of admixture proportion of Western Eurasian and Eastern Eurasian in Mongolian intermediated between Mongols and previously studied Mongolians.

To obtain a more elaborate genetic structure of Mongolians, we conducted the IBD (identified by descent) analysis and pairwise f_4 statistics of all individuals (**Supplementary Figures S2–S8**). Taking results from PCA, admixture, pairwise IBD, and pairwise f_4 statistics into careful consideration, we grouped the Mongolian population into three subgroups for subsequent analysis, marked as Mongolian_inner who clustered with Mongolian speakers in China, Mongolian_mid, and Mongolian_outer clustered with Mongols and closed with Tungusic populations.

The Differentiated Genetic Affinity and Continuity Within Mongolian Subgroups

To quantitatively evaluate the genetic differences among three Mongolian subgroups and other modern and ancient Eurasian populations, we calculated the pairwise Fst genetic distances using the smartpca program (**Supplementary Table S1**). The genetic structure was confirmed by Neighbor-Joining Trees based on Fst (**Supplementary Figure S10**) results (Gautier et al., 2017), showing the different genetic affinities with other modern populations among those three Mongolian subgroups. Overall, three Mongolian subgroups showed lower genetic differences with other Mongolic-speaking groups and Tungusic populations. The Mongolian_inner was prone to cluster with Mongola_HGDP and Mongolian_BCET (Zhao et al., 2020) that belongs to Inner Mongolians and shares more genetic drift with East Asians, as shown in a previous study, and the Mongolian_outer group possessed a much closer genetic affinity to Mongols and Mongolian_BX who is Mongolian_Chahar and harbors more Western Eurasian-related ancestry than Mongolian_BCET, which was consistent with results of f_4(*Mbuti.DG*, *X*; *Mongolian_sub*, *Mongolian_BXBC/Mongolian_HNT/Mongolian_TE*) reflecting as no significant Z (**Supplementary Tables S4, S5**). The Mongolian_outer showed a similar genetic profile to Mongols with a higher genetic difference with Sino-Tibetan population and southern East Asian populations and lower genetic difference with populations harboring Western Steppe-related ancestry compared to Mongolian_inner and Mongolian_mid (**Supplementary Table S1**). Consistent with the pattern of genetic variations that showed in PCA and Fst and the shared ancestral components observed in ADMIXTURE, the result of outgroup f_3 statistics (**Figure 2**) in the form of f_3 (Mongolian_sub, modern Eurasian; Mbuti) showed that Mongolian_inner possessed the most shared ancestry with modern Han groups and Mongolian_outer had strong genetic drift with Tungusic populations, while Mongolian_mid shared closer genetic affinity with Han and Tungusic populations. The genetic affinity profile also demonstrated that, in outgroup f_3 (Mongolians, ancient Eurasian; Mbuti) (**Supplementary Figure S9**), Mongolian_outer shared the most significant genetic drift with ancient Northern Asian hunter-gatherers (previously called ANA or AEA), while Mongolian_inner had a closer genetic affinity with populations harbored Neolithic farmers related ancestry, suggesting an extent of long-term genetic continuity in Northeast Asia and the communication between Northeast Asian hunter-gatherers and millet farmers of North China. In addition, the shared genetic drift with three Mongolian subgroups in the Eastern Eurasian populations was stronger than that in Western Eurasians, indicating the deeper Eastern Eurasian lineage of Mongolian. The phylogenetic relationships between the studied three Mongolian subpopulations and modern Eurasian populations were further confirmed by a TreeMix-based phylogenetic tree. Among a large reference population set consisting of 47 Eurasian populations as representatives from the main language families and Mbuti as the root, we also identified a gene flow event from Tungusic into the Buryat population but not into Mongolian (**Supplementary Figure S11A**). When including fewer reference populations, one western gene influx flow into Mongol and the other gene flow from Western Eurasian into Eastern Eurasian was identified in the Tuvinian population of Siberia (**Supplementary Figure S11B**).

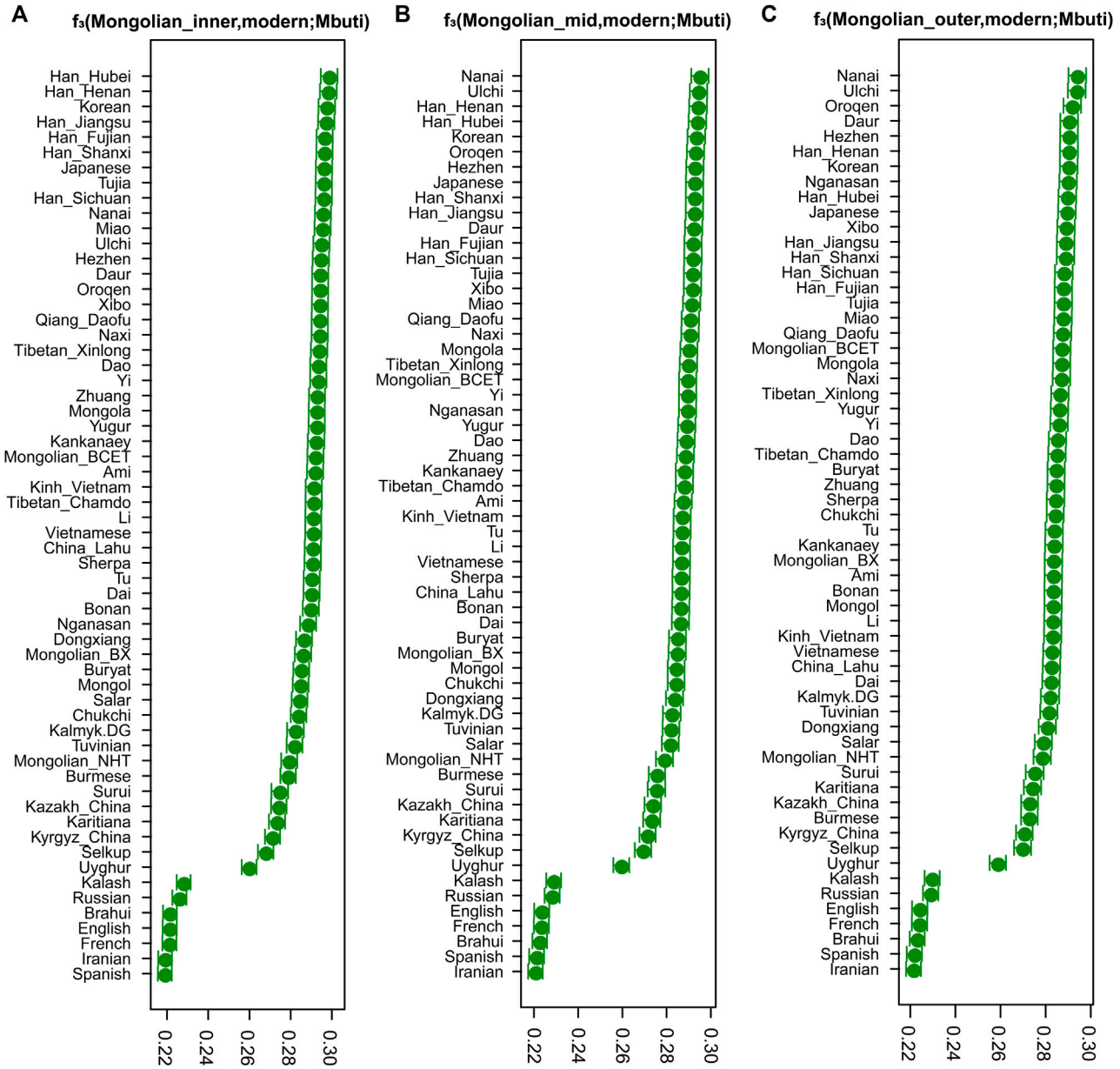

FIGURE 2 | The results of three-population statistic. The shared genetic drift between modern Eurasian populations and Mongolian subgroups.

The genetic differentiation and affinity profile among three Mongolian subgroups was further certified by f_4 statistics test in the form of f_4(*Mbuti.DG*, *X*; *Mongolian_sub1*, *Mongolian_sub2*) (**Supplementary Table S3**), showing the significant difference in sharing affinity with ancient and modern East Asians in China between Mongolian_inner and Mongolian_outer. Mongolian_inner harbored more ancestry related to millet and rice farmers than Mongolian_mid and Mongolian_outer. The result provided evidence that Mongolian_outer harbored more Western Steppe-related ancestral components than Mongolian_inner and Mongolian_mid. Interestingly, there were differences in sharing genetic affinity to WSHG (Western Siberian Hunter-Gatherers) and Mesolithic hunter-gatherers in Japanese population (Japan_Jomon) and Iranian Neolithic farmers among these Mongolian subgroups. The additional Iranian-related ancestry was detected in ancient Mongolian populations after the Bronze Age and decreased in the modern Mongolian subgroups; notably, the level of Iranian-related ancestry in Mongolian_outer and Mongolian_mid was roughly equal to populations associated with the Late Bronze Age Ulaanzuukh (1450–1150 BCE) and Early Iron Age Slab Grave (1000–300 BCE) cultures in eastern and southern Mongolia (**Supplementary Tables S6, S7D**). The results of f_4(*Mbuti.DG*, *X*; *Mongolian_sub1*, *Mongolian_sub2*) (**Supplementary Table**

S3), f_4(*Mbuti.DG*, *X*; *Mongolian_sub*, *Mongol/Mongola_HGDP*) (**Supplementary Table S4**), and f_4(*Mbuti.DG*, *X; Mongolian_sub*, *Mongolian_BCET/Mongolian_BX/Mongolian_NHT*) (**Supplementary Table S5**) did provide a robust evidence of the differentiation of sharing genetic affinity with Mongols and Inner Mongolians among three Mongolian subgroups, showing a similar genetic profile with Mongola_HGDP and Mongolian_BCET of Mongolian_inner and analogical genetic structure with Mongols and Mongolian_BX of Mongolian_outer.

To further reveal the different genetic affinities of Mongolian-related populations, we used a distantly related set of outgroups. We observed a significant population stratification in three Mongolian subgroups and genetic heterogeneity in modern and ancient Mongolian-related populations except for Mongolian_inner that showed the genetic homogeneity with Inner Mongolians (Mongola_HGDP and Mongolian_BCET), Mongolian_BX that showed the genetic homogeneity with Mongolian_outer/Mongol/Mongolia_Medieval, and Buryat that showed the genetic continuity with Mongolia_Medieval (**Supplementary Table S8A**). We obtained a subtler population structure of Mongolian-related populations when we repeated the qpWave analysis adding outgroups that are genetically closer to the test groups. With this more powerful set of outgroups, Mongol and Buryat also provided evidence of not being pairwise clades with the remaining groups (**Supplementary Table S8B**), while Mongolian_BCET still displayed a close relationship with Mongola_HGDP/Mongolian_inner. Thus, beyond the broad observation of genetic affinities between three Mongolian subgroups, we also observed subtle ancestry heterogeneity in Mongolia since Bronze Age. Mongolian_inner showed continuity with Xiongnu populations in Iron Age and Mongolian_mid and Mongolian_outer showed some extent of continuity with Xiongnu, which was further confirmed in the results of f_4(*Mbuti*, *X*; *Mongolian_inner*, *Mongolia_XiongNu.SG*) ($|Z| < 3$) and f_4 (*Mbuti*, *X*; *Mongolian_mid/Mongolian_outer*, *Mongolia_XiongNu.SG*) (part of $|Z| < 3$) (**Supplementary Table S7**). In addition, three Mongolian subgroups showed evident genetic continuity with Medieval Mongolian and the ancestry related to Han increased in modern Mongolians since the Yuan Dynasty.

The phased Mongolian and Eurasian populations data were also used to conduct haplotype-based fineSTRUCTURE and the finer-scale population structure of Mongolian was further comprehensively characterized. The inferred polygenetic tree showed that Mongolian_inner clustered with Mongola_HGDP and Mongolian_BCET, one part of Mongolian_mid clustered with Mongola, and the others clustered with Mongol, while Mongolian_outer was clustered with Mongolian_BX and Mongol (**Figure 3B**). Besides, the pattern of shared haplotypes based on the ChromoPainter showed prominent sharing haplotypes among Mongolian_outer, Mongolian_BX, and Mongol and remarkable sharing haplotypes among Mogolian_inner and Mongola (**Figure 3A**). PCA calculated from the coancestry matrix generated by fineSTRUCTURE also confirmed the west-east cline of Eurasians and the north-south cline of Eastern Eurasians (**Supplementary Figure S13**).

The Admixture History of the Mongolian Population Based on Allele Frequency and Haplotype-Based GLOBETROTTER

We performed allele frequency-based three-population (f_3) tests to characterize the admixed gene pools of three Mongolian subgroups. Testing all possible pairs of 115 present-day "source" groups and 117 ancient "source" groups, we detected highly significantly negative f_3 statistics ($f_3 \leq -3$ standard error; **Supplementary Table S2**), providing unambiguous evidence that the target population is a mixture of groups related, perhaps deeply, to the source populations. Reference pairs with the most negative f_3 statistics, for the most part, involved one Eastern and one Western Eurasian group (including Neolithic Iranian farmers and Chalcolithic Iranians to represent West/South Asian-related ancestry), supporting the qualitative impression of east-west admixture from PCA and ADMIXTURE analyses. To highlight the difference among Mongolian subgroups, we looked into f_3-results with representative reference pairs comprising ancient Eurasians (Sintashta to represent the steppe Middle and Late Bronze Age ancestry and Chalcolithic Iranians to represent South Asian-related ancestry, Ulchi and Han, and ancient Mongolia to represent Eastern Eurasian-related ancestry). Farmer-related ancestry was the best representation of Eastern Eurasian ancestry for Mongolian_inner compared to Ulchi; farmer-related and Neolithic hunter-gatherers–related ancestry (Ulchi is regarded as the most genetic homogeneous population with Neolithic hunter-gatherers of DevilsCave) both represented ancestries related to Eastern Eurasian well in Mongolian_mid and Mongolian_outer. Considering the admixture events and sources that we observed in Mongolian subgroups, we applied qpWave/qpAdm to validate different proposed admixture scenarios and ancestral proportions. In the two-way mixture model of Western Steppe populations and Eastern Eurasians (**Figure 4**, **Supplementary Table S9A**), Russian_Sitashta_MLBA and WLR_BA, a mixture of Neolithic hunter-gatherers and millet farmers, approximated the Mongolian populations well ($\chi 2$ $p \geq 0.05$), while the model of Eastern Eurasian simply represented by Neolithic hunter-gatherers (Mongolia_N_North and DevilsCave_N, AR_EN) or millet farmers (YR_LN) and farmers in West Liao River (WLR_MN) mostly failed, indicating that Neolithic hunter-gatherers, millet farmers, and Western Steppe populations contributed to the formation of Mongolian population together and the gene flow from the population related to millet farmers into the gene pool of Mongolian continued to today. The ancestral proportion of Western Steppe in those Mongolian subgroups was distinct, showing the parallel genetic makeup of Mongolian_outer and Mongolian_BX harboring a higher level of Western Steppe ancestry (10.9%, 12.8% Russian_Sitashata_MLBA/11.6%, 11.5% Mongolia_EBA_2_Chemurchek, a mixture population with Western Steppe), and the proportion of the ancestry in Mongolian_inner, Mongola_HGDP, and Mongolian_BCET were similar (5.6%, 5.2%, and 5% Russian_Sitashata_MLBA, respectively),

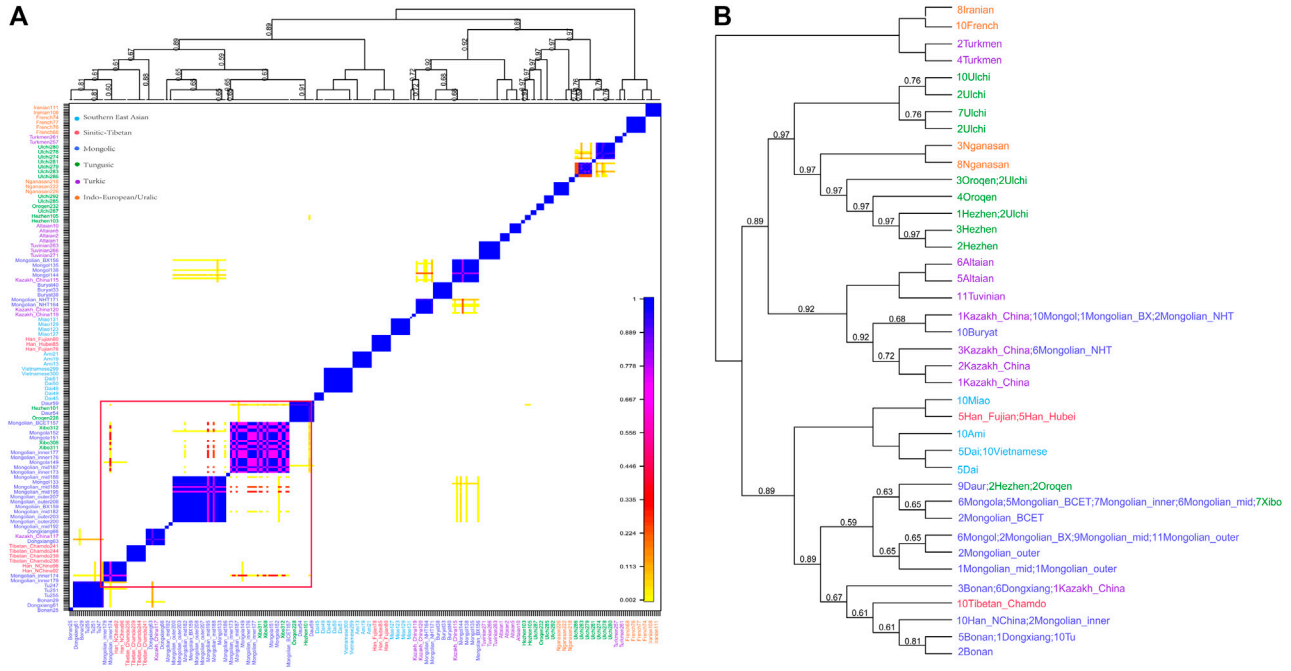

FIGURE 3 | Pheatmap of sharing haplotypes and clustering dendrogram and by FineSTRUCTURE based on the chunk length.

the proportion in Mongolian_mid intermediated between Mongolian_inner and Mongolian_outer, coinciding with the population structure mentioned above. A more complex three-way model of YR_LN + Mongolia_N_North + Russia_Sintashta_MLBA fitted all Mongolian groups ($\chi 2$ $p \geq 0.05$) (**Supplementary Table S9B**) but showed prominently various proportions of YR_LN and Mongolia_N in Mongolian subgroups, which also shown in two admixture models of millet farmers (YR_LN) + Russian_Sitashta_MLBA ($\chi 2$ P (Mongolian_inner/Mongola_HGDP/Mongolian_BCET) > 0.01), reflecting minor heterogeneity in the Eastern Eurasian source of Mongolians. Considering that we observed a gene flow signal from Iranian-related populations, all subpopulations were fitted by three models with YR_LN + Mongolia_Khovsgol_LB + Turkmenistan_Gonur_BA_1 (3.8–6%) when we added the third ancestral source of Turkmenistan_Gonur_BA_1 where is the key EBA site of the Bactria-Margiana Archaeological Complex (BMAC) culture. The legacy of the spread during the Early Iron Age was mediated by increased contact and mixture with agropastoralist populations in the region of Turan and then introduced into northwestern Mongolia along the Inner Asian Mountain Corridor. Overall, several ancestral sources contributed to the formation of modern Mongolian and the population structure was the result of different proportions of ancestries.

We reconstructed the deep demographic history using qpGraph. Mbuti, Denisovan, Onge and Tianyuan were included to explore the basal model; Early Bronze Age Afanasievo and Chemurchek, Neolithic hunter-gatherers in Mongolia Plateau, millet farmers (YR_LN), Tibetan Plateau, and Iron Age Hanben were used as ancestral source proxies from Western Eurasian, Mongolia, millet farmers in Yellow River, and southern populations. We found that Mongolian subgroups could be modeled as the mixture of EBA_Chemurchek (34–37%) derived from Western Steppe herders (47–55%) and Mongolia's Neolithic hunter-gatherers related ancestry and Han-related ancestry (63–66%) (**Supplementary Figure S12**). Our qpGraph models were compatible with qpAdm results and further supported the fact that Western Eurasian herders, ANA, and millet farmers contributed to the genetic formation of modern Mongolians.

The ALDER method based on weighted linkage disequilibrium statistics also provided evidence of population structure within Mongolians (**Supplementary Table S10**). ALDER demonstrated multiple admixture sources from southern populations, Han, Tungusic speakers, and populations harboring Western Eurasian-related ancestry. Overall, the admixture events of Eastern and Western Eurasians occurred in a historic period (~400–~1700 years ago), which were consistent with the extensive western-eastern communication along the Silk Road (Yao et al., 2004; Liu et al., 2021) and the western expansion of the Mongol empire. ALDER detected extra admixture events between Tungusic/Turkic/Indo-European speakers and southern populations around ~170–~1700 years ago. Intriguingly, the admixture signal from Han was just detected in Mongolian_outer with admixture time ranging from ~600 to ~1000 years ago, inferring that the recent Han-related ancestry flowed into the Mongolians during the Late Tang Dynasty to the Yuan Dynasty when the Khitans controlled large areas of the Eastern Steppe and the Khitan empire fell to the Jurchen's Jin Dynasty, which was then conquered in turn by the Mongols in 1234 CE. Companied by the expedition to the West by Mongol nobles, the flow of people groups was more frequent than ever before in Eurasia in the 13th century.

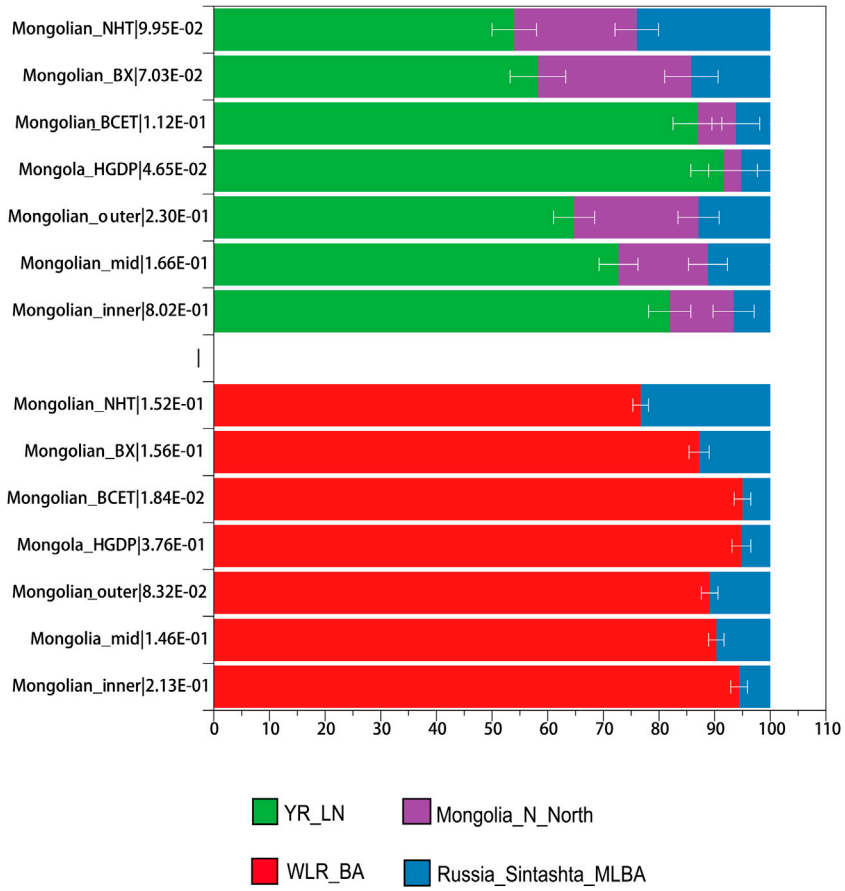

FIGURE 4 | qpAdm-based admixture models for Mongolian subgroups.

We further performed haplotype-based GLOBETROTTER to obtain a high-resolution characterization of the admixture landscaped of three Mongolian subgroups. All targets showed robust signals of west-east admixture (**Supplementary Table S11**). The west-east admixture event in subgroups could be traced back to 29–40 generations, with the inferred majority contributing Eastern Eurasian sources ranging from 77 to 87%. Mongolian_inner derived Eastern Eurasian ancestry from Han-related ancestry, while Mongolian_mid and Mongolian_outer retained Eastern Eurasian ancestry from Northeast Asia. The different Eastern Eurasian ancestral surrogates in Mongolian subgroups were in line with admixture models of qpAdm/ALDER. Meanwhile, GLOBETROTTER identified the second less strongly signaled north-south admixture event.

The Paternal/Maternal Lineages of Mongolian

We assigned 39 mitochondrial genomes based on 4,198 maternal lineage-informative SNPs and 33 Y-chromosomal genomes based on 22,512 paternal lineage-informative SNPs (**Supplementary Table S12**). The maternal mtDNA lineages of Mongolians were diverse, with lineages significantly enriched in present-day East Asian populations (A, B4, C4, D4, F1, G, M, and N), showing terminal lineage frequencies ranging from 0.0256 to 0.0513 (G2a5: 2); B4, C4, D4, and F1 were prevalent in the Mongolian population. From the paternal perspective, 24 different terminal paternal lineages with frequencies ranging from 0.0303 to 0.1212 (C2b1a3b~: 4). Siberian-dominant paternal lineage was detected (C2b1a and C2c1a). In addition, more East Asian Y-chromosomal founding lineages were identified in Mongolians with dominant lineage O2a2b1a2. To further validate the potential sex bias admixture in the Mongolian population, we used qpAdm to estimate the sex bias Z-score. We observed positive Z sex bias scores in different two-way admixture models focused on Mongolians, which suggested a male-dominated admixture of Han-related ancestry.

The Natural Selection Signal and Functional Genes in Mongolian

We employed the iHS test to identify recent natural signatures of positive selective sweeps in the Mongolian population. Some differences of loci under natural selection detected by iHS among Mongolian subgroups existed, the GO enrichment of

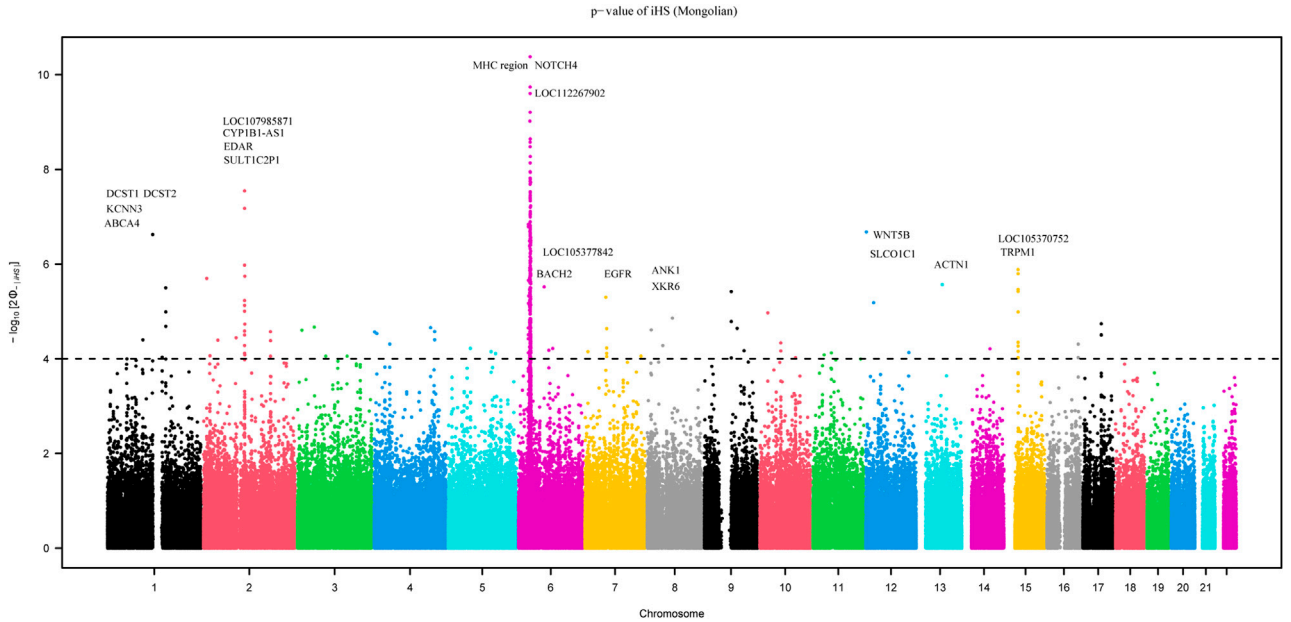

FIGURE 5 | The result of recent natural signatures of positive selective sweeps in Mongolian population based on iHS showed the strongest positive selection region in MHC region.

Mongolian subgroups' genes with significant natural selection, however, all showed mainly enriched in cellular component with the membrane (**Supplementary Table S13B–D**; **Supplementary Figure S14**). Therefore, considering the small sample size of subgroups that is likely to cause the deviation of detected selection signals and the homogenous Mongolian relative to other populations, we performed the natural selection related analysis on the whole Mongolian group subsequently (**Figure 5**, **Supplementary Table S13A**). We observed the highest $-\log_{10}$p (iHS) score in the Major Histocompatibility Complex (MHC) region, indicating that genes in this region might experience strong positive selection, which has been already found in previous studies. In addition, in the gene *TRPM1* located in chromosome 15, more than 30 SNPs showed strong selection signatures ($-\log_{10}p > 4$), which indicated significant enrichment of selection in this genomic region. The *EDAR* gene (rs922452) was identified with higher |iHS|, which has shown the strong signatures of positive selection in East Asians (Kamberov et al., 2013). Notably, the alcohol dehydrogenase (*ADH*) gene cluster was not identified. The derived allele frequencies of the *ADH* gene family in those Mongolian subgroups, however, were higher and associated with the genetic affinity to Han (adjusted $R^2 > 0.5$, adjusted p < 0.01) (**Supplementary Table S14**). In addition to iHS, XP-EHH was also used to indicate the effect of local positive selection. The results of XP-EHH (southern Altaic/southern Tibetan-Burman vs. Mongolian) (**Supplementary Table S13E–F**) showed overlapping positive selection signals in northern Mongolian population relatively to southern Altaic populations (Mongolian_Guizhou and Manchu_Guizhou) and southern Tibetan-Burman population, including *SLC28A3*, *SLC47A1*, *LOC100506499*, *ZFPM2*, *AGBL4*, and *MHC* regions. However, there still were differences in positive selection between Mongolian relative to southern Altaic populations and Mongolian relative to southern Tibetan-Burman population. The number of loci that experienced positive selective sweeps in Mongolian relative to southern Altaic was less than that in Mongolian relative to southern Tibetan-Burman population, indicating a diverse local selection and adaption in regions. Genes subjected to natural selection were concentrated in a membrane-associated cellular component, while genes enriched in molecular function and biological processes were associated with immune response (**Figure 6**). Furthermore, the related traits from the GeneATLAS dataset in chromosome 6 showed immune-related traits. Gene expression of those genes was mostly focused on such immune tissue as brain, reproductive organ, skin, stomach, and spleen (**Supplementary Table S15**).

Animal husbandry is the main means of livelihood of Eastern Steppe herders; therefore, dairy livestock is a staple food and traditional diet style. We found that despite a pastoralist lifestyle started in Late Bronze Age, the Mongolians did not have a higher frequency of derived mutations associated with lactase persistence (*LCT/MCM6*, frequency < 0.07143), which showed a strong positive correlation with the genetic affinity to Western Eurasian (adjusted $R^2 > 0.7$, adjusted p < 0.05) (**Supplementary Table S14**). Given the dairy habit of Mongolians, we observed the derived allele frequencies of the FADS1 gene (**Supplementary Table S14**) intermediated between northern Han and southern Han when fatty acid desaturase (*FADS*) gene family which plays vital role in the biosynthesis of polyunsaturated fatty (Schaeffer et al., 2006; Nakayama et al., 2010; Song et al., 2013; Wu et al., 2017) has been taken into account. Due to the absence of a

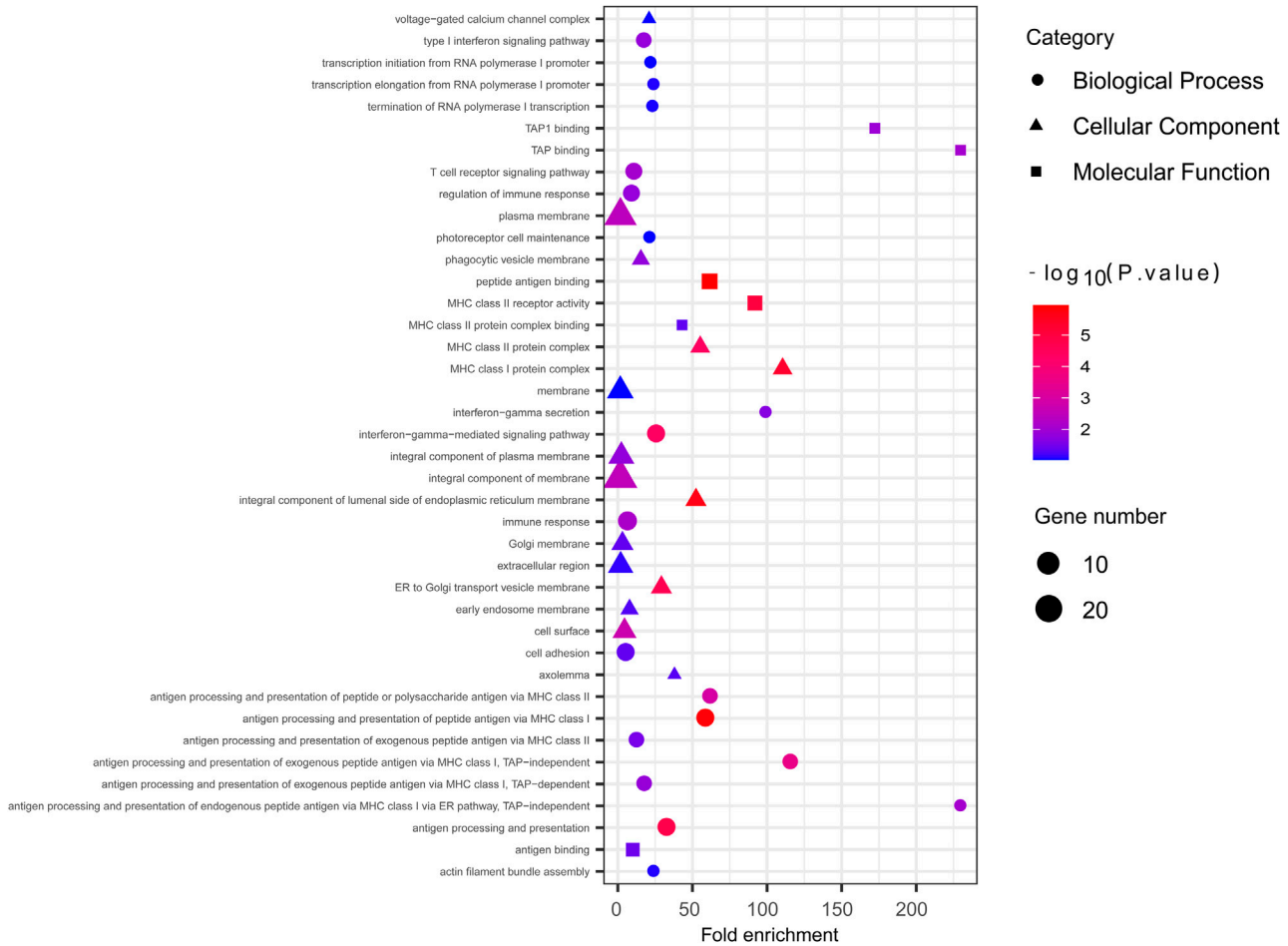

FIGURE 6 | The GO enrichment analysis of Mongolian showed genes with significant natural signal were mostly enriched in membrane-associated cellular component.

phenotype dataset, we could not further analyze the association of *FADS* with the high-fat dairy consumption of Mongolians.

DISCUSSION

We provided newly generated genome-wide SNP data of the Mongolian population from the Inner Mongolia Autonomous Region and performed a comprehensive population genetic analysis to investigate the genetic origin and admixture history. Findings from IBD segments among pairwise individuals, approximate ancestral composition differences from ADMIXTURE result, and pairwise f_4(studied individual1, studied individual2; Atayal/Han_NChina/Tibetan_Chamdo/Ulchi/Mongol/Mongola, Mbuti) suggested that our focused Mongolian existing population stratification was genetically separated into three subgroups. Overall, even though three Mongolian subgroups had a closer genetic relationship with Tungusic populations, which might result from Altaic-speaking populations—the common ancestor of Tungusic and Mongolian provided by linguistic information, there were differences in sharing genetic affinity with Eurasian populations among Mongolian subgroups. The grouped Mongolian subpopulations showed significant distinction of genetic affinity with previously studied Mongolians of Inner Mongolia Autonomous Region and Mongols. That was, Mongolian_inner had a similar genetic profile with Mongola_HGDP and Mongolian_BCET showing the most shared ancestry with modern Han groups, while Mongolian_outer genetically closed to Mongols showed the higher genetic difference with Sino-Tibetan and southern East Asian populations and lower genetic difference with populations harboring Western Steppe pastoralists related ancestry than Mongolian_inner and Mongolian_mid, and the genetic profile of Mongolian_mid intermediated between Mongolian_inner and Mongolian_outer. The f_4(*Mbuti*, *X*; *Mongolian_sub1*, *Mongolian_sub2*), f_4(Mbuti, X; Mongolian_sub, Mongol/Mongola_HGDP), and f_4(Mbuti, X; Mongolian_sub, Mongolian_BCET/Mongolian_BX/Mongoliann_NHT), qpWave homogenous test did further provide evidence of the genetic structure and the diverse sharing genetic affinity to modern/ancient Mongolians among three Mongolian subgroups.

Paleogenomic studies demonstrated that the disparate genetic profile of ancient Mongolian existed at different times and geographic regions and multiple ancestral sources flowed into Mongolia Plateau shaped the higher genetic heterogeneity of ancient Mongolian: the local ANA ancestry, the ephemeral ANE ancestry, the eastward movement of Western Steppe herders in a different period, limited gene flow of Iranian-related ancestry, and recent Han-related ancestry. The intercontinental expansion of Mongols established the genetic structure that characterized the present-day Mongolic-speaking population in North Asia. Model-based populations clustering analysis of ADMIXTURE and admixture f_3 tentatively suggested that the differentiated genetic profile of Mongolians might be the results of various ancestral sources and proportions: the Eastern Eurasian including Neolithic hunter-gatherers related ancestry (ANA, represented by DevilsCave_N/Mongolia_N_North), millet farmers related ancestry (represented by YR_LN), and relative low proportion of ancestry related to Western Steppe herders contributed to the gene pool of modern Mongolian, in agreement with previous studies (Zhao et al., 2020). The gene flow from Western Eurasian was preliminarily detected in Mongol population of TreeMix-based phylogenetic tree; the ancestral source was finally identified in qpAdm, ranging from 5.6 to 11.6% in those Mongolian subgroups; ALDER and GLOBETROTTER supported that the west-east admixture event was recently estimated in the period ranging from Tang Dynasty to Yuan Dynasty. One important point is that the truth admixture scenarios might be continuous, complicated admixture and estimated admixture only provide simply a single event, and the recent date should be paid attention. The admixture between Western Steppe pastoralists and ancient Eastern Eurasians in the Mongolia Plateau has been attested in paleogenomics studies, including Early Bronze Age Yamnaya and Afanasiveo populations showing primary culture influence and limited genetic impact and Middle and Late Bronze Age Andronovo and Sintashta with visible genetic contribution to Eastern Steppe populations and historic nomadic pastoral. What is more, the Silk Road, connecting the Eurasian continent, promoted not only prosperous western-eastern population communication and culture exchange but also genetic material flow. The rise of the nomadic empire in the historic period facilitated the population interaction of western-eastern Eurasian and farmers-pastoralists.

Neolithic hunter-gatherers and millet farmers in East Asia made a large genetic contribution to the formation of Mongolian matched by the two-way admixture model of WLR_BA that is a mixed population of Neolithic hunter-gatherers and millet farmers and Western Steppe herders or adequately modeled as YR_LN + Mongolia_N_North/AR_EN + Russia_Sintashta_MLBA or YR_LN + Russia_Sintashta_MLBA + Mongolia_N_North + Turkmenistan_Gonur_BA_1. The proportion of Neolithic hunter-gatherers contributing to Mongolian subgroups increased with the genetic affinity with Mongols; in contrast, the ancestry of Neolithic farmers dedicated to Mongolian subgroups increased with the genetic affinity with Han. The derived Eastern Eurasian ancestry (ANA) from a gene pool was similar to contemporary Tungusic speakers from Amur River Basin, suggesting a genetic connection among the speakers of languages belonging to the Altaic macrofamily (Turkic, Mongolic, and Tungusic language families) (Yunusbayev et al., 2015; Pugach et al., 2016; Chen et al., 2021; Zhang et al., 2021). The genetic connection of Mongolic and Tungusic populations was also shown in a similar pattern of the paternal Y chromosomes (Huang et al., 2018a; Huang et al., 2018b; Wen et al., 2019; Wei et al., 2018a; Yan et al., 2015; Wei et al., 2018b). Trans-Eurasian language origin hypothesis asserted that the language subfamily of Mongolic, Tungusic, Turkic, and Japonic-Korean originated from Neolithic Hongshan culture in West Liao River Basin; the Hongshan farmers in West Liao River Basin migrated westward to the Mongolia Plateau and gradually developed into nomadic style, leading to the separation of Proto-Turkic and Proto-Mongolic-Tungusic languages. However, our findings did not observe the Hongshan related ancestry in Mongolic speakers and supported the Trans-Eurasian agricultural origin and diffusion hypothesis (the two-way admixture of WLR_MN + Russia_Sintashta_MLBA failed, **Supplementary Table S9A**). Considering the genetic similarity continuity in ancient Northeast Asian, our established genetic landscape in Mongolians supported the potential Northeast Asian origin of the Altaic language. What's more, the genetic contribution of Han-related ancestry might be mediated by the gene flow into ancient populations in Mongolia started in the Xiongnu Regime of the Early Iron Age (Jeong et al., 2020; Wang et al., 2021). The unique geographic position of the Inner Mongolia Autonomous Region has always been the boundary between the agriculture of the Han population and the pastoral husbandry of herders. Therefore, the recorded communication between the populations related to Han and Eastern Steppe pastoralists started in Han Dynasty when the rise of the Xiongnu Regime often invaded the boundary of the Han Dynasty, which facilitated the cultural and genetic exchanges. Since the confrontation between the Han and the nomads opened up the historical situation, this kind of exchange between the agricultural people and the nomads has continued until Genghis Khan's cavalries swept across the whole Eastern Eurasia and the exchanges between the agricultural people and the nomads reached the peak; our ALDER results also suggested gene flow from Han into Mongolian during the rise of the Mongol empire. The Han-related ancestry increased with the time transection. Sex-biased patterns of genetic admixture could be informative about gendered aspects of migration, social kinship, and family structure. We observed a clear signal of male-biased Han admixture in the Mongolian population, corresponding to the Y chromosome lineage O2a in some Mongolian individuals.

The additional ancestral source related to populations of Central Asia (Caucasus/Iranian Plateau/Transoxiana regions) flowed into Eastern Eurasian initiated in the Early Iron Age along the Inner Asian Mountain Corridor/the Tian Shan Mountains, which is detected in the Iron Age groups such as TianShan Saka, Mongolia_Chandman_IA (Jeong et al., 2019; Jeong et al., 2020). This genetic influx continued to the Xiongnu Empire and even the Early Medieval period. The westward disseminating Turkic language influenced the group in the south-eastern side of the Tian Shan Mountains, such as

Wusun and Kangju (Damgaard et al., 2018). The Xiongnu population and in a later Uyghur period, Wusun and Kangju in the Tian Shan Mountains received an Iranian-related ancestry (BMAC related or Neolithic Iranian-related). Although the Iranian-related ancestry component did not largely contribute to the gene pool of the Mongolic-speaking population, it has been detected in modern Mongolians. Our modern Mongolian populations also showed a minor genetic affinity to Iranian-related populations; the genetic affinity in Mongolian populations was inferior to that in ancient populations in Mongolia Plateau since the Late Bronze Age. The qpAdm results further provided robust evidence that the subtle genetic influx was dedicated to the gene pool of modern Mongolians.

The Eastern Steppe has served as a crossroad for human population migration and cultural exchanges: the eastward expansion of Western Steppe herders since the Bronze Age (Allentoft et al., 2015; de Barros Damgaard et al., 2018; Narasimhan et al., 2019; Ning et al., 2019; Wang et al., 2019); the WSHG (West Siberian hunter-gatherers) in Central/South Asia (Jeong et al., 2019; Narasimhan et al., 2019; Wang et al., 2021); Iran-related ancestry flowed into northern Mongolia since Early Iron Age (Jeong et al., 2020). More recent historical migrations are companied by the opening of the Silk Road and the westward expansions of Turkic and Mongolic groups. The flourishing population movement facilitated the intricate formation history of the Mongolian population. Our sample was collected from Darham Mau Mingan Union Flag of Baotou of Inner Mongolia Autonomous Region, which is located in the hinterland of the Bohai Rim and the Yellow River Economic Belt and has functioned as a conduit for human migration and cultural transfer between Mongols and China so that also be characterized as an immigrant city with flourished migrations. Prosperous economic and trade activity promotes the population exchange between China and Mongols, which is also shown in the genetic profile of three Mongolian subgroups.

The detailed population origin and admixture history provide clues to understanding natural selection and functional genes. In Mongolians, we detected the strong selection signal from the MHC region, which is a key point of the human immune response. Gene enrichment analysis also supported the most enrichment related to the human immune response in terms of cellular component and molecular function. The positive selective sweeps in this region have been already identified in Han populations (Zheng et al., 2021). However, the alcohol dehydrogenase (*ADH*) gene cluster that underwent regional selective sweeps in East Asia (Ma et al., 2005; Li et al., 2008; Li et al., 2011; Allentoft et al., 2015) was not identified, and the derived allele frequency of *ADH* genes in three Mongolian subgroups showed a strong correlation with the genetic affinity to Han, indicating the possibility of introducing genes into Mongolians. The fact that Mongolians started milk consumption in the Late Bronze Age (Jeong et al., 2018; Wilkin et al., 2020) suggested that ruminant dairy pastoralism was adopted on the Eastern Steppe by local hunter-gatherers through a process of cultural transmission and minimal genetic exchange with outside groups. Ancient populations in the Eastern Steppe of different periods have a negligibly low frequency of the derived mutation with no increase in frequency over time (Jeong et al., 2018; Jeong et al., 2020). The derived mutation in modern Mongolians was still at low frequency, even if the frequency increased with the genetic affinity to Western Eurasian in subgroups. Therefore, the ability to digest large quantities of lactose for millennia in the absence of lactase persistence is remarkable, which may be related to their reportedly unusual gut microbiome structure.

CONCLUSION

We generated genome-wide data from 42 Mongolians of the Inner Mongolia Autonomous Region. We first identified a significant genetic differentiation among Mongolians, who were structured into three distinct genetic clusters harboring various Western and Eastern Eurasian ancestries. Findings based on the *f*-statistics demonstrated that Mongolian subgroups possessed different Chinese Mongolian/Mongols/Tungusic/East Asian affinities, indicating successful population migration in a frontier city. The successfully fitted four-way admixture model revealed that Eastern Eurasian ancestry included Northeast Asian Neolithic hunter-gatherers related ancestry and East Asian millet farmers related ancestry and Western Eurasian ancestry included Western Steppe herders related ancestry and small Iran-related ancestry. Furthermore, the natural selection analysis of Mongolian showed that the MHC region underwent significant positive selective sweeps and the functional *ADH* and *LCT* were not identified. This study characterized the complex population admixture history of Chinese Mongolians, which shed light on the intensified interaction and mixture history of farmers and pastoralists in the boundary between agriculture of contemporaneous imperial Han and pastoral husbandry of herders. Moreover, it revealed intricate genetic structure in a frontier industrial city. The genetical structure of populations inspired that the regional positive selection with allele frequency change might be associated with the genetic affinity. It will be extremely important to expand the set of available ancient and modern genomes across the Eastern Steppe to fully reveal the population structure and history of the Eurasian Steppe and further investigate the local natural selection of functional genes.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://zenodo.org/record/5067504, doi: 10.5281/zenodo.5067504.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Medical Ethics Committee of Xiamen University (Approval Number: XDYX2019009). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

C-CW designed this study. XY wrote the manuscript. S collected the samples. XY, GH, KZ, HM, JZ, MY, JC, XZ, LT, and YL conducted the experiment and analyzed the data. All authors reviewed the manuscript.

FUNDING

The work was funded by the National Natural Science Foundation of China (31801040), 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, 0310/X2106027), Nanqiang Outstanding Young Talents Program of Xiamen University (X2123302), the Major Project of National Social Science Foundation of China (20&ZD248), the European Research Council (ERC) grant to Dan Xu (ERC-2019-ADG-883700-TRAM), and China Postdoctoral Science Foundation of China (2021M691882).

ACKNOWLEDGMENTS

S. Fang and Z. Xu from Information and Network Center of Xiamen University are acknowledged for the help with the high-performance computing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.735786/full#supplementary-material

Supplementary Figure S1 | **(A)**. The geographic distribution of Mongolians. **(B)**. The PCA results focused on Eastern Eurasians displayed a clearer genetic structure of Mongolians.

Supplementary Figure S2 | The pairwise f4 statistics in forms of f4 (studied individual1, studied individual2; Atayal, Mbuti.DG). Z scores with 2<|Z|<6 were labeled as "+/–," significant Z scores(|Z| > 6) were labeled as "++/––."

Supplementary Figure S3 | The pairwise f4 statistics in forms of f4 (studied individual1, studied individual2; Han_NChina, Mbuti.DG). Z scores with 2<|Z|<6 were labeled as "+/–," significant Z scores (|Z| > 6) were labeled as "++/––".

Supplementary Figure S4 | The pairwise f4 statistics in forms of f4 (studied individual1, studied individual2; Tibetan_Chamdo, Mbuti.DG). Z scores with 2<|Z|<6 were labeled as "+/–"; significant Z scores(|Z|>6) were labeled as "++/––".

Supplementary Figure S5 | The pairwise f4 statistics in forms of f4 (studied individual1, studied individual2; Ulchi, Mbuti.DG). Z scores with 2<|Z|<6 were labeled as "+/–"; significant Z scores(|Z|>6) were labeled as "++/––".

Supplementary Figure S6 | The pairwise f4 statistics in forms of f4 (studied individual1, studied individual2; Mongol, Mbuti.DG). Z scores with 2<|Z|<6 were labeled as "+/–"; significant Z scores(|Z| > 6) were labeled as "++/––".

Supplementary Figure S7 | The pairwise f4 statistics in forms of f4 (studied individual1, studied individual2; Mongola, Mbuti.DG). Z scores with 2<|Z|<6 were labeled as "+/–"; significant Z scores(|Z|>6) were labeled as "++/––".

Supplementary Figure S8 | The heatmap of pairwise IBD (identified by descent) segments among Mongolian individuals.

Supplementary Figure S9 | The shared genetic drift between ancient Eurasian populations and Mongolian subgroups.

Supplementary Figure S10 | The genetic distance (Fst) based on smartpca showed the genetic difference in Eastern Eurasian among Mongolian subgroups.

Supplementary Figure S11 | The phylogenetic relationships between the studied Mongolian populations and modern Eurasian populations based on Treemix. **(A)**. TreeMix based on relative genetic drift showed the polygenetic relationship among global populations and four-gene flow events. One gene flow occurs between Ulchi and Mongolian speaker Buryat. **(B)**. TreeMix-based phylogenetic tree including fewer references showed one western gene influx flow into Mongol.

Supplementary Figure S12 | The best-fitted qpGraph-based deep population admixture history of Mongolian subgroups.

Supplementary Figure S13 | PCA patterns based on the coancestry matrix of linked SNP markers.

Supplementary Figure S14 | The GO enrichment analysis of Mongolian subgroups.

REFERENCES

Agranat-Tamir, L., Waldman, S., Martin, M. A. S., Gokhman, D., Mishol, N., Eshel, T., et al. (2020). The Genomic History of the Bronze Age Southern Levant. *Cell* 181 (5), 1146–1157.e1111. doi:10.1016/j.cell.2020.04.024

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast Model-Based Estimation of Ancestry in Unrelated Individuals. *Genome Res.* 19 (9), 1655–1664. doi:10.1101/gr.094052.109

Allentoft, M. E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., et al. (2015). Population Genomics of Bronze Age Eurasia. *Nature* 522 (7555), 167–172. doi:10.1038/nature14507

Bai, H., Guo, X., Narisu, N., Lan, T., Wu, Q., Xing, Y., et al. (2018). Whole-genome Sequencing of 175 Mongolians Uncovers Population-specific Genetic Architecture and Gene Flow throughout North and East Asia. *Nat. Genet.* 50 (12), 1696–1704. doi:10.1038/s41588-018-0250-5

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. (2021). Genomic Insights into the Admixture History of Mongolic- and Tungusic-Speaking Populations from Southwestern East Asia. *Front. Genet.* 12 (880). doi:10.3389/fgene.2021.685285

Damgaard, P. d. B., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., Korneliussen, T., et al. (2018). 137 Ancient Human Genomes from across the Eurasian Steppes. *Nature* 557 (7705), 369–374. doi:10.1038/s41586-018-0094-2

de Barros Damgaard, P., Martiniano, R., Kamm, J., Moreno-Mayar, J. V., Kroonen, G., Peyrot, M., et al. (2018). The First Horse Herders and the Impact of Early Bronze Age Steppe Expansions into Asia. *Science* 360 (6396), eaar7711. doi:10.1126/science.aar7711

Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved Whole-Chromosome Phasing for Disease and Population Genetic Studies. *Nat. Methods* 10 (1), 5–6. doi:10.1038/nmeth.2307

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 (1), 78–90. doi:10.1111/1755-0998.12634

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., et al. (2015). Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe. *Nature* 522 (7555), 207–211. doi:10.1038/nature14317

He, G., Adnan, A., Rakha, A., Yeh, H.-Y., Wang, M., Zou, X., et al. (2019). A Comprehensive Exploration of the Genetic Legacy and Forensic Features of Afghanistan and Pakistan Mongolian-descent Hazara. *Forensic Sci. Int. GeneticsGenetics* 42, e1–e12. doi:10.1016/j.fsigen.2019.06.018

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 (6172), 747–751. doi:10.1126/science.1243518

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists. *Nucleic Acids Res.* 37 (1), 1–13. doi:10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* 4 (1), 44–57. doi:10.1038/nprot.2008.211

Huang, Y.-Z., Pamjav, H., Flegontov, P., Stenzl, V., Wen, S.-Q., Tong, X.-Z., et al. (2018a). Dispersals of the Siberian Y-Chromosome Haplogroup Q in Eurasia. *Mol. Genet. Genomics* 293 (1), 107–117. doi:10.1007/s00438-017-1363-8

Huang, Y.-Z., Wei, L.-H., Yan, S., Wen, S.-Q., Wang, C.-C., Yang, Y.-J., et al. (2018b). Whole Sequence Analysis Indicates a Recent Southern Origin of Mongolian Y-Chromosome C2c1a1a1-M407. *Mol. Genet. Genomics* 293 (3), 657–663. doi:10.1007/s00438-017-1403-4

Jeong, C., Balanovsky, O., Lukianova, E., Kahbatkyzy, N., Flegontov, P., Zaporozhchenko, V., et al. (2019). The Genetic History of Admixture across Inner Eurasia. *Nat. Ecol. Evol.* 3 (6), 966–976. doi:10.1038/s41559-019-0878-2

Jeong, C., Wang, K., Wilkin, S., Taylor, W. T. T., Miller, B. K., Bemmann, J. H., et al. (2020). A Dynamic 6,000-Year Genetic History of Eurasia's Eastern Steppe. *Cell* 183 (4), 890–904.e29. doi:10.1016/j.cell.2020.10.015

Jeong, C., Wilkin, S., Amgalantugs, T., Bouwman, A. S., Taylor, W. T. T., Hagan, R. W., et al. (2018). Bronze Age Population Dynamics and the Rise of Dairy Pastoralism on the Eastern Eurasian Steppe. *Proc. Natl. Acad. Sci. USA* 115 (48), E11248–e11255. doi:10.1073/pnas.1813608115

Kamberov, Y. G., Wang, S., Tan, J., Gerbault, P., Wark, A., Tan, L., et al. (2013). Modeling Recent Human Evolution in Mice by Expression of a Selected EDAR Variant. *Cell* 152 (4), 691–702. doi:10.1016/j.cell.2013.01.016

Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of Population Structure Using Dense Haplotype Data. *Plos Genet.* 8 (1), e1002453. doi:10.1371/journal.pgen.1002453

Li, H., Gu, S., Cai, X., Speed, W. C., Pakstis, A. J., Golub, E. I., et al. (2008). Ethnic Related Selection for an ADH Class I Variant within East Asia. *PLoS One* 3 (4), e1881. doi:10.1371/journal.pone.0001881

Li, H., Gu, S., Han, Y., Xu, Z., Pakstis, A. J., Jin, L., et al. (2011). Diversification of the ADH1B Gene during Expansion of Modern Humans. *Ann. Hum. Genet.* 75 (4), 497–507. doi:10.1111/j.1469-1809.2011.00651.x

Liu, Y., Yang, J., Li, Y., Tang, R., Yuan, D., Wang, Y., et al. (2021). 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 (834). doi:10.3389/fgene.2021.626710

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 (4), 1233–1254. doi:10.1534/genetics.112.147330

Lu, Y., Quan, C., Chen, H., Bo, X., and Zhang, C. (2017). 3DSNP: a Database for Linking Human Noncoding SNPs to Their Three-Dimensional Interacting Genes. *Nucleic Acids Res.* 45 (D1), D643–d649. doi:10.1093/nar/gkw1022

Ma, L., Xue, Y., Liu, Y., Wang, Z., Cui, X., Li, P., et al. (20052005). Polymorphism Study of Seven SNPs at ADH Genes in 15 Chinese Populations. *Hereditas* 142, 103–111. doi:10.1111/j.1601-5223.2005.01910.x

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust Relationship Inference in Genome-wide Association Studies. *Bioinformatics* 26 (22), 2867–2873. doi:10.1093/bioinformatics/btq559

Mathieson, I., Alpaslan-Roodenberg, S., Posth, C., Szécsényi-Nagy, A., Rohland, N., Mallick, S., et al. (2018). The Genomic History of southeastern Europe. *Nature* 555 (7695), 197–203. doi:10.1038/nature25778

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., et al. (2015). Genome-wide Patterns of Selection in 230 Ancient Eurasians. *Nature* 528 (7583), 499–503. doi:10.1038/nature16152

Nakayama, K., Bayasgalan, T., Bayasgalan, T., Tazoe, F., Yanagisawa, Y., Gotoh, T., et al. (2010). A Single Nucleotide Polymorphism in the FADS1/FADS2 Gene Is Associated with Plasma Lipid Profiles in Two Genetically Similar Asian Ethnic Groups with Distinctive Differences in Lifestyle. *Hum. Genet.* 127 (6), 685–690. doi:10.1007/s00439-010-0815-6

Narasimhan, V. M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., Mallick, S., et al. (2019). The Formation of Human Populations in South and Central Asia. *Science* 365 (6457), eaat7487. doi:10.1126/science.aat7487

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 (1), 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 (15), 2526–2532.e2524. doi:10.1016/j.cub.2019.06.044

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., et al. (2012). Ancient Admixture in Human History. *Genetics* 192 (3), 1065–1093. doi:10.1534/genetics.112.145037

Patterson, N., Price, A. L., and Reich, D. (2006). Population Structure and Eigenanalysis. *Plos Genet.* 2 (12), 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 (11), e1002967. doi:10.1371/journal.pgen.1002967

Pugach, I., Matveev, R., Spitsyn, V., Makarov, S., Novgorodov, I., Osakovsky, V., et al. (2016). The Complex Admixture History and Recent Southern Origins of Siberian Populations. *Mol. Biol. Evol.* 33 (7), 1777–1795. doi:10.1093/molbev/msw055

Schaeffer, L., Gohlke, H., Müller, M., Heid, I. M., Palmer, L. J., Kompauer, I., et al. (2006). Common Genetic Variants of the FADS1 FADS2 Gene Cluster and Their Reconstructed Haplotypes Are Associated with the Fatty Acid Composition in Phospholipids. *Hum. Mol. Genet.* 15 (11), 1745–1756. doi:10.1093/hmg/ddl117

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 (2), e1601877. doi:10.1126/sciadv.1601877

Song, Z., Cao, H., Qin, L., and Jiang, Y. (2013). A Case-Control Study between Gene Polymorphisms of Polyunsaturated Fatty Acid Metabolic Rate-Limiting Enzymes and Acute Coronary Syndrome in Chinese Han Population. *Biomed. Res. Int.* 2013, 1–7. doi:10.1155/2013/928178

Wang, C.-C., Reinhold, S., Kalmykov, A., Wissgott, A., Brandt, G., Jeong, C., et al. (2019). Ancient Human Genome-wide Data from a 3000-year Interval in the Caucasus Corresponds with Eco-Geographic Regions. *Nat. Commun.* 10 (1), 590. doi:10.1038/s41467-018-08220-8

Wang, C.-C., Yeh, H.-Y., Popov, A. N., Zhang, H.-Q., Matsumura, H., Sirak, K., et al. (2021). Genomic Insights into the Formation of Human Populations in East Asia. *Nature* 591 (7850), 413–419. doi:10.1038/s41586-021-03336-2

Wei, L.-H., Wang, L.-X., Wen, S.-Q., Yan, S., Canada, R., Gurianov, V., et al. (2018a). Paternal Origin of Paleo-Indians in Siberia: Insights from Y-Chromosome Sequences. *Eur. J. Hum. Genet.* 26 (11), 1687–1696. doi:10.1038/s41431-018-0211-6

Wei, L.-H., Yan, S., Lu, Y., Wen, S.-Q., Huang, Y.-Z., Wang, L.-X., et al. (2018b). Whole-sequence Analysis Indicates that the Y Chromosome C2*-Star Cluster Traces Back to Ordinary Mongols, rather Than Genghis Khan. *Eur. J. Hum. Genet.* 26 (2), 230–237. doi:10.1038/s41431-017-0012-3

Wen, S.-Q., Yao, H.-B., Du, P.-X., Wei, L.-H., Tong, X.-Z., Wang, L.-X., et al. (2019). Molecular Genealogy of Tusi Lu's Family Reveals Their Paternal Relationship with Jochi, Genghis Khan's Eldest Son. *J. Hum. Genet.* 64 (8), 815–820. doi:10.1038/s10038-019-0618-0

Wilkin, S., Ventresca Miller, A., Taylor, W. T. T., Miller, B. K., Hagan, R. W., Bleasdale, M., et al. (2020). Dairy Pastoralism Sustained Eastern Eurasian Steppe Populations for 5,000 Years. *Nat. Ecol. Evol.* 4 (3), 346–355. doi:10.1038/s41559-020-1120-y

Wu, Y., Zeng, L., Chen, X., Xu, Y., Ye, L., Qin, L., et al. (2017). Association of the FADS Gene Cluster with Coronary Artery Disease and Plasma Lipid Concentrations in the Northern Chinese Han Population. *Prostaglandins, Leukot. Essent. Fatty Acids* 117, 11–16. doi:10.1016/j.plefa.2017.01.014

Yan, S., Tachibana, H., Wei, L.-H., Yu, G., Wen, S.-Q., and Wang, C.-C. (2015). Y Chromosome of Aisin Gioro, the imperial House of the Qing Dynasty. *J. Hum. Genet.* 60 (6), 295–298. doi:10.1038/jhg.2015.28

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 (6501), 282–288. doi:10.1126/science.aba0909

Yao, Y.-G., Kong, Q. P., Wang, C. Y., Zhu, C. L., and Zhang, Y. P. (2004). Different Matrilineal Contributions to Genetic Structure of Ethnic Groups in the Silk

Road Region in china. *Mol. Biol. Evol.* 21 (12), 2265–2280. doi:10.1093/molbev/msh238

Yunusbayev, B., Metspalu, M., Metspalu, E., Valeev, A., Litvinov, S., Valiev, R., et al. (2015). The Genetic Legacy of the Expansion of Turkic-Speaking Nomads across Eurasia. *Plos Genet.* 11 (4), e1005068. doi:10.1371/journal.pgen.1005068

Zhang, X., He, G., Li, W., Wang, Y., Li, X., Chen, Y., et al. (2021). Genomic Insight into the Population Admixture History of Tungusic-Speaking Manchu People in Northeast China. *Front. Genet.* 12, 754492. doi:10.3389/fgene.2021.754492

Zhao, J., WurigemuleSun, J., Sun, J., Xia, Z., He, G., Yang, X., et al. (2020). Genetic Substructure and Admixture of Mongolians and Kazakhs Inferred from Genome-wide Array Genotyping. *Ann. Hum. Biol.* 47 (7-8), 620–628. doi:10.1080/03014460.2020.1837952

Zheng, H., Cong, P., Bai, W., Li, J., Li, N., Gai, S., et al. (2021). Genomic analyses of 10,376 individuals provides comprehensive map of genetic variations, structure and reference haplotypes for Chinese population. *Research Square*. doi:10.21203/rs.3.rs-184446/v1

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.

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.

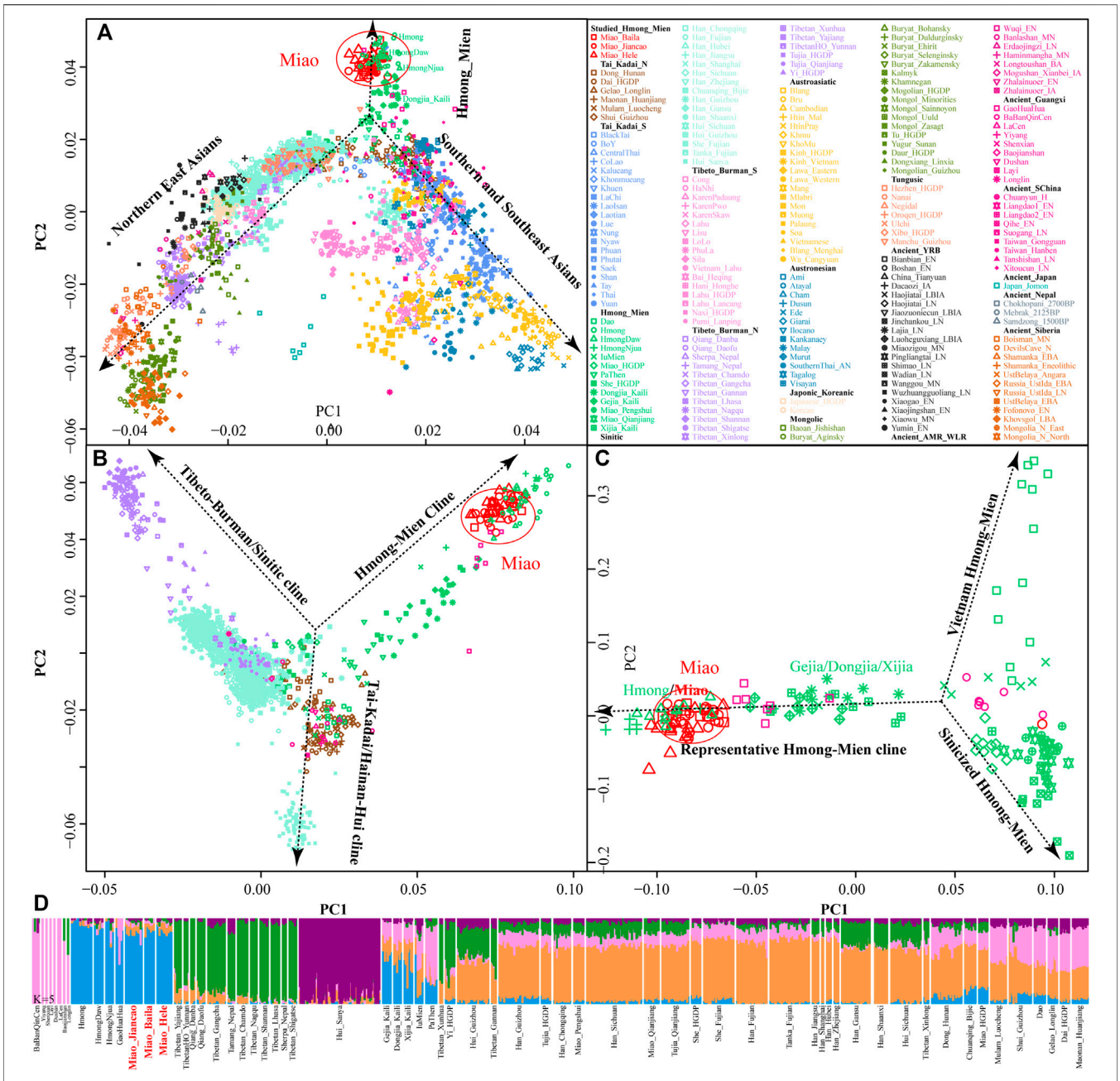

FIGURE 1 | Genetic affinity of HM people in the context of modern and ancient eastern Eurasians. **(A–C)**, Principal component analyses focused on the genetic diversity from East Asian, South Chinese populations, and HM-speaking populations. Included East Asian ancient populations were projected onto 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.

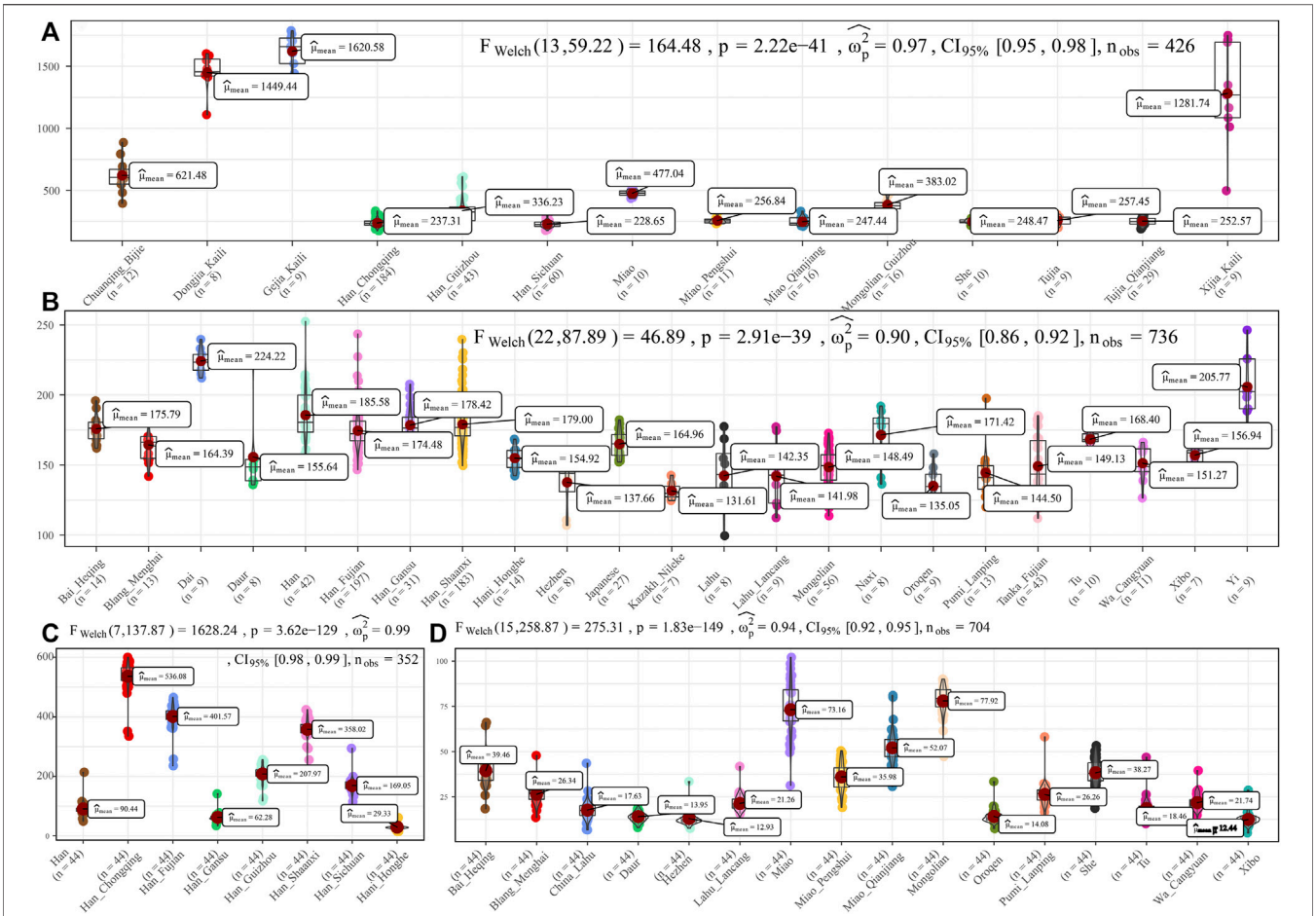

FIGURE 2 | Chromosome painting between Hmong–Mien people and other East Asian reference populations. **(A,B)** Amount of total length of DNA fragments of modern East Asians copied from donor chromosomes related to Sichuan Miao. **(C,D)** Average DNA chunk of Sichuan Miao copied from other East Asians. 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

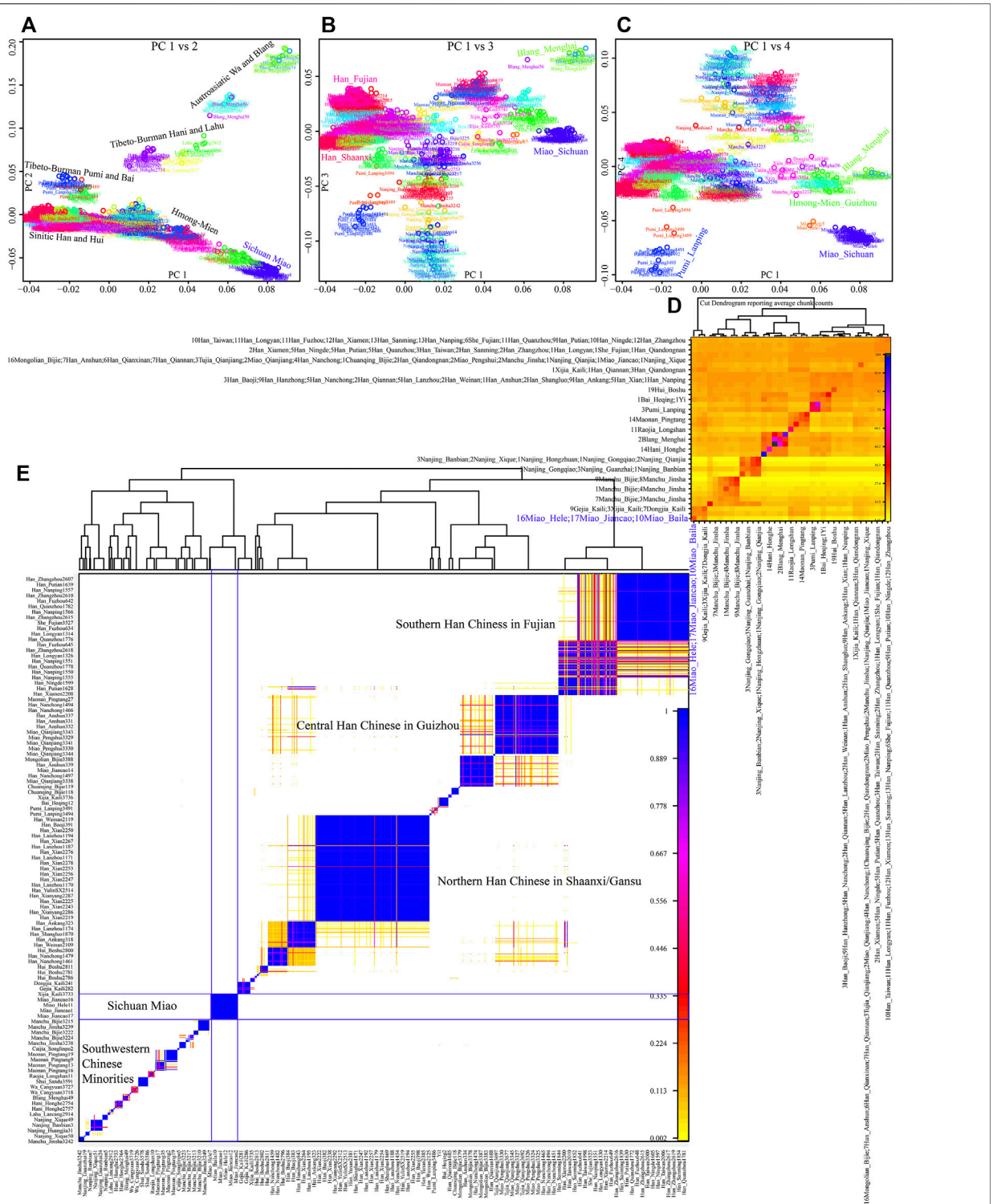

FIGURE 3 | Fine-scale population genetic structure based on the shared haplotype data. **(A–C)** PCA results based on the coancestry matrix showed a genetic relationship among modern East Asians. The color showed the re-classification of the homogenous population label. **(D,E)** Clustering patterns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

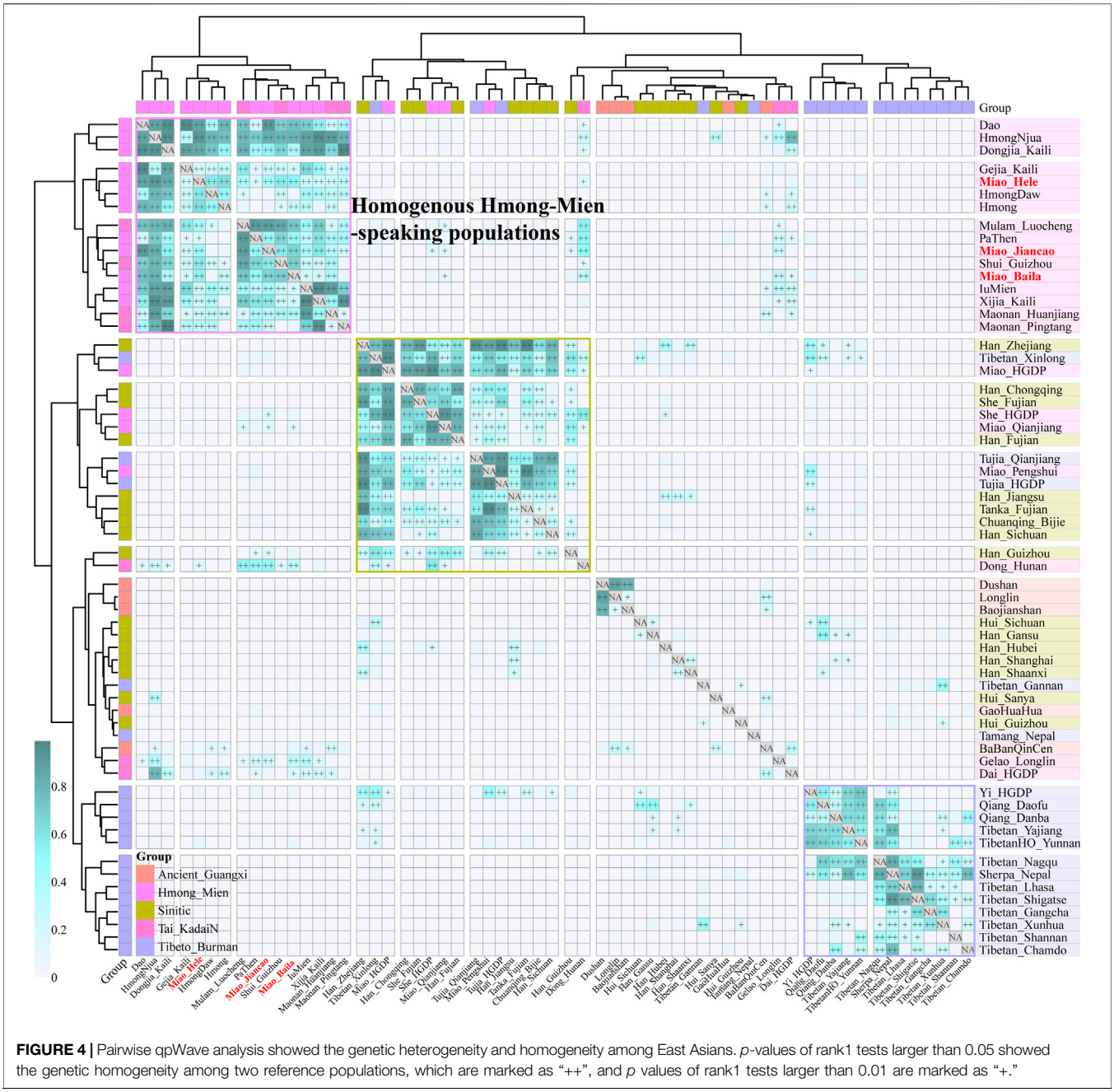

FIGURE 4 | Pairwise qpWave analysis showed the genetic heterogeneity and homogeneity among East Asians. *p*-values of rank1 tests larger than 0.05 showed the genetic homogeneity among two reference populations, which are marked as "++", and *p* values of rank1 tests larger than 0.01 are marked as "+."

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.

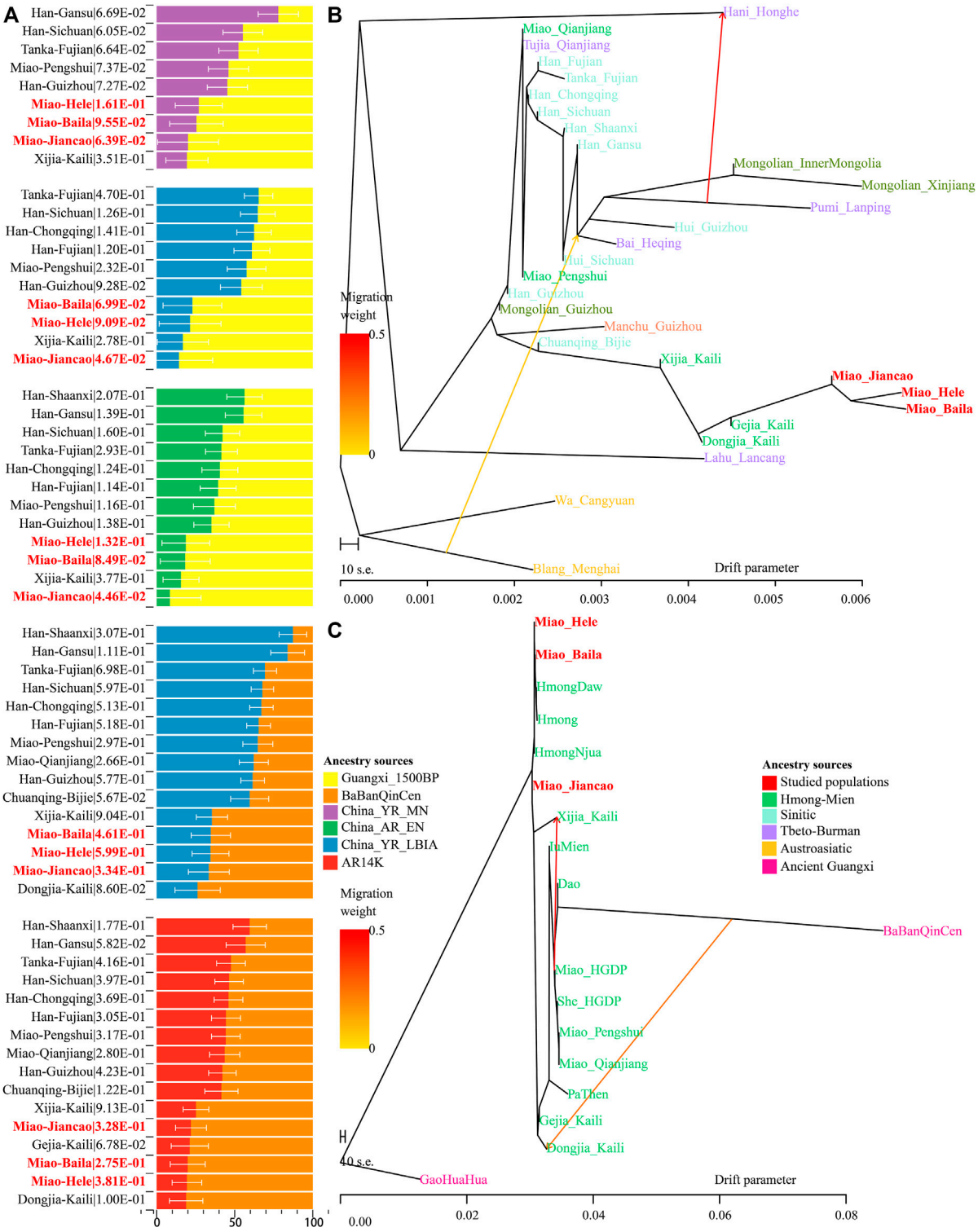

FIGURE 5 | Results of qpAdm models and TreeMix-based phylogenies. **(A)** Two-way admixture models showed ancestry comparison in different ancestral source pairs. **(B,C)** TreeMix-based phylogenetic tree with two migration events showed the genetic relationship between East Asians.

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

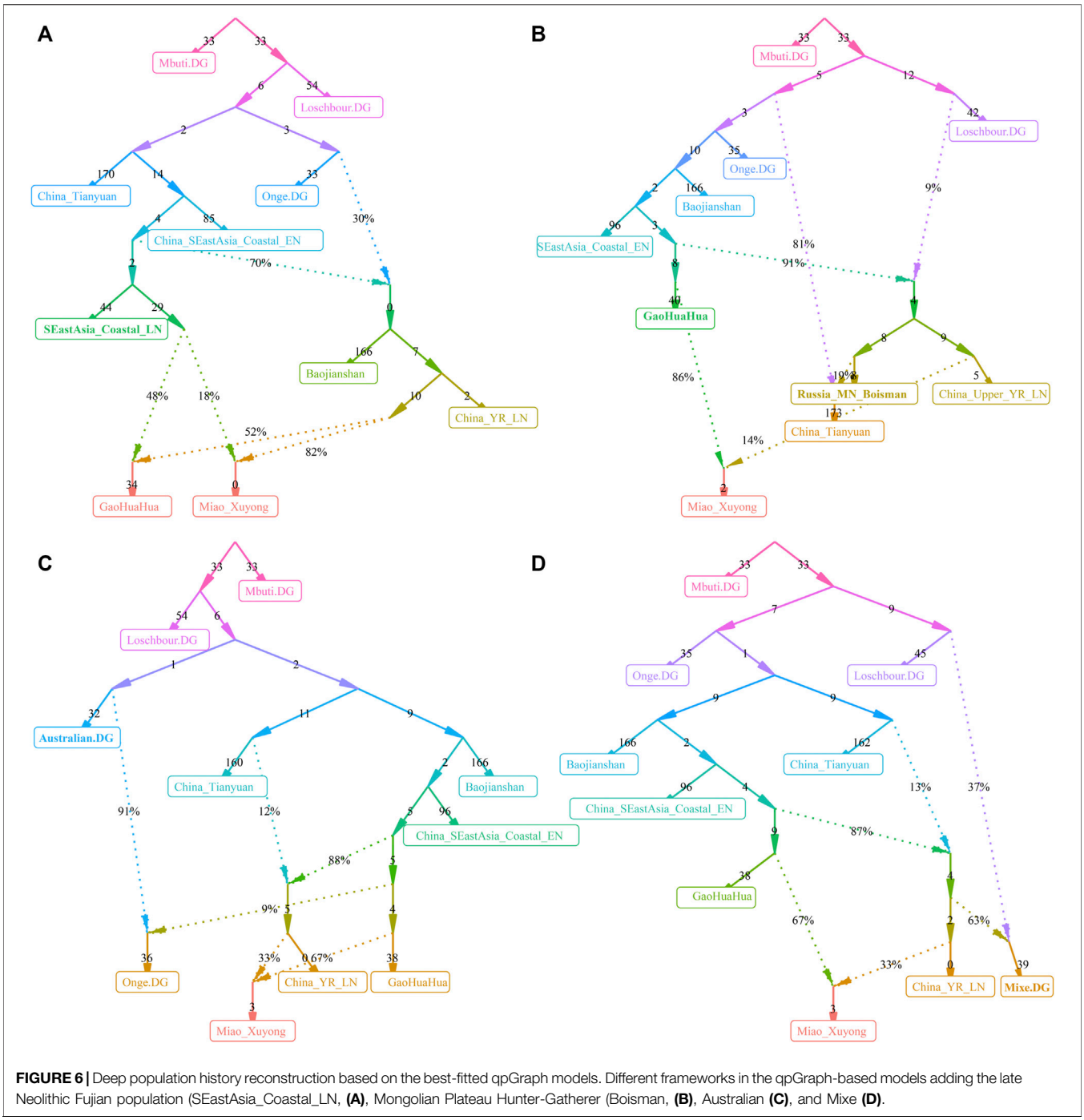

FIGURE 6 | Deep population history reconstruction based on the best-fitted qpGraph models. Different frameworks in the qpGraph-based models adding the late Neolithic Fujian population (SEastAsia_Coastal_LN, **(A)**, Mongolian Plateau Hunter-Gatherer (Boisman, **(B)**, Australian **(C)**, and Mixe **(D)**.

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*

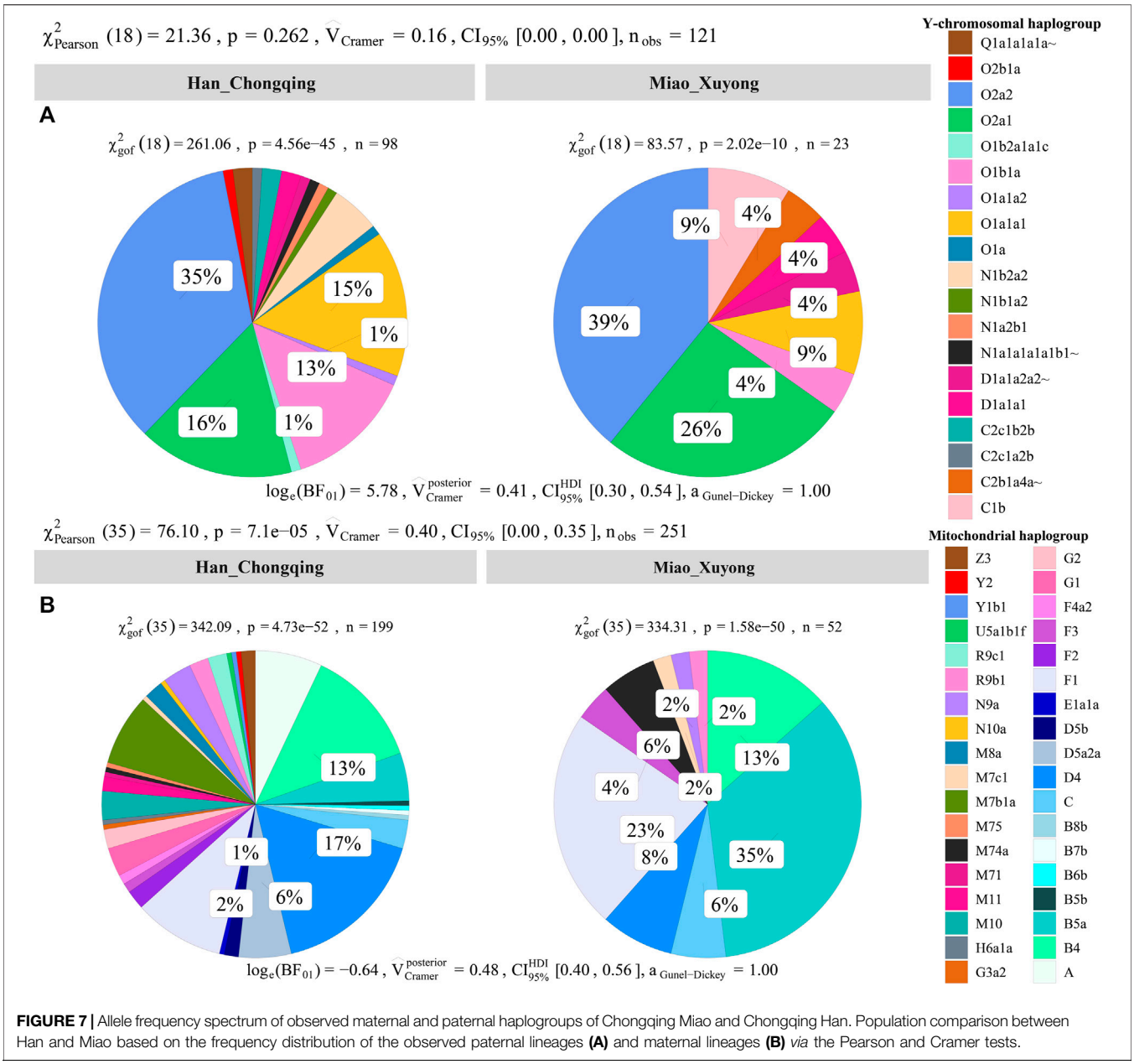

FIGURE 7 | Allele frequency spectrum of observed maternal and paternal haplogroups of Chongqing Miao and Chongqing Han. Population comparison between Han and Miao based on the frequency distribution of the observed paternal lineages **(A)** and maternal lineages **(B)** *via* the Pearson and Cramer tests.

(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

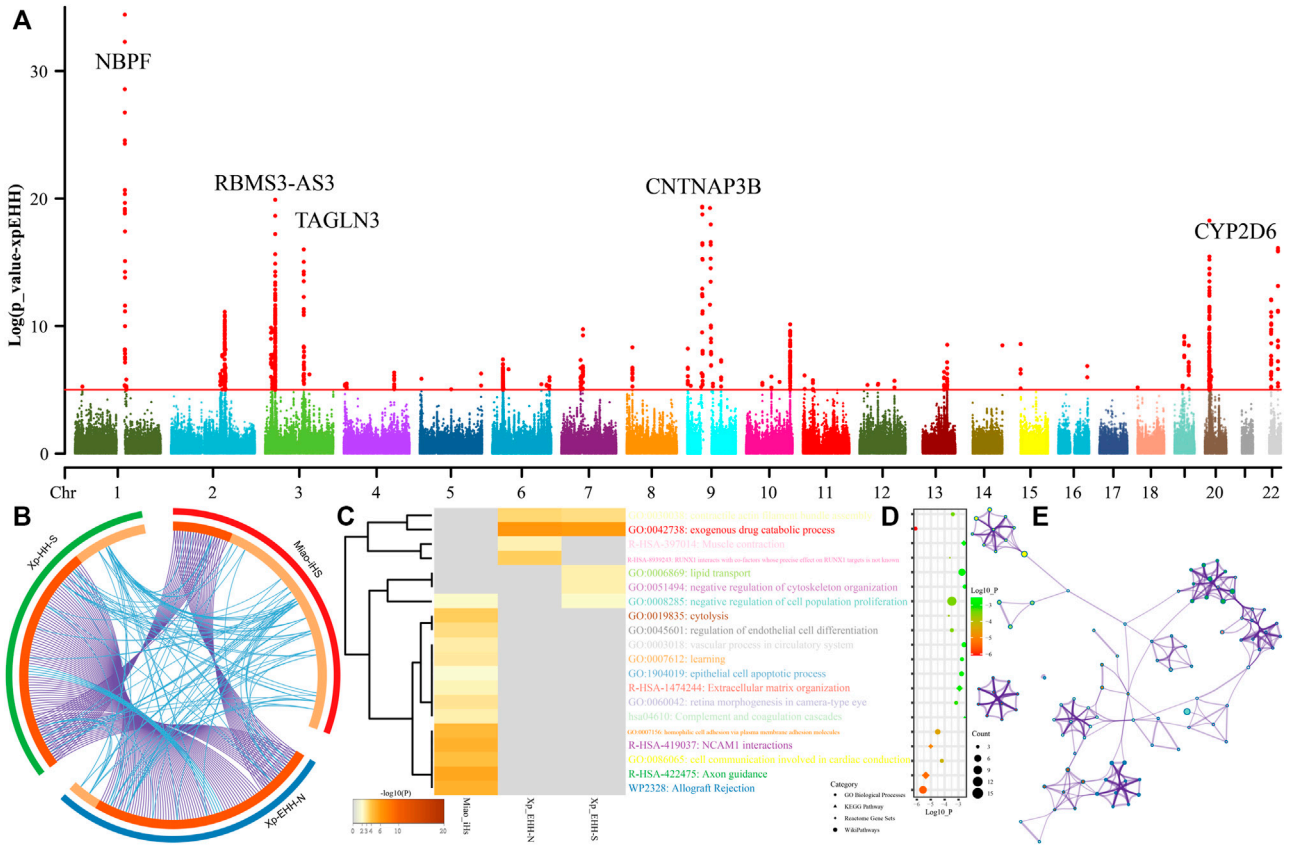

FIGURE 8 | Manhattan showed the natural selection signatures and enrichment analysis. **(A)** *p*-values of XPEHH in the Miao population using northern Han as the reference population. **(B)** Overlap among three gene lists based on gene-level and shared-term level, where blue curves link genes that 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 January 2022
doi: 10.3389/fgene.2021.805936

Genetic Structure and Forensic Feature of 38 X-Chromosome InDels in the Henan Han Chinese Population

Lin Zhang[1,2†], Zhendong Zhu[3†], Weian Du[4], Shengbin Li[1] and Changhui Liu[5]**

[1]Bio-evidence Science Academy, Xi'an Jiaotong University, Xi'an, China, [2]Department of Forensic Medicine, Xinxiang Medical University, Xinxiang, China, [3]Department of Human Anatomy, School of Basic Medical Sciences, Xinxiang Medical University, Xinxiang, China, [4]HOMY GeneTech Incorporation, Foshan, China, [5]Guangzhou Forensic Science Institute, Guangzhou, China

Insertion/deletion (InDel) polymorphisms, as ideal forensic markers, show useful characteristics of both SNPs and STRs, such as low mutation rate, short amplicon size and general applicability of genotyping platform, and have been used in human identification, population genetics and biogeographic research in recent years. X-chromosome genetic markers are significant in population genetic studies and indispensable complements in some complex forensic cases. However, the population genetic studies of X-chromosome InDel polymorphisms (X-InDels) still need to be explored. In this study, the forensic utility of a novel panel including 38 X-InDel markers was evaluated in a sample of Han population from Henan province in China. It is observed that the heterozygosities ranged from 0.0054 to 0.6133, and the combined discrimination power was 1–9.18×10^{-17} for males and 1–7.22×10^{-12} for females respectively. The mean exclusion chance in trios and duos were 0.999999319 and 0.999802969 respectively. Multiple biostatistics methods, such as principal component analysis, genetic distances analysis, phylogenetic reconstruction, and structure analysis was used to reveal the genetic relationships among the studied Henan Han group and other 26 reference groups from 1,000 Genomes Project. As expected, the Henan Han population was clustered with East Asian populations, and the most intimate genetic relationships existed in three Han Chinese populations from Henan, Beijing and South China, and showed significant differences compared with other continental groups. These results confirmed the suitability of the 38 X-InDel markers both in individual identification and parentage testing in Han Chinese population, and simultaneously showed the potential application in population genetics.

Keywords: Henan Han population, forensic features, population genetics, InDel, X-chromosome

OPEN ACCESS

Edited by:
Chuan-Chao Wang,
Xiamen University, China

Reviewed by:
Pingming Qiu,
Southern Medical University, China
Jun Yao,
China Medical University, China
Shujin Li,
Hebei Medical University, China
Hongyu Sun,
Sun Yat-sen University, China

****Correspondence:***
Shengbin Li
shbinlee@xjtu.edu.cn
Changhui Liu
68120968@qq.com

[†]These authors have contributed equally to this work

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

Received: *31 October 2021*
Accepted: *25 November 2021*
Published: *03 January 2022*

Citation:
Zhang L, Zhu Z, Du W, Li S and Liu C (2022) Genetic Structure and Forensic Feature of 38 X-Chromosome InDels in the Henan Han Chinese Population. Front. Genet. 12:805936. doi: 10.3389/fgene.2021.805936

INTRODUCTION

Forensic genetic research is devoted to the pursuit of ideal genetic markers that can be used to the routine forensic casework including individual identification, paternity and kinship testing, race and species identification, biological ancestry analysis et al. Short tandem repeat markers (STRs) have been widely used in routine forensic casework. However, due to the defects of high spontaneous mutation rate, the limitation of detecting degraded samples, stutter peak et al., STRs occasionally produced inconclusive results, especially in some deficient paternity cases or some cases with only degraded tissue as DNA source (Huang et al., 2002; Fan and Chu, 2007; Giardina et al., 2011; Caputo et al., 2017). To overcome the abovementioned defects of STRs, new genetic markers have been

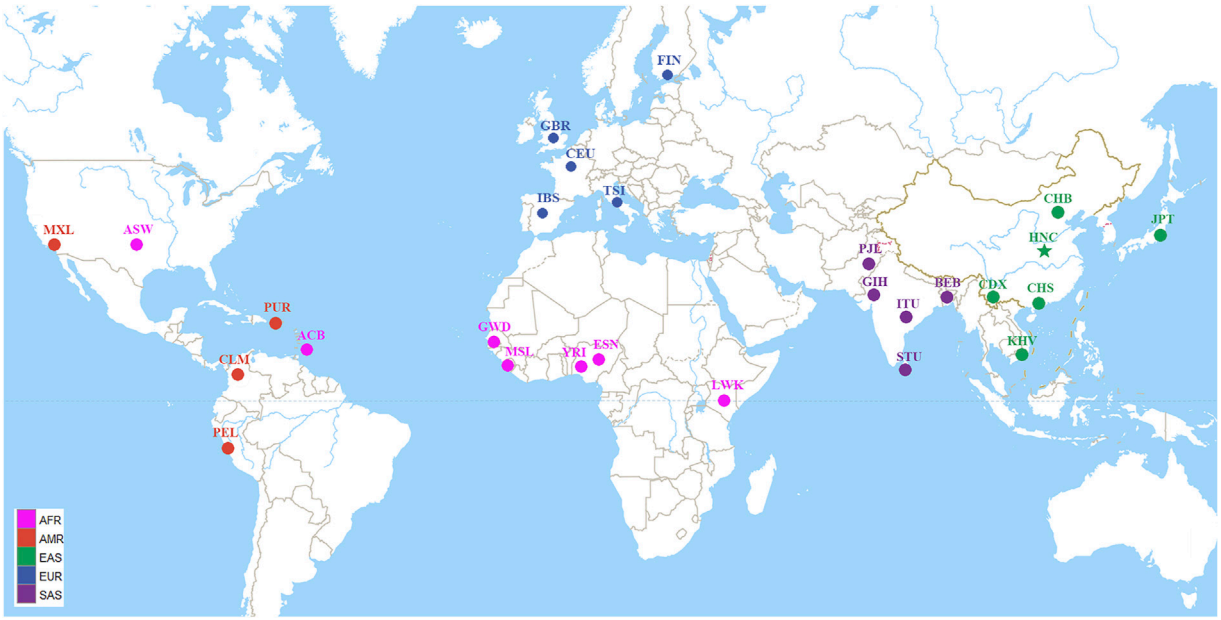

FIGURE 1 | Geographic location information of the studied population and other 26 worldwide reference populations. The 27 populations were divided into five continental groups: African (AFR) including African ancestry in Southwest United States (ASW), African Caribbean in Barbados (ACB), Esan in Nigeria (ESN), Gambian in Western Division-Mandinka (GWD), Luhya in Webuye, Kenya (LWK), Mende in Sierra Leone (MSL), Yoruba in Ibadan, Nigeria (YRI); American (AMR) including Colombian in Medellín, Colombia (CLM), Mexican Ancestry in Los Angeles CA United States (MXL), Peruvian in Lima, Peru (PEL), Puerto Rican in Puerto Rico (PUR); East Asian (EAS) including Chinese Dai in Xishuangbanna (CDX), Han Chinese in Beijing, China (CHB), Han Chinese in South China (CHS), Henan Han Chinese (HNC), Japanese in Tokyo, Japan (JPT), Kinh in Ho Chi Minh City, Vietnam (KHV); European (EUR) including Utah residents with Northern and Western European ancestry (CEU), Finnish in Finland (FIN), British from England and Scotland (GBR), Iberian populations in Spain (IBS), Toscani in Italia (TSI); South Asian (SAS) including Bengali in Bangladesh (BEB), Gujarati Indians in Houston, Texas, United States (GIH), Indian Telugu in the United Kingdom (ITU), Punjabi in Lahore, Pakistan (PJL), Sri Lankan Tamil in the United Kingdom (STU).

searching by scholars worldwide in recent years. As a kind of special dimorphic marker, Insertion/Deletion polymorphisms (InDels) are generated in natural populations by inserting or deleting DNA fragments of different sizes and widely distributed throughout the genome (Weber et al., 2002; Mills et al., 2006). Although InDel markers have fewer genetic diversity than STRs, they show lower mutation rates and are more suitable for detecting degraded DNA samples because analysis can be based on short amplicons which increases the chance to generate more accurate data with high resolution (Gomes C. et al., 2020). Furthermore, InDel polymorphisms are length polymorphism markers which can be easily separated in capillary electrophoresis technology. Hence, InDels are considered to be ideal forensic markers and have attracted attentions from forensic researchers worldwide (Sheng et al., 2018). InDels have been increasingly explored and used in forensic genetics and biogeographical ancestry inferences, and much more extensive forensic and population genetic studies are ongoing (Bastos-Rodrigues et al., 2006; Larue et al., 2014; Du et al., 2019; He et al., 2019).

X-chromosome markers recombine along the whole chromosome during female meiosis in females and are transmitted to both female and male descendants, however, are entirely transmitted to female offspring in males (Gomes I. et al., 2020). Due to their distinctive transmission properties, X-chromosome markers have emerged as a powerful complementary tool of parentage testing, especially in some special parentage cases such as “half-siblings”, “avuncular” or “grandparent-grandchild” (Gomes et al., 2012). In addition, for the smaller effective population size, the X-chromosome shows faster genetic drift than autosomes, hence genetic distances between populations are significantly larger on the X chromosome (Schaffner, 2004). These specific properties of X-chromosome makes it a good system for assessing admixture of population and investigating evolutionary anthropology.

Han Chinese who makes up 91.50% of Chinese population has shown the regional genetic diversities because of national amalgamation in history (Lin et al., 2017; Cao et al., 2020). Han Chinese originated from Huaxia tribes or the Neolithic Yan Huang tribes along the upper and middle Yellow River in northern China (Wen et al., 2004). As the center birthplace of Huaxia civilization, Henan province, with a total population of 99.37 million, of which 98.84% are Han Chinese, enjoys the representative group of Han Chinese in China. In this study, a novel multiplex amplification system including 38 X-InDel loci was tested in Henan Han Chinese population. We firstly genotyped 38 X-InDel loci in 268 individuals of Han Chinese from Henan province and calculated the forensic parameters to provide basic group data for paternity identification and individual identification. Moreover, using the 38 X-InDel loci, we analyzed the population genetic differentiations between

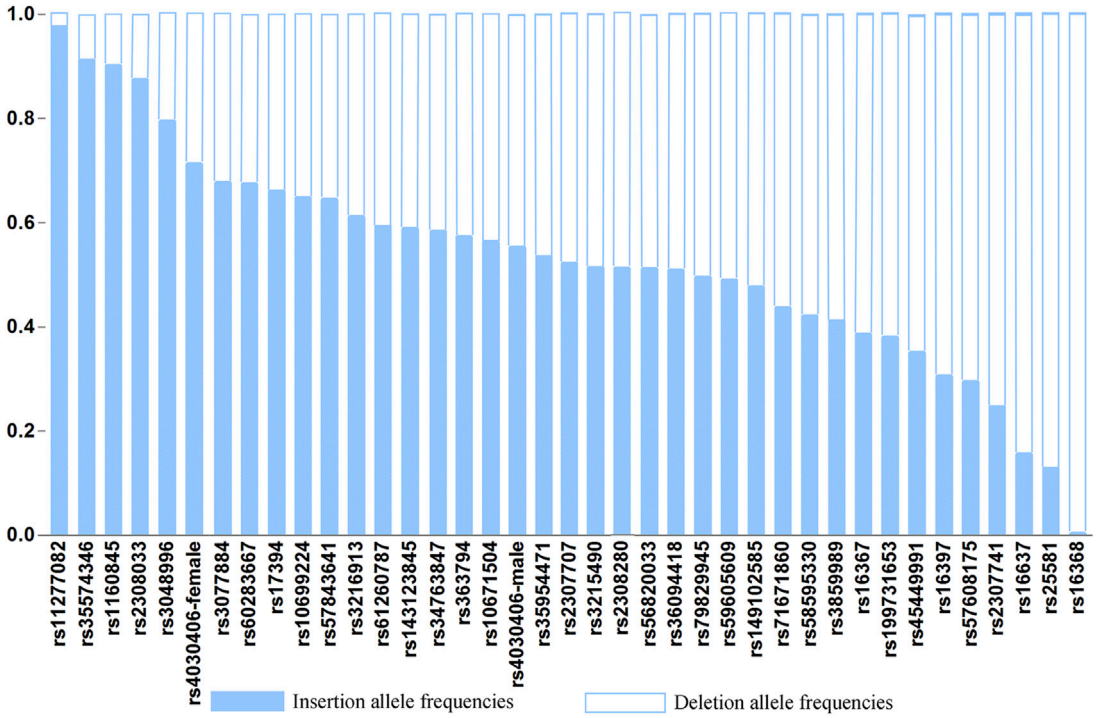

FIGURE 2 | Histogram of allele frequencies at the 38 X-InDel loci of the Henan Han population.

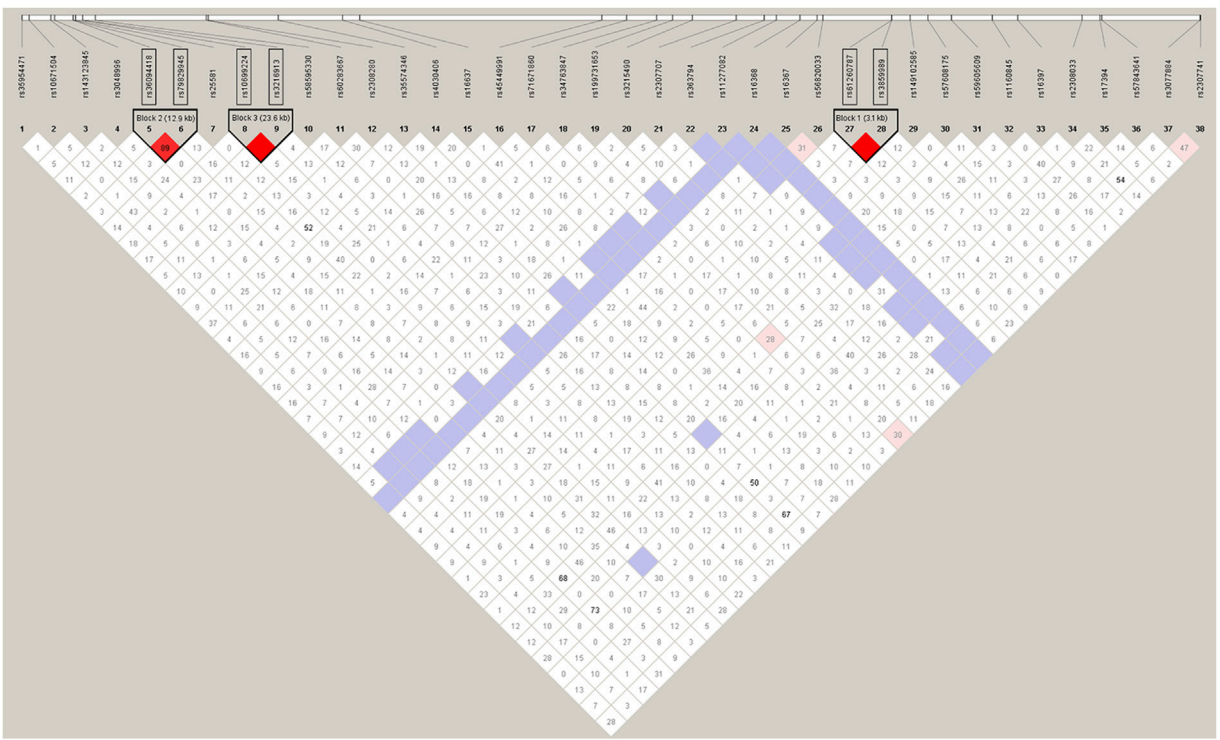

FIGURE 3 | The degree of LD among the 38 InDel loci in Henan Han population. The red color indicates a high level of linkage between two loci.

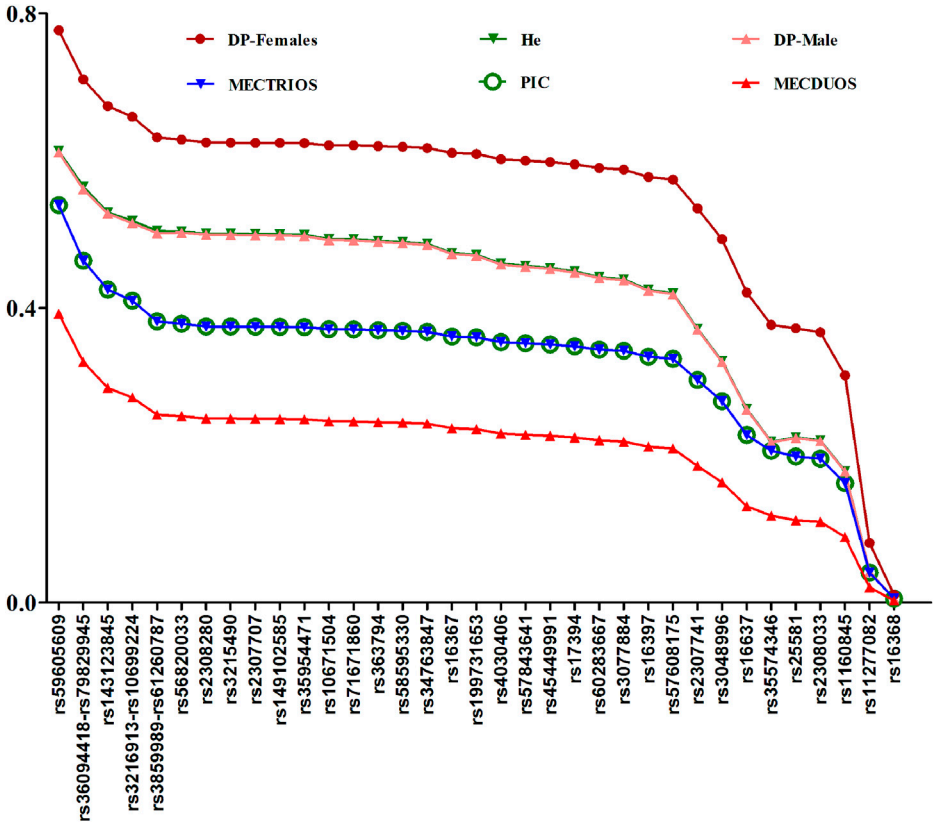

FIGURE 4 | Plots of forensic parameters at the 38 InDel loci in Henan Han population.

Henan Han population and other 26 reference populations from 1,000 Genomes Project.

MATERIALS AND METHODS

Sample Collection

Our study was approved by the Ethics Committee of Xi'an Jiaotong University Health Science Center (No.2021-1,444). Each volunteer signed an informed consent, giving his permission to the analysis and publishing of anonymized genetic data obtained from his biological sample. A total of 268 fingertip blood samples were gathered from unrelated individuals (106 females and 162 males) of Chinese Han population in Henan province (HNC). The authentication was performed to ensure those volunteers are indigenous people of Henan province. All the volunteers meet the conditions that the past three generations are Han group and have no trans-regional migration.

The blood samples were placed in the sample acquisition card and then were dried and stored at room temperature. Meanwhile, we collected 38 InDels raw data and allele frequencies of 26 populations worldwide spanning five continents from the 1,000 Genomes Project (Ensembl Genome Browser, http://grch37.ensembl.org/Homo_sapiens/Info/Index), as shown in **Figure 1** (The 1000 Genomes Project Consortium et al., 2010).

PCR Amplification

All blood samples were amplified directly without DNA extraction. The 268 unrelated samples were genotyped using our new established multiplex amplification system on the Thermo 96-Well PCR System (Thermo Fisher Scientific Company, Carlsbad, United States). The analyzed panel including 38 X-InDel markers, namely rs10671504, rs11277082, rs1160845, rs143123845, rs149102585, rs16367, rs16368, rs16397, rs16637, rs17394, rs199731653, rs2307707, rs2307741, rs2308033, rs2308280, rs25581, rs3048996, rs3077884, rs3215490, rs34763847, rs35574346, rs35954471, rs363794, rs4030406, rs45449991, rs56820033, rs57608175, rs57843641, rs58595330, rs59605609, rs60283667, rs71671860, rs3216913, rs10699224, rs3859989, rs61260787, rs36094418 and rs79829945. The localizations of the different markers were previously described in our published article (Chen et al., 2021). Separation of PCR-amplified products were performed on the ABI 3130xL DNA Analyzer (Applied Biosystems, Foster City, CA, United States). Electropherogram analysis and allele assignment were performed with GeneMapper v 4.0.

Statistical Analysis

Arlequin v3.5 software was used to test for the linkage disequilibrium (LD) and Hardy-Weinberg Equilibrium (HWE) between all pairs of the 38 X-InDel loci, and p values were corrected by the Bonferroni procedure (Excoffier et al., 2007). HWE was evaluated in females,

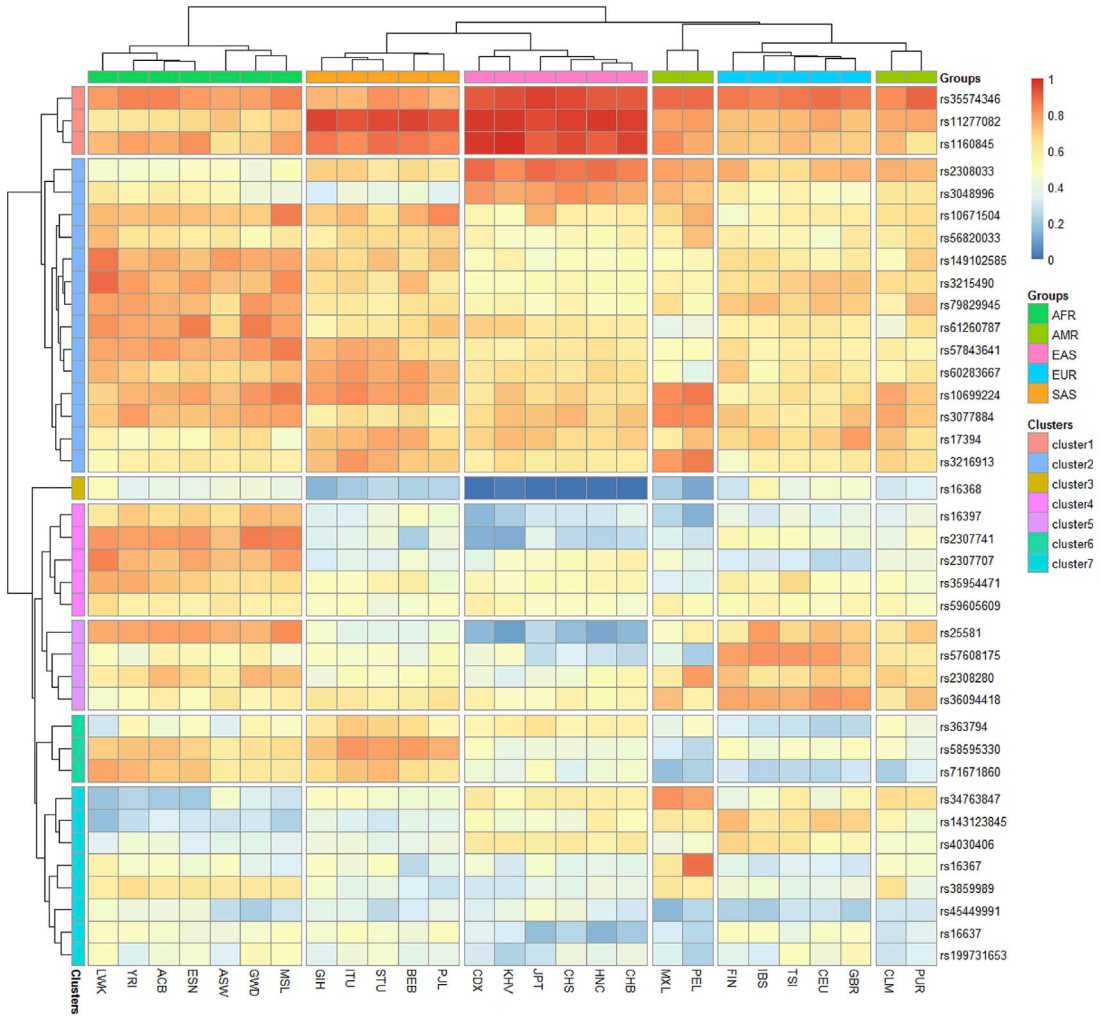

FIGURE 5 | Heatmap on the basis of the insertion allele frequency distributions for Henan Han group and other 26 populations worldwide.

whereas LD was tested by an extension of Fisher's exact test on contingency tables, D′ and Chi-square values from male haplotype counts (Caputo et al., 2017). LD was considered positive when D'≥0.8 and at distances of ≤60 kb (Reich et al., 2001). Allele frequencies and forensic parameters such as discrimination power (DP), probability of exclusion of trios-testing (PEtrio), probability of exclusion of duos-testing (PEduos), polymorphic information content (PIC) and expected heterozygosity (He) of the 38 X-InDel loci were calculated using StatsX v2.0 software (Lang et al., 2019). Fisher's exact test were used by SPSS 25.0 software package for comparing allele frequencies between males and females. A value of $p < 0.05$ was considered statistically significant. Chi-square tests were used to compare the allele frequencies between reference populations and Henan Han population.

Unbiased Nei's genetic distances were calculated in Genalex v6.5 (Peakall and Smouse, 2012) based on allele frequencies of the 38 X-InDel loci and heatmaps were plotted using Pheatmap package of R Statistical Software v4.0.2. The principal component analyses (PCA) were performed using the online tool (https://www.omicstudio.cn/tool). Phylogenetic neighbor-joining (NJ) tree was reconstructed by MEGA software v7.0 using Nei genetic distance matrices. On the basis of raw data of female individuals, population genetic structure was analyzed by the Structure version 2.3.4.21 (Evanno et al., 2005). Structure parameter was set to run 15 replicates from K = 2 to K = 8 with 10,000 burn-ins and 10,000 Markov Chain Monte Carlo (MCMC). CLUMPP version 1.1.2 was used to align the different replicates of STRUCTURE analysis (Jakobsson and Rosenberg, 2007). DISTRUCT 1.1 was used for the graphical display of population structure (Rosenberg, 2004).

RESULTS AND DISCUSSION

Allele Frequencies

The raw genotype data of 38 X-InDel loci for 268 individuals of Henan Han population are shown in **Supplementary Table S1**. The insertion allele frequencies of the 38 X-InDel loci are shown in **Figure 2**. The numerical values of insertion and deletion allele frequencies are shown in **Supplementary Table S2**.

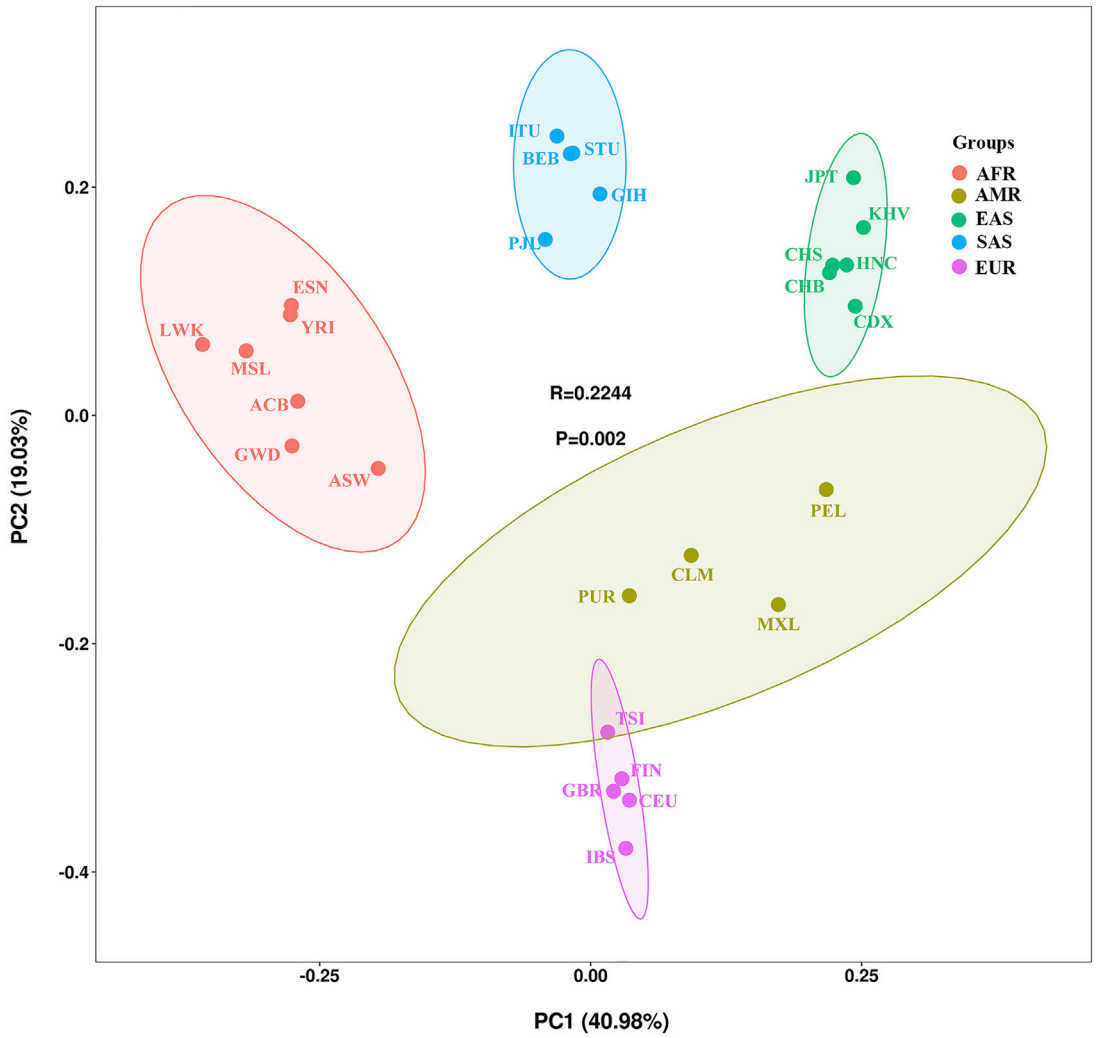

FIGURE 6 | PCA among the 27 populations based on the allele frequencies of 38 X-InDel markers.

To evaluate the differences in allele frequencies of the X-InDel loci between males and females, we first carried out fisher's exact test, and found that there were no statistical significance between males and females ($p > 0.05$) except for rs4030406 locus ($p = 0.007$). The male and female allele frequencies of rs4030406 locus are shown respectively, and the male and female allele frequencies of the other 37 X-InDel loci were combined for calculation. Unless otherwise specified, the female allele frequency of rs4030406 locus is used for calculation. The 29 insertion allele frequencies out of 38 loci for Henan Han population range from 0.3000 to 0.7000. As a whole, the allele frequency distributions of the 38 X-InDel loci are relatively balanced in Henan Han population.

Linkage Disequilibrium Analysis of 38 X-InDels

The recombination rate for the X chromosome is almost exactly two-thirds of the genome average (Kong et al., 2002). It is expected that linkage disequilibrium (LD) will be greater on the X chromosome, especially among younger loci, which have had less time for recombination to break down LD. Therefore, we carried out LD tests firstly and illustrated the degree of LD between the polymorphic loci, as shown in **Figure 3**. With the condition of D'≥0.8, the linkage Chi-square test was significant (after Bonferroni adjustments) and the distance between the loci is less than 60kb, three disequilibrium linkage blocks containing two linked loci were found in the tests, namely block 1 (rs3859989 and rs61260787 3.1 kb apart), block 2 (rs36094418 and rs79829945 12.9 kb apart), block 3 (rs3216913 and rs10699224 23.6 kb apart). The two loci in each block were combined for the subsequent forensic studies. Meanwhile, we performed LD tests within continental groups (**Supplementary Figure S1**), and achieved consistent results in East Asian, European and American groups. In South Asian and African groups, the block 1 (rs3859989 - rs61260787) and block 3 (rs3216913 - rs10699224) were detected as well, but no significant association was observed between InDels rs36094418 and rs79829945 (block 2). This discrepancy could be explained by other factors than linkage, such as admixture and population

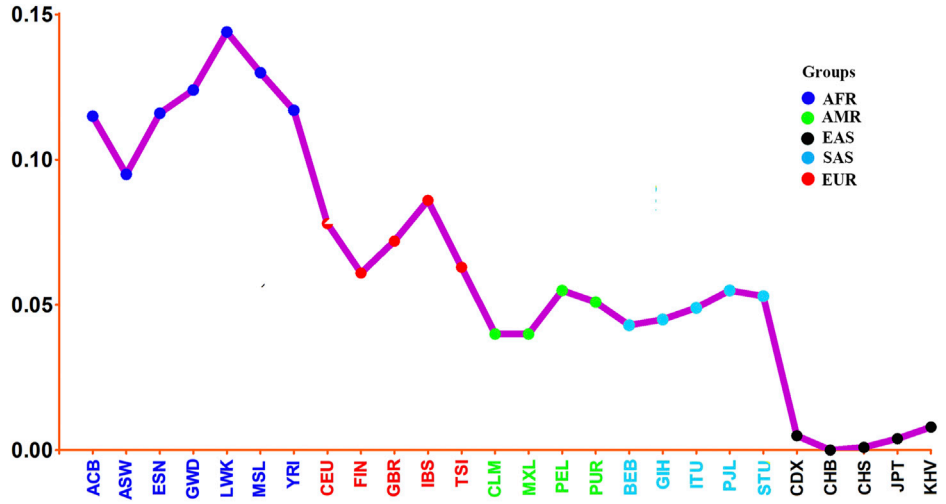

FIGURE 7 | Line chart on the basis of Nei's genetic distances between studied Henan Han population and the other 26 reference populations.

substructure (Martinez et al., 2019). Drawing on the methods described by Ferragut et al. (2017) we retained all the associated loci in the subsequent genetic analysis.

Forensic Parameters

To ensure samples can represent the population, Hardy-Weinberg equilibrium (HWE) testing was performed, and no deviations from HWE ($p > 0.05$) were found among the 38 X-InDel loci, which indicated that we can estimate both the forensic characteristics of the single locus (except for the three blocks) and the combined forensic efficiency of the 38 X-InDel loci in our following analysis. Forensic parameters for the 38 X-InDel loci in Henan Han population are shown in **Figure 4** and **Supplementary Table S3**. Most of He values for Henan Han population are in the range of 0.4249–0.6133 (29 out of 35), and most of PIC values for Henan Han population are in the range of 0.3–0.5399 (29 out of 35). The combined DP (CDP) for males and females are 1–9.18E-17 and 1-7.22E-12 respectively. The mean exclusion chance in trios and duos are 0.999999319 and 0.999802969, respectively. Based on the standard of 0.9999, the results indicated that the panel of 38 X-InDel loci meets the efficiency of individual identification and parentage testing of trios in Henan Han population. It is a pity that the relative standard deviations of the mean exclusion chance in duos is 0.0001 lower. We speculate that replacing the low heterozygous loci and linked loci or adding more high heterozygous loci may achieve the standard of 0.9999. On the whole, the panel of 38 X-InDel loci can be used as an effective supplementary tool for human identity and parentage testing in China.

Allele Frequency Distributions in 27 Populations

To investigate the allele frequency distributions of the 38 X-InDel loci in different continental populations, we further collected the raw data of the 38 X-InDel loci in 26 populations worldwide and conducted a heatmap based on the insertion allele frequencies of the 38 X-InDel loci of the 26 populations and Henan Han population, as shown in **Figure 5**. Red represented high insertion allele frequencies, on the contrary, blue represented low insertion allele frequencies. The populations from Africa, America, East Asia, Europe and South Asia were divided into five groups roughly according to the continent, except for the Puerto Rican (PUR) and Colombian (CLM) which were clustered with the European populations, and Henan Han population was classified to East Asian population. After clusters analyzing, we divided the 38 X-InDel loci into seven clusters (1–7). Cluster 1, including rs35574346, rs11277082 and rs1160845 loci, shows high allele frequencies in East and South Asian population (>0.9). Instead, the cluster three that contains only one locus of rs16368 is observed at low insertion allele frequencies in five groups especially in East Asian population (<0.1). Cluster two contains 14 loci, which have high heterozygosity and polymorphism in the 27 populations and most of the insertion allele frequencies of the 14 X-InDel loci are between 0.4–0.6. The insertion allele frequencies of clusters 4–7 show more differences among groups. Markers showing large allele frequency differences between different ancestral or geographically distant populations were considered to be ideal ancestry informative markers (AIMs). We found rs11277082 and rs1160845 loci in Asian population, rs2307741 locus in African population, rs16367 locus in Peruvian and rs34763847 locus in Peruvian and Mexican showed high frequencies. In contrast, rs16368 locus and rs25581 locus showed low frequencies in East Asian population. These results mean that the seven loci could be eminently suitable for AIMs.

Overall, the vast majority of loci in the panel including 38 X-InDel loci shows better heterozygosities and polymorphisms, especially in African, American and European populations, and the difference among groups

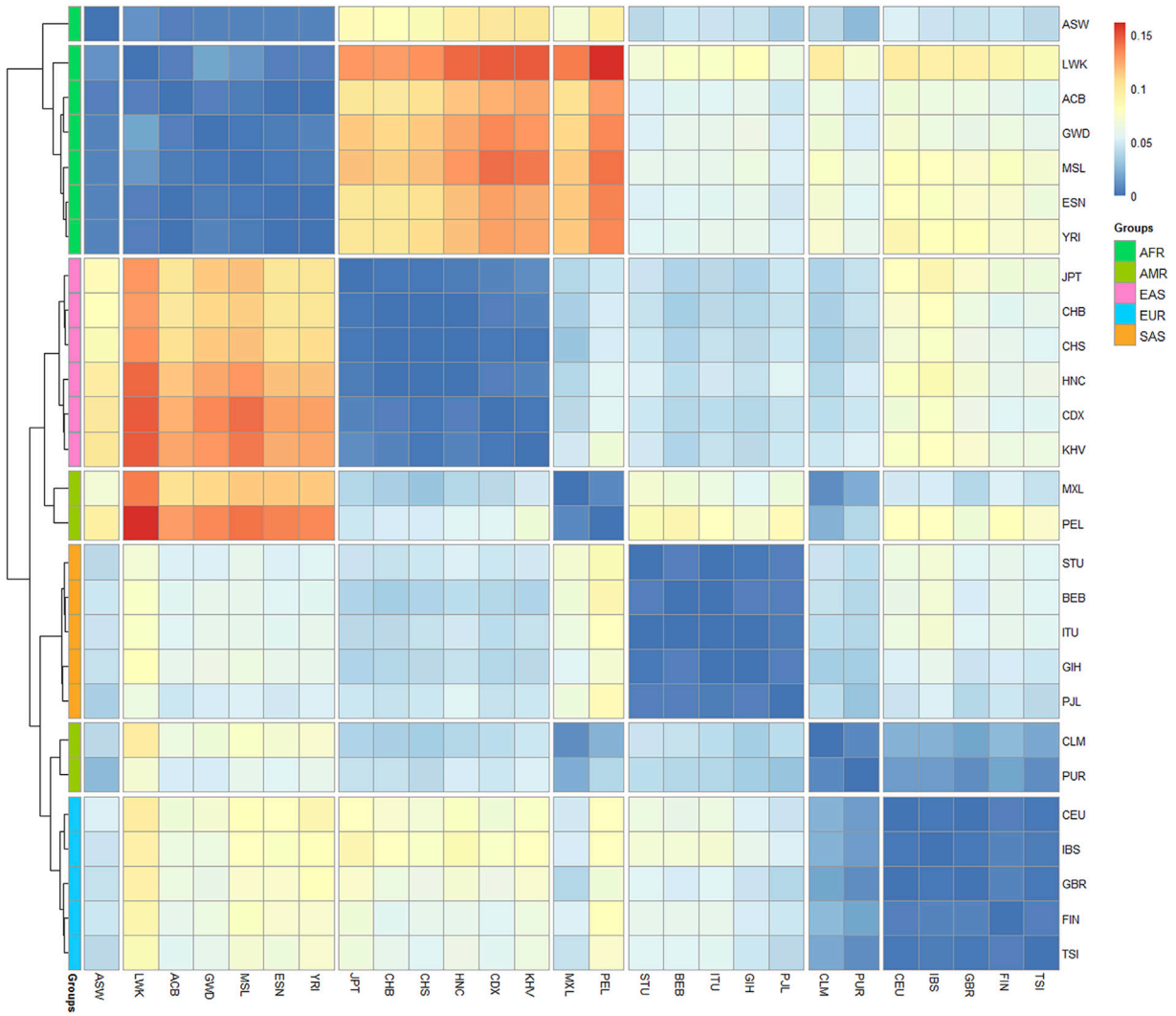

FIGURE 8 | Heatmap on the basis of Nei's genetic distances among the 27 populations.

are valuable in understanding the ethnic aggregation and migration and forensic identifications.

Genetic Affinity Analysis Among the 27 Populations Using the Panel of 38 X-InDel Loci

To illustrate the genetic background and population relationship of the 27 populations using the 38 X-InDel markers, the PCA was performed based on insertion allele frequencies. As shown in **Figure 6**, the 27 populations were labeled with five different colors according to the large continental-level groups, and the contribution rates of PC1 and PC2 were 40.98 and 19.03% respectively, which explained aggregately about 60% of the genetic structure variances among populations. Consistent with the large continental-level groups, we observed the five main population clusters. We noticed that the American populations clustered not so well as other four groups and showed genetic affinities to the European populations. Li et al. also showed that Colombian was clustered with European populations and other American populations were dispersedly distributed in MDS plot (Li et al., 2019). For historical reasons, most of the Puerto Rican and Colombian from the four American populations are European or European-Indian mixed-race population (Via et al., 2011; Ossa et al., 2021), which could be the reason that American populations clustered not obviously.

Although East Asian populations cluster tightly in the upper right corner, they showed substantial substructure in the PC2. Han Chinese (CHB, CHS, HNC), Japanese, Chinese Dai and Vietnamese Kinh were separated each other. The studied Henan Han population was plotted in the middle of the East Asian populations and presented close genetic affinities to Han Chinese in Beijing and Han Chinese in South China.

To further reveal population genetic similarities and divergences of Henan Han population and other 26 reference populations, the pairwise Nei's genetic distances between studied

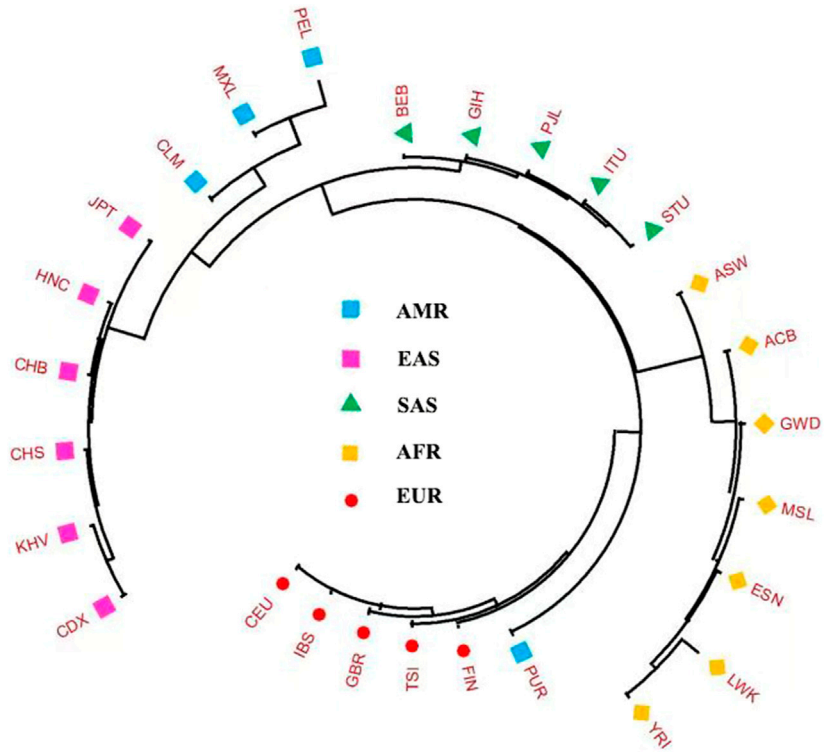

FIGURE 9 | Neighbor-Joining phylogenetic tree constructed on the basis of the Nei's D matrix among the 27 populations.

Henan Han population and the other 26 reference populations were calculated and the result was shown in the form of line chart (**Figure 7**). We observed that Henan Han population had close affinities with East Asian populations (with smallest genetic distances), especially with Han Chinese in Beijing (CHB, 0.000) and Han Chinese in South China (CHS, 0.001), and showed most distant affinities with African populations (with largest genetic distances), especially with Luhya in Webuye, Kenya (with largest genetic distances, 0.144). It should be noted that Henan Han population showed closer affinity with Japanese (JPT, 0.004) than Chinese Dai (CDX, 0.005) and Vietnamese Kinh (KHV, 0.008), and even some research found in PCA that the Japanese individuals are overlapped with Chinese Han populations in North China (Cao et al., 2020). Genetic ancestry is strongly correlated with geography as well as linguistic affiliations (Su et al., 1999; HUGO Pan-Asian SNP Consortium et al., 2009; Watanabe et al., 2019). Chinese Dai in Xishuangbanna is geographically and linguistically close to Southeast Asian. Meanwhile, multiple migration in history affect population composition of East Asian but not Southeast Asian, which maybe the reason why the Japanese showed closer genetic relationship with Henan population in comparison with Chinese Dai and Vietnamese Kinh. Surprisingly, we also noted that American populations and South Asian populations showed similar genetic distances with East Asian, and even more close genetic distances of Colombian (CLM) and Mexican Ancestry in Los Angeles United States (MXL) with East Asian. It happens that there is a similar phenomenon was shown in a recent study (Cao et al., 2020).

Furthermore, the matrix of Nei's genetic distances was applied to draw a heatmap plot. As shown in **Figure 8**, red color denotes the larger genetic distances and blue color denotes the smaller genetic distances. We found that the 27 populations were divided into 7 clusters: African populations were split into cluster 1 and 2, and the cluster one contains only one population of African Ancestry in Southwest United States (ASW); American populations were divided into cluster 4 (MXL and PEL) and cluster 6 (CLM and PUR); The clusters 3, five and seven were composed of East Asian populations, South Asian populations and European populations separately. We noted that larger genetic distances were presented between cluster two and clusters 3 and 4, which indicate that most of African populations show distant affinities with East Asian populations and American populations. Meanwhile, we found the smaller genetic distances were presented between cluster six and clusters 4 and 7, indicating that PUR and CLM populations present similar affinities with American populations and European populations.

To more visually exhibit the genetic similarities and divergences among the 27 populations, a Neighbor-Joining (N-J) phylogenetic tree was constructed based on the allele frequencies of 38 X-InDel loci, as shown in **Figure 9**. Neighbor-Joining tree supported the population relationship patterns that HNC population was gathered with CHB, and then gathered together with CHS and other East Asian populations. We also found that the Puerto Rican

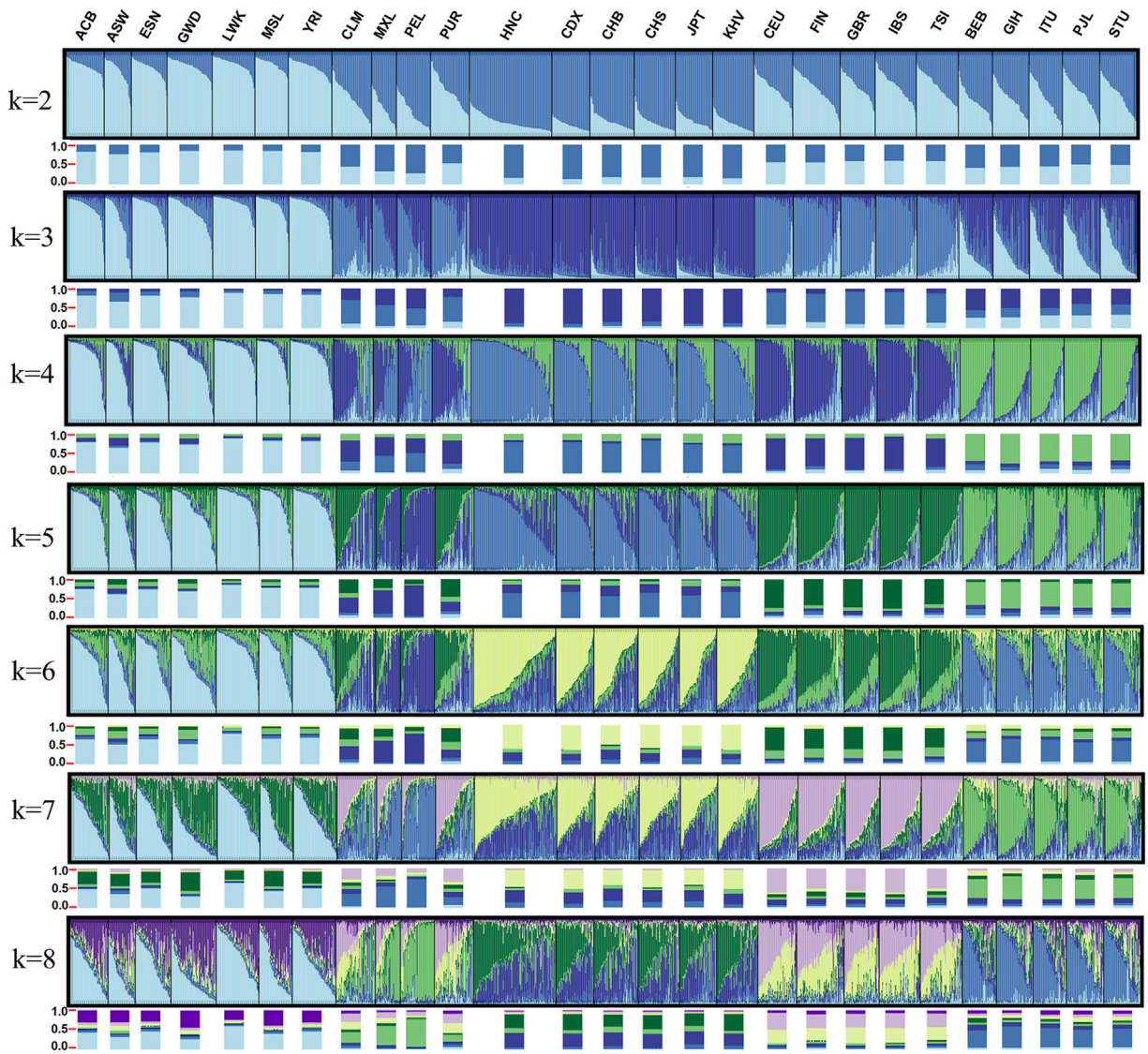

FIGURE 10 | Structure analysis results for females of the 27 populations.

in American populations clustered with the European populations, and the other three American populations showed distant genetic affinities each other. This result supported the aforementioned PCA and Nei's genetic distances and further proved Puerto Rican have close genetic affinities with the European populations. Similarly, Li et al. showed the close affinities of American populations with European populations and Colombian was clustered with European populations, however, different with our results, the Gujarati Indians in Houston (GIH) was clustered with American populations (Li et al., 2019). In African populations, the population of ASW also showed distant affinities with the other six populations.

To analyze the population ancestry component among the 27 populations, we conducted structure analysis based on 34 X-InDels (excluding three linked loci and the locus with different allele frequency between males and females) raw genotype of 1,377 females. The detection of the number of genetic groups that best fit the data was performed by an online program STRUCTURE HARVESTER (Earl and Vonholdt, 2012) and the best K-value was observed at K = 4. As shown in **Figure 10**, K values were set at the range of two–8. Each vertical line represents an individual and all individuals were clustered by the populations. Colors represent the inferred ancestry from K ancestral populations. Histograms below each population show proportions of ancestry components in the population. When K = 2, East Asian-dominant component and African-dominant component were distinguished clearly. When K = 3, European-dominant component was identified, and we could distinguish African, East Asian and South Asian groups based on the proportions of

ancestry components, but failure to distinguish European and American group. When K = 4 (the optimum K-value), South Asian-dominant component was further separated clearly. We also observed that the identified four ancestry components were generally in accord with geographic patterns. In addition, when K = 4, Mexican and Peruvian population could be separated with European East Asian, South Asian and African groups. The ancestry components of Colombian and Puerto Rican population were considered to be intermediate between Mexican and Peruvian populations and European populations. When K > 4, no more substructures were found in the 27 populations. The studied Henan Han population shared a similar ancestry component with the other East Asian populations. It has long been recognized that markers with relatively low mutation rates (SNP, InDels) serve as best loci for the analysis of human history over longer time scales (Moriot et al., 2018). However, not as Yuchen Wang et al. described that Han Chinese, Japanese and Korean can be distinguished using a series of single nucleotide polymorphism (SNP) AIMs (Wang et al., 2018), there are no sub-populations were separated in East Asian populations, African populations, European populations and South Asian populations in our study, which could be ascribed to the relative stability of X-Chromosome or more representative loci were needed to distinguish sub-populations within continental populations. We also conducted structure analysis based on raw genotype of 1,395 males and achieved familiar results, as shown in **Supplementary Figure S2**.

CONCLUSION

In the present study, we first described a novel panel of 38 X-InDels and investigated the forensic efficiency in Henan Han population. The results showed that this panel is powerful as a complementary tool for forensic individual identification and parentage testing of trios. Based on the allele frequencies and raw data of the 38 X-InDels, we conducted principal component analysis, calculated Nei's genetic distances, and constructed phylogenetic tree and structure analysis to illustrate the genetic affinity and population ancestry component among Henan Han population and 26 populations worldwide. It is observed that the grouping from ancestry component analysis and genetic affinity revealed by the panel of 38 X-InDels are generally consistent with geographical classifications. Meanwhile, we noticed that mixed-race populations could be distinguished clearly from the continental population groups by using the panel of 38 X-InDels. However, sub-population identification within continental population groups was failed due to the limitation of X-InDels. It is expected that compound markers which are adopted to infer biogeographic ancestry would improve accurate classification of populations and sub-populations. In conclusion, this study makes a solid foundation of the population data for the application of the 38 X-InDel markers panel in individual identification, parentage testing and biogeographic ancestry analysis, though much more regional and national population data still need to be collected.

DATA AVAILABILITY STATEMENT

The raw data used during the current study were uploaded as supplementary materials, and all data are fully available without restriction.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of Xi'an Jiaotong University Health Science Center. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

LZ and SL: Conceptualization and Resources; WD and CL: Investigation and Resources; LZ and ZZ: Data Curation, Writing—Original Draft, Writing—Review and Editing, Visualization; SL and CL: Supervision, Project administration; ZZ, SL, and CL: Funding acquisition.

FUNDING

This study was supported by grants from the Science and Technology Program of Guangzhou, China (grant no. 201607010016 and 2019030016), the Ministry of Science and Technology of the People's Republic of China (grant no. 2013FY114300) and the Doctor Scientific Research Foundation of Xinxiang Medical University (grant no. XYBSKYZZ201819).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.805936/full#supplementary-material

REFERENCES

Bastos-Rodrigues, L., Pimenta, J. R., and Pena, S. D. J. (2006). The Genetic Structure of Human Populations Studied through Short Insertion-Deletion Polymorphisms. *Ann. Hum. Genet.* 70, 658–665. doi:10.1111/j.1469-1809.2006.00287.x

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

Caputo, M., Amador, M. A., Santos, S., and Corach, D. (2017). Potential Forensic Use of a 33 X-InDel Panel in the Argentinean Population. *Int. J. Leg. Med.* 131, 107–112. doi:10.1007/s00414-016-1399-z

Chen, L., Pan, X., Wang, Y., Du, W., Wu, W., Tang, Z., et al. (2021). Development and Validation of a Forensic Multiplex System with 38 X-InDel Loci. *Front. Genet.* 12, 670482. doi:10.3389/fgene.2021.670482

Du, W., Feng, C., Yao, T., Xiao, C., Huang, H., Wu, W., et al. (2019). Genetic Variation and Forensic Efficiency of 30 Indels for Three Ethnic Groups in Guangxi: Relationships with Other Populations. *PeerJ* 7, e6861. doi:10.7717/peerj.6861

Earl, D. A., and Vonholdt, B. M. (2012). STRUCTURE HARVESTER: a Website and Program for Visualizing STRUCTURE Output and Implementing the Evanno Method. *Conserv. Genet. Resour.* 4, 359–361. doi:10.1007/s12686-011-9548-7

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the Number of Clusters of Individuals Using the Software STRUCTURE: a Simulation Study. *Mol. Ecol.* 14, 2611–2620. doi:10.1111/j.1365-294x.2005.02553.x

Excoffier, L., Laval, G., and Schneider, S. (2007). Arlequin (Version 3.0): an Integrated Software Package for Population Genetics Data Analysis. *Evol. Bioinform Online* 1, 47–50. doi:10.1177/117693430500100003

Fan, H., and Chu, J.-Y. (2007). A Brief Review of Short Tandem Repeat Mutation. *Genomics Proteomics Bioinformatics* 5, 7–14. doi:10.1016/s1672-0229(07)60009-6

Ferragut, J. F., Bentayebi, K., Pereira, R., Castro, J. A., Amorim, A., Ramon, C., et al. (2017). Genetic Portrait of Jewish Populations Based on Three Sets of X-Chromosome Markers: Indels, Alu Insertions and STRs. *Forensic Sci. Int. Genet.* 31, e5. doi:10.1016/j.fsigen.2017.09.008

Giardina, E., Spinella, A., and Novelli, G. (2011). Past, Present and Future of Forensic DNA Typing. *Nanomedicine* 6, 257–270. doi:10.2217/nnm.10.160

Gomes, C., Magalhães, M., Alves, C., Amorim, A., Pinto, N., and Gusmão, L. (2012). Comparative Evaluation of Alternative Batteries of Genetic Markers to Complement Autosomal STRs in Kinship Investigations: Autosomal Indels vs. X-Chromosome STRs. *Int. J. Leg. Med* 126, 917–921. doi:10.1007/s00414-012-0768-5

Gomes, C., Quintero-Brito, J. D., Martínez-Gómez, J., Pereira, R., Baeza-Richer, C., Aler Gay, M., et al. (2020a). Spanish Allele and Haplotype Database for 32 X-Chromosome Insertion-Deletion Polymorphisms. *Forensic Sci. Int. Genet.* 46, 102262. doi:10.1016/j.fsigen.2020.102262

Gomes, I., Pinto, N., Antão-Sousa, S., Gomes, V., Gusmão, L., and Amorim, A. (2020b). Twenty Years Later: A Comprehensive Review of the X Chromosome Use in Forensic Genetics. *Front. Genet.* 11, 926. doi:10.3389/fgene.2020.00926

He, G., Ren, Z., Guo, J., Zhang, F., Zou, X., Zhang, H., et al. (2019). Population Genetics, Diversity and Forensic Characteristics of Tai-Kadai-Speaking Bouyei Revealed by Insertion/deletions Markers. *Mol. Genet. Genomics* 294, 1343–1357. doi:10.1007/s00438-019-01584-6

Huang, Q.-Y., Xu, F.-H., Shen, H., Deng, H.-Y., Liu, Y.-J., Liu, Y.-Z., et al. (2002). Mutation Patterns at Dinucleotide Microsatellite Loci in Humans. *Am. J. Hum. Genet.* 70, 625–634. doi:10.1086/338997

HUGO Pan-Asian SNP Consortium, 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

Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a Cluster Matching and Permutation Program for Dealing with Label Switching and Multimodality in Analysis of Population Structure. *Bioinformatics* 23, 1801–1806. doi:10.1093/bioinformatics/btm233

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., et al. (2002). A High-Resolution Recombination Map of the Human Genome. *Nat. Genet.* 31, 241–247. doi:10.1038/ng917

Lang, Y., Guo, F., and Niu, Q. (2019). StatsX v2.0: the Interactive Graphical Software for Population Statistics on X-STR. *Int. J. Leg. Med.* 133, 39–44. doi:10.1007/s00414-018-1824-6

Larue, B. L., Lagacé, R., Chang, C.-W., Holt, A., Hennessy, L., Ge, J., et al. (2014). Characterization of 114 Insertion/deletion (INDEL) Polymorphisms, and Selection for a Global INDEL Panel for Human Identification. *Leg. Med.* 16, 26–32. doi:10.1016/j.legalmed.2013.10.006

Li, L., Ye, Y., Song, F., Wang, Z., and Hou, Y. (2019). Genetic Structure and Forensic Parameters of 30 InDels for Human Identification Purposes in 10 Tibetan Populations of China. *Forensic Sci. Int. Genet.* 40, e219–e227. doi:10.1016/j.fsigen.2019.02.002

Lin, Z., Kejie, W., Hongyan, W., Aiying, F., Xupeng, S., and Zhendong, Z. (2017). Analysis of Mutation of 20 Autosomal Short Tandem Repeat Loci in Henan Han Population. *Chin. J. Forensic Med.* 32, 33–35. doi:10.13618/j.issn.1001-5728.2017.01.009

Martinez, J., Polverari, F. S., Silva, F. A. D. J., Braganholi, D. F., Ferraz, J., Gusmão, L., et al. (2019). Genetic Characterization of 32 X-InDels in a Population Sample from São Paulo State (Brazil). *Int. J. Leg. Med.* 133, 1385. doi:10.1007/s00414-018-01988-w

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., et al. (2006). An Initial Map of Insertion and Deletion (INDEL) Variation in the Human Genome. *Genome Res.* 16, 1182–1190. doi:10.1101/gr.4565806

Moriot, A., Santos, C., Freire-Aradas, A., Phillips, C., and Hall, D. (2018). Inferring Biogeographic Ancestry with Compound Markers of Slow and Fast Evolving Polymorphisms. *Eur. J. Hum. Genet.* 26, 1697–1707. doi:10.1038/s41431-018-0215-2

Ossa, H., Posada, Y., Trujillo, N., Martínez, B., Loiola, S., Simão, F., et al. (2021). Patterns of Genetic Diversity in Colombia for 38 Indels Used in Human Identification. *Forensic Sci. Int. Genet.* 53, 102495. doi:10.1016/j.fsigen.2021.102495

Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: Genetic Analysis in Excel. Population Genetic Software for Teaching and Research-Aan Update. *Bioinformatics* 28, 2537–2539. doi:10.1093/bioinformatics/bts460

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., et al. (2001). Linkage Disequilibrium in the Human Genome. *Nature* 411, 199–204. doi:10.1038/35075590

Rosenberg, N. A. (2004). Distruct: a Program for the Graphical Display of Population Structure. *Mol. Ecol. Notes* 4, 137. doi:10.1046/j.1471-8286.2003.00566.x

Schaffner, S. F. (2004). The X Chromosome in Population Genetics. *Nat. Rev. Genet.* 5, 43–51. doi:10.1038/nrg1247

Sheng, X., Bao, Y., Zhang, J. S., Li, M., Li, Y. N., Xu, Q. N., et al. (2018). Research Progress on InDel Genetic Marker in Forensic Science. *Fa Yi Xue Za Zhi* 34, 420–427. doi:10.12116/j.issn.1004-5619.2018.04.016

Su, B., Xiao, J., Underhill, P., Deka, R., Zhang, W., Akey, J., et al. (1999). Y-chromosome Evidence for a Northward Migration of Modern Humans into Eastern Asia during the Last Ice Age. *Am. J. Hum. Genet.* 65, 1718–1724. doi:10.1086/302680

The 1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., et al. (2010). A Map of Human Genome Variation from Population-Scale Sequencing. *Nature* 467, 1061–1073. doi:10.1038/nature09534

Via, M., Gignoux, C. R., Roth, L. A., Fejerman, L., Galanter, J., Choudhry, S., et al. (2011). History Shaped the Geographic Distribution of Genomic Admixture on the Island of Puerto Rico. *PloS one* 6, e16513. doi:10.1371/journal.pone.0016513

Wang, Y., Lu, D., Chung, Y.-J., and Xu, S. (2018). Genetic Structure, Divergence and Admixture of Han Chinese, Japanese and Korean Populations. *Hereditas* 155, 19. doi:10.1186/s41065-018-0057-5

Watanabe, Y., Naka, I., Khor, S.-S., Sawai, H., Hitomi, Y., Tokunaga, K., et al. (2019). Analysis of Whole Y-Chromosome Sequences Reveals the Japanese Population History in the Jomon Period. *Sci. Rep.* 9, 8556. doi:10.1038/s41598-019-44473-z

Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C., and Marth, G. (2002). Human Diallelic Insertion/deletion Polymorphisms. *Am. J. Hum. Genet.* 71, 854–862. doi:10.1086/342727

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

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.

The handling editor declared a past co-authorship with one of the authors CL.

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.

frontiers in Genetics

REVIEW
published: 27 January 2022
doi: 10.3389/fgene.2022.767018

The Peopling and Migration History of the Natives in Peninsular Malaysia and Borneo: A Glimpse on the Studies Over the Past 100 years

Boon-Peng Hoh[1]*, *Lian Deng*[2] and *Shuhua Xu*[2,3,4,5,6,7,8,9]

[1]Faculty of Medicine and Health Sciences, UCSI University, UCSI Hospital, Port Dickson, Malaysia, [2]State Key Laboratory of Genetic Engineering, Center for Evolutionary Biology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China, [3]Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China, [4]Department of Liver Surgery and Transplantation Liver Cancer Institute, Zhongshan Hospital, Fudan University, Shanghai, China, [5]Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Human Phenome Institute, Fudan University, Shanghai, China, [6]School of Life Science and Technology, ShanghaiTech University, Shanghai, China, [7]Jiangsu Key Laboratory of Phylogenomics and Comparative Genomics, School of Life Sciences, Jiangsu Normal University, Xuzhou, China, [8]Henan Institute of Medical and Pharmaceutical Sciences, Zhengzhou University, Zhengzhou, China, [9]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China

OPEN ACCESS

Edited by:
Chuan-Chao Wang, Xiamen University, China

Reviewed by:
Guanglin He, Nanyang Technological University, Singapore
Xiaoming Zhang, Kunming Institute of Zoology (CAS), China

***Correspondence:**
Boon-Peng Hoh
hoh.boonpeng@gmail.com

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

Received: *30 August 2021*
Accepted: *07 January 2022*
Published: *27 January 2022*

Citation:
Hoh B-P, Deng L and Xu S (2022) The Peopling and Migration History of the Natives in Peninsular Malaysia and Borneo: A Glimpse on the Studies Over the Past 100 years. Front. Genet. 13:767018. doi: 10.3389/fgene.2022.767018

Southeast Asia (SEA) has one of the longest records of modern human habitation out-of-Africa. Located at the crossroad of the mainland and islands of SEA, Peninsular Malaysia is an important piece of puzzle to the map of peopling and migration history in Asia, a question that is of interest to many anthropologists, archeologists, and population geneticists. This review aims to revisit our understanding to the population genetics of the natives from Peninsular Malaysia and Borneo over the past century based on the chronology of the technology advancement: 1) Anthropological and Physical Characterization; 2) Blood Group Markers; 3) Protein Markers; 4) Mitochondrial and Autosomal DNA Markers; and 5) Whole Genome Analysis. Subsequently some missing gaps of the study are identified. In the later part of this review, challenges of studying the population genetics of natives will be elaborated. Finally, we conclude our review by reiterating the importance of unveiling migration history and genetic diversity of the indigenous populations as a steppingstone towards comprehending disease evolution and etiology.

Keywords: Orang Asli, natives, admixture, divergence, peopling history, migration

INTRODUCTION

Southeast Asia (SEA) is believed to be one of the earliest regions of hominin habitation recorded outside Africa nearly 2 million years ago, following the arrival of the ancient "Java Man" known as the *Homo erectus* (Jin et al., 2001). Archeological evidence from sites such as Tam Pa Ling in Laos (Demeter et al., 2012), Callao Cave in the Philippines (Mijares et al., 2010), and Niah Cave in Borneo Malaysia (Barker et al., 2007; Curnoe et al., 2016), suggest that SEA may have occupied by anatomically modern humans (AMH) at least 50–70 thousand years ago (kya). Today this landmass is home to ~600 million people, enriched with cultural, linguistic, and genetic diversity.

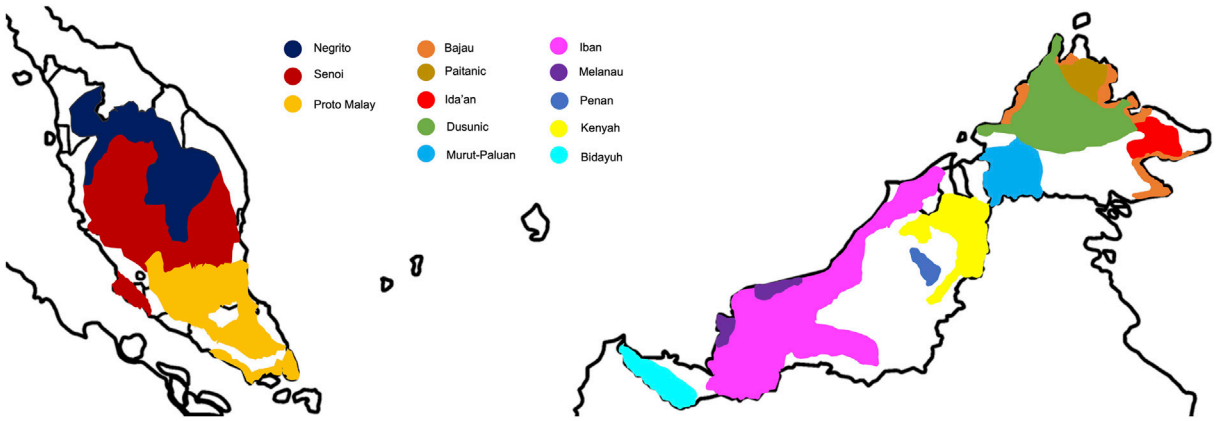

FIGURE 1 | Distribution of the major native populations from Peninsular Malaysia and Borneo.

SEA was once a great landmass bridging the Eurasia during the last glacial maximum peaked ~20 kya (Bellwood, 2007). This landmass, called the Sundaland, linked the present day Peninsular Malaysia and Borneo, the southern Philippines, and the west and south Indonesia islands. Subsequently, the Sundaland was hit by three great sea-level surges in approximately 14, 11, and 8 kya, respectively. The sea level rose by 120 m, thus forming the present-day map of SEA. These climate and geographical changes had drastically changed the flora and fauna in SEA, and plausibly the prehistorical peopling of modern human to the land of SEA (Macaulay et al., 2005; Bellwood, 2007; HUGO Pan-Asian SNP Consortium et al., 2009; Arenas et al., 2020).

Malaysia has been an important piece of puzzle to the global map of peopling and migration history, owing to its strategic location at the crossroad of the mainland and islands of SEA. It is a multi-ethnic, multi-lingual, and multi-cultural nation with diverse socio-economic practices. The Malay ethnic group is the predominant native in Malaysia, making up ~61% of the total population; whereas the natives from Borneo comprise ~9% of the total population (**Figure 1**). The indigenous populations from Peninsular Malaysia, locally known as the Orang Asli, constitute ~0.5% of the total population. They are heterogeneous and are categorized into Negrito, Senoi, and Proto Malay, each sub-classified into different sub-tribes (**Table 1**). Traditionally, the Negritos are hunters-gatherers that practice egalitarianism, semi-nomadic and patrilineal descent system. The Senoi was primitive slash-and-burn farmers, and the Proto-Malays are primarily agriculturists and fishermen.

The Malaysian section of Borneo comprises the states of Sabah and Sarawak. The Sabah state (formerly known as North Borneo) is home to culturally diverse populations, comprising of more than 40 ethnicities that are broadly categorized into five major groups: Dusunic, Paitanic, Murutic, Ida'anic, and Sama-Bajaw (Combrink et al., 2008), with Kadazandusun being the largest ethnicity. Most of the native Sabahan inhabit the inland of northern Borneo. Traditionally, they were agriculturists and hunter-gatherers, except for a few populations scattering along the coastal lines of northern Borneo who are seafarers (mainly the Bajaus). The Ida'anic populations used to practice hunting and gathering as their subsistence. The Muruts were once head-hunters but the practice has been ceased during the British colonization. The Sarawak native population consists of some 40 ethnicities. They are either farmers or fishermen; whereas some rely on forest resources as their subsistence (e.g., Penan). The Iban (also known as Sea Dayak) is the major ethnic group, follow by Bidayuh (or known as Land Dayak), Melanau, and many others (**Table 1**). Like the Muruts, the Iban were formerly reputed as the head-hunters. They were seafarers or agriculturists. The Sarawak natives are believed to have migrated from the neighboring islands of SEA (e.g., Bisaya are thought to be linked with the Visayan of the Philippines) (Williams, 1965; Ibrahim et al., 1985); whereas some have little historical evidence regarding their origin (e.g., the Kenyah and Punan Bah tribes) (Nicolaisen, 1976). The ethnicities of the Sabah and Sarawak are primarily distinguished by their linguistic and sociocultural practices, and has been a long-standing debate.

The SEA populations are enriched with their linguistics spectrum, primarily represented by five language families: Austroasiatic, Austronesian, Tai-Kadai, Hmong-Mien and Sino-Tibetan (Wang, 2001; Enfield, 2005). Nonetheless the natives from Malaysia speak languages that fall under either Austroasiatic, or Austronesian families. The Negrito and Senoi tribes basically speak the "Aslian" language that belongs to the Mon-Khmer linguistic branch that falls under Austroasiatic family, each tribe has their own dialects; whereas the rest of the populations are primarily categorized as the Austronesian language speaking family. The Proto-Malay and the modern Malays speaks the language that derived from the Malayan group; while the Borneo natives' languages fall under the Malayo-Polynesian group.

Given such rich ethnological diversity and complex prehistorical events in this landmass, addressing migration and peopling history of Peninsular Malaysia and Borneo is therefore a pertinent question.

In this article, we attempt to summarize an overview on our understanding of migration and peopling histories of the natives from Peninsular Malaysia and Borneo over the past century,

TABLE 1 | Classification of Orang Asli and some major sub-ethnic populations from Borneo. Each Orang Asli group is further divided into six sub-tribes. The natives Sabahan are primarily classified into five language groups (Dusunic, Paitanic, Murutic, Ida'anic, and Sama-Bajaw). The major Sarawak natives include the two major communities namely, the Dayaks and the Orang Ulu.

Geographical distributions	Group	Sub-tribe	Language family (Aslian language branch)	Traditional socioeconomic activity
Northern Peninsular Malaysia	Negrito	Bateq	Austroasiatic (Northern Aslian)	Hunting and gathering
		Mendriq	Austro-Asiatic (Northern Aslian)	Hunting and gathering
		Jehai	Austroasiatic (Northern Aslian)	Hunting and gathering
		Kensiu	Austroasiatic (Northern Aslian)	Hunting and gathering
		Kintak	Austroasiatic (Northern Aslian)	Hunting and gathering
		Lanoh	Austroasiatic (Central Aslian)	Swine-and-burn; hunting and gathering
Central Peninsular Malaysia	Senoi	Semai	Austroasiatic (Central Aslian)	Swine-and-burn
		Temiar	Austroasiatic (Central Aslian)	Swine-and-burn
		Che Wong	Austroasiatic (Northern Aslian)	Swine-and-burn; hunting and gathering
		Mah Meri	Austroasiatic (Southern Aslian)	Swine-and-burn; hunting and gathering
		Semaq Beri	Austroasiatic (Southern Aslian)	Swine-and-burn
		Jahut	Austroasiatic (Central Aslian)	Swine-and-burn
Central- to Southern Peninsular Malaysia	Proto-Malay	Jakun	Austronesian	Farming and fishing
		Temuan	Austronesian	Agriculture
		Semelai	Austroasiatic (Southern Aslian)	Swine-and-burn
		Orang Kanaq	Austronesian	Agriculture
		Seletar	Austronesian	Fishing; hunting and gathering
		Orang Kuala	Austronesian	Fishing
West coast Sabah	Dusunic	Dusun	Austronesian	Agriculture
West coast Sabah		Kadazan	Austronesian	Agriculture; hunting and gathering
Northern Sabah		Rungus	Austronesian	Agriculture
Northwest coastal Sabah		Bisaya	Austronesian	Agriculture
Northern Sabah		Bonggi	Austronesian	Agriculture; fishing
West coast Sabah		Lotud	Austronesian	Agriculture
Northern Sabah		Kimaragang	Austronesian	Agriculture; fishing
Western coastal Sabah		Tatana	Austronesian	Agriculture; fishing
East coast Sabah		Minokok	Austronesian	Agriculture
Northern Sabah		Sonsogon	Austronesian	Agriculture; hunting and gathering
Northeast coast Sabah	Paitanic	Sungai	Austronesian	Agriculture; fishing
Northern Sabah		Lingkabau	Austronesian	Agriculture; fishing
Southwestern Sabah	Murutic	Murut	Austronesian	Agriculture; fishing; hunting-and-gathering
		Paluan	Austronesian	Agriculture; fishing; hunting-and-gathering
		Tagal	Austronesian	Agriculture; fishing; hunting-and-gathering
		Selungai	Austronesian	Agriculture; fishing; hunting-and-gathering
East coast Sabah	Ida'anic	Ida'an	Austronesian	Hunting-and-gathering
		Begahak	Austronesian	Agriculture; fishing; hunting-and-gathering
		Subpan	Austronesian	Agriculture; fishing; hunting-and-gathering
East coast Sabah	Sama-Bajaw	Bajau	Austronesian	Seafaring
Throughout Sarawak state	Dayak	Iban (Sea Dayak)	Austronesian	Seafaring; agriculture
Southwest Sarawak (adjacent to West Kalimantan)		Bidayuh (Land Dayak)	Austronesian	Agriculture
Northeast Sarawak	Orang Ulu	Kelabit	Austronesian	Agriculture
Interior areas of Sarawak		Penan	Austronesian	Hunting-and-gathering
Interior areas of Sarawak		Kenyah	Austronesian	Swine-and-burn; hunting-and-gathering; fishing
Interior areas of Sarawak		Kayan	Austronesian	Agriculture; fishing
Northern Sarawak		Lun Bawang; or Lundayeh	Austronesian	Fishing; agriculture; hunting-and-gathering
Northern Sarawak		Bisaya	Austronesian	Agriculture
Interior areas of Sarawak		Sebop	Austronesian	Agriculture
Coastal area of central Sarawak	Melanau		Austronesian	Agriculture; fishing

(Continued on following page)

TABLE 1 | (*Continued*) Classification of Orang Asli and some major sub-ethnic populations from Borneo. Each Orang Asli group is further divided into six sub-tribes. The natives Sabahan are primarily classified into five language groups (Dusunic, Paitanic, Murutic, Ida'anic, and Sama-Bajaw). The major Sarawak natives include the two major communities namely, the Dayaks and the Orang Ulu.

Geographical distributions	Group	Sub-tribe	Language family (Aslian language branch)	Traditional socioeconomic activity
Interior areas of Sarawak	Punan		Austronesian	Swine-and-burn
Southwest Sarawak (adjacent to West Kalimantan)	Kedayan		Austronesian	Agriculture; fishing

based on five chronological eras of genomic technological advancement: 1) Anthropological and Physical Characterization; 2) Blood Group Markers; 3) Protein Markers; 4) Mitochondrial and Autosomal DNA Markers; and 5) Whole Genome Analysis. We also provide some hints to the "missing puzzles" of the peopling history in SEA yet to be uncovered, and finally reiterate the implications of migration and peopling history of SEA populations on human health and diseases.

TABLE 2 | The proportion of blood group and population classification based on Biochemical Racial Index.

Blood group	Population
Proportion higher of "A"	European
Proportion higher of "B"	Asio-African
Equal distribution of "A" and "B"	Intermediate type of population
Proportion higher of "O"	Island folk; isolated population

Anthropological and Physical Characterization

The study of human populations in Malaysia began in the 19th century when the colonial scholars attempted to classify "race". Since genetic knowledge was rather limited, hypotheses were built merely based on physical and anthropological observations. The population of the Peninsular Malaysia was first grossly classified into two geographical races: "Negrito-like" and "East Asian-like" (Crawfurd, 1820). Populations originated from the Malay Peninsula were firstly categorized into the "black" and the "brown", but was refined by John Anderson as the "Negrito" and "non-Negrito" groups, respectively. The non-Negrito group includes, for instance, the "Sakei", "Orang Bukit" (means "People of the Hill"), and "Orang Laut" (means "People of the Sea"). Anderson considered the Malays as the "native" but not "indigenous" people (Anderson, 1824). These classifications were primarily rationalized by outward physical attributes including skin pigmentation (dark, brown light), hair morphology (woolly, wavy, straight, brownie, black), eye color, and stature (tall, short). Other considerations included language, culture, and subsistence (forager, farmer). These physical characteristics are still very much in use to date.

The basic "three-way division" framework was raised by the British anthropologists Skeat and Blagden (1906). They named these groups of people as the "woolly-haired" Negrito, "wavy-haired" Sakai, or "straight-haired" Jakun. Nonetheless, disagreements remained, and new competing theories continued to be proposed. Some did not agree with the separation between Negrito and Sakai (Annandale and Robinson, 1902); and one distinguished them into: Semang, the Northern Sakai, the Central Sakai, the Bersisi, and the Jakun (Wilkinson, 1926). The tripartite classification of Negrito, Senoi, and Proto Malay, was only firmed in the 1930s. Subsequent sub-tribes of the Orang Asli populations were categorized based on their linguistic dialects.

The Negritos from the SEA was once argued to be the remnants of the wrecked slave ship from Africa (Crawfurd, 1820), yet the possibility that they had a much deeper link hinting at the earliest inhabitant in SEA was not refuted. The connections between these "eastern" Negrito (as labeled by Crawfurd) and Africa became a hot topic of argument. The alternative Pan-Negrito theory was proposed, which argued that all indigenous people of Malaya were Negrito origin and were linked to others with Negrito physical characteristics, but was rejected (Skeat and Blagden, 1906). In the late 1930s, the role of SEA Negritos in human migration history was heavily debated, and the idea pertaining the connections between the African and Asian "pygmies" were revived among the anthropologists (Endicott, 2016).

The population classification for the Bornean native populations has been far more complicated because most of these populations have indistinguishable physical appearance and economic activities; yet a rich variety of cultural heritage and linguistics. These populations were classified as the East Asian populations, as exemplified by several native populations in the Philippines and Taiwan (Williams, 1965). The classification of the Bornean populations is more complex hence inconclusive and still under debate to date.

The distinct phenotypic characteristics of these SEA inhabitants along with a handful of archaeological and linguistic evidence thence provided hints to the early "Two-Layer" migration hypothesis, which argued that SEA was occupied (the first layer) by the direct descendants of the early modern humans out-of-Africa, and subsequently admixed with the later immigrants (the second layer) from North and/or East Asia leading to the present day SEA human diversity (Bellwood et al., 2007; Matsumura et al., 2008).

Blood Groups Markers

It was not until the 1950's that a new form of anthropological measurement was introduced into the consideration of the "race"

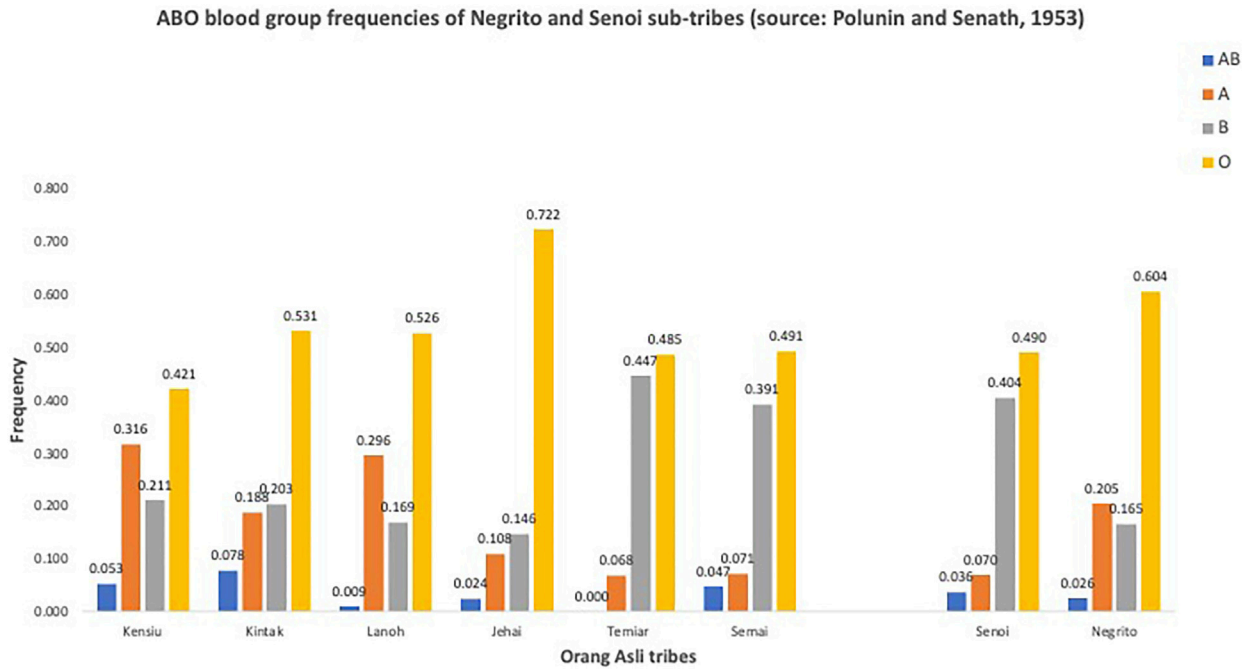

FIGURE 2 | ABO blood group frequencies of Negrito and Senoi sub-tribes (adapted from: Polunin and Sneath, 1953).

classification – the blood group (**Table 2**). The study for blood group anthropology was claimed to be the "first generation of human population genetics" (Endicott, 2016).

The first known blood group test on Orang Asli was carried out in the 117 Semai sub-tribe, which showed a higher frequency of the "O" blood group (Green, 1949). This report had led to a postulation that the Semai was an isolated population and somehow connected to other indigenous populations such as Tho and Muongs in mainland SEA and Tobias in Sumatera.

Further investigations into blood groups were carried out, attempting to address some contentious questions like classifications of the Orang Asli; relationships between the Orang Asli Negrito with the Negritos from Africa, SEA, and Australia, and the links between the Orang Asli groups with the rest of the populations in the world. Many of these studies were carried out by Ivan Polunin (Polunin, 1953; Polunin and Sneath, 1953). The categorization of the populations was primarily based on the tripartite classification and was very much relied on the physical anthropological measurements. The study showed that the populations classified under Negrito had lower frequency for the B blood group compared to those classified as Senoi (**Figure 2**), and noted similarities between the Senoi ABO frequencies to those people from India and Burma. On the other hand, the complete absence of sickle cell trait in Negrito, and the frequency for B blood group in the native Iban in Sarawak that was found similar to the Negrito (Graydon et al., 1952; Tan, 1978), had posted doubts on their connection to the African ancestry.

Protein Markers

The period between the 1960s–1980s marked the advancement of population genetics with the popularity of protein electrophoretic and isozyme variations from erythrocytes, serum, cerumen, human placenta, and saliva. The majority of these works were carried out by Lie-Injo Luan Eng and Tan Soon-Guan (Lie-Injo and Chin, 1964; Luan Eng, 1965; Lie-Injo, 1969; Steinberg and Eng, 1972; Lie-Injo et al., 1976; Tan, 1978; Tan, 1979; Tan and Teng, 1978; Tan et al., 1982) (**Supplementary Table S1**). Some findings worth acknowledging: 1) Gamma globulin variant (Gm) in Northern Orang Asli (presumably the Negrito) varied markedly from other Orang Asli groups; 2) Temiar and Semai had similar Gm variant frequency; 3) some malaria-related genes from the Negritos were different from the Southern Orang Asli (presumably the Proto Malays); 4) gene variant ADA-2 (Adenosine deaminase) and Pep B-6 (Peptidase B) in Semelai were similar to Temuan; 5) analysis based on five isozyme markers indicated that Sabahan native Kadazan was closer to their neighboring population as well as the Taiwanese and Philippine aborigines. In general, these biochemical markers were able to distinguish the East Asian from the non-East Asian groups of populations. Among the East Asian cluster, the Proto-Malay was found to cluster with the indigenous populations from Taiwan; whilst the natives from Sarawak formed another cluster (Tan, 2001).

Other intriguing observations included the frequencies Hb E, G6PD deficiency, and ovalocytosis among the native populations from Malaysia that were parallel with past distributions of

malaria posted the theory that malaria is selective for these hematological traits (Luan Eng, 1965; Lie-Injo, 1969).

Although the questions pertaining to the history of peopling and migrations in Malaysia and SEA were not noticeably addressed in these early studies, the clarification of population relationships using these conventional technologies has essentially laid an important foundation for future population genetics, or subsequent studies in Malaysia could not have been made successful.

Mitochondrial and Autosomal DNA Markers

The availability of molecular markers in particular maternal lineage mitochondrial DNA (mtDNA) markers, and subsequently molecular clocking approach have completely changed the way we appreciate the human migration and peopling history in Peninsular Malaysia. Several postulations on the migration history of SEA were initially proposed but posited considerable doubts. For instance, Ballinger et al. (1992) proposed a genetic continuity of Southern China migration of the Orang Asli, but the samples classification was questionable (Baer, 1999). In other studies, genetic trees produced by Cavalli-Sforza et al. (1994) and Saha et al. (1995) were inconsistent with the accepted historical events (Endicott, 2016), leading to an inconclusive genetic history of Orang Asli. Fix (1995) on the other hand, illustrated the introduction of an autosomal adaptive allele of SEA ovalocytosis (SAO) to the Orang Asli from the islands instead of mainland SEA, therefore postulated that the present Orang Asli is a product of local acculturation and differentiation.

Despites the said controversial, several hallmark studies had provided further insights into the prehistorical peopling in SEA. Based on mtDNA variation, Macaulay et al. (2005) suggested a seminal publication, that there was a single dispersal of the modern humans out-of-Africa ~65 kya, migrated along the southern coastal route through present-day India towards SEA and Australasia. Hill et al. (2006) claimed that the Orang Asli Negrito carry deep ancestry haplogroups dating to the initial settlement out-of-Africa more than 50 kya; whereas half of the maternal lineages of Senoi traced back to the ancestors of the Negrito and the remaining to Indochina, supporting the postulation that they represent the descendants of early Austroasiatic speaking migrants. The Proto Malay were more diverse and had closer affinity to the populations from the islands of SEA. These studies found that the ancient haplogroup M21a and R21 was most frequent among the Negrito and Senoi; whilst N21 and N22 that are thought to be specific to SEA, largely occurred in the Proto Malays. The M21 haplogroup branches directly from the Eurasian founder haplotypes, hence implied that the Negrito is probably the most direct descendants of the original inhabitants of the Malay Peninsula. Their findings also suggested that there were at least four migration events that shaped the genetic make-up of the indigenous populations in Peninsular Malaysia. The fact that the Senoi carried the Indo-Chinese haplogroup F1a1a and N9a6a (date ~5,500 years ago), suggested that they could have been descended from the admixture of local hunter-gatherers with the later agricultural invaders (Hill et al., 2006; Oppenheimer, 2011).

Whole Genome Analysis

The completion of the human genome project has revolutionized and accelerated the progress of population genetic studies at a tremendous pace. Genome-wide genotyping array (SNP array) and next-generation sequencing technologies were subsequently made available for large genome mapping initiatives to be carried out, including Human Genome Diversity Project (Cavalli-sforza, 2005), the International HapMap Project (The International HapMap Project, 2003), and the more recent 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010). Genotyping on the modern Malay populations were carried out (Teo et al., 2009), followed by large scale whole genome sequencing (Wong et al., 2013; Wu et al., 2019).

Surprisingly however, the indigenous populations from SEA were underrepresented in these large initiatives. Acknowledging this notion, the HUGO Pan-Asian SNP (PASNP) Consortium initiated the mapping of the Asian human genetic diversity with ~50,000 genome-wide autosomal SNPs, and revealed that the ancestry of Asian populations harbors genetic contributions derived from five language groups, and that the migrations of SEA populations were more complex than the anticipated. Most notably, the study supports an initial single migration wave of the ancestors of East Asian (EA) populations via "Southern route" into SEA, followed by multiple subsequent migrations thence shaped the complex genetic diversity of SEA (HUGO Pan-Asian SNP Consortium et al., 2009). The groundbreaking findings have placed the Orang Asli as an important piece of puzzle in mapping modern human migration.

Subsequently a growing body of evidence thenceforth refined the peopling history in Malaysia. The SEA Austronesian populations generally carried ancestry components derived from four primary admixture events: 1) aboriginal Taiwan; 2) Austroasiatic; 3) Melanesian; and 4) Negrito (Lipson et al., 2014); whilst the western islands of SEA populations was found to carry the ancestral components originated from the present-day mainland SEA Austroasiatic populations (Hudjashov et al., 2017). This migration model is consistent with the alternative "Early Train" migration hypothesis proposed by Jinam et al. (2012), which argues that there was a migration originating from Indochina or South China ~30–10 kya. Both hypotheses are supported at least in part, by several other lines of evidence: 1) close genetic affinity between the Sabahan natives and the Taiwanese aborigines and the Philippines aborigines but distantly related to the populations from mainland SEA (Yew et al., 2018a); 2) putative signals of positive selection driven by malaria infection found in the Sabahan natives occurred ~5 kya, which coincides with the period during Austronesian expansion (Hoh et al., 2020); 3) inference divergence time between the Negrito and Senoi coincides with the proposed period of "Early Train" hypothesis, posing the plausibility of the swiddening Austroasiatic agriculturist migration to Peninsular Malaysia, which resulted in declined effective population size of Negrito (Yew et al., 2018b); 4) inference using both uniparental and autosomal markers suggested primarily common ancestry for Taiwan or islands of SEA populations established before the Neolithic period (Soares

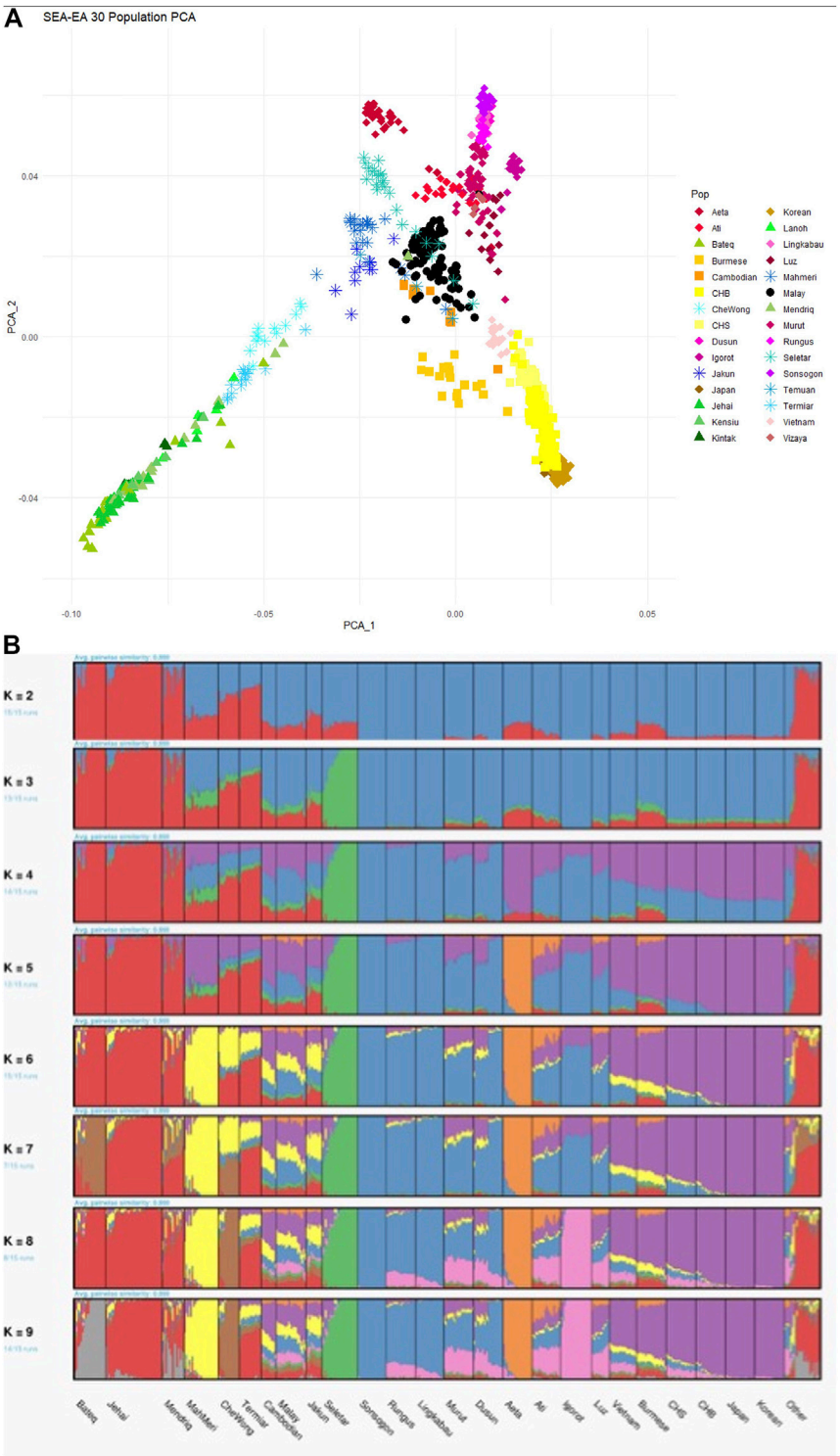

FIGURE 3 | PCA and ADMIXTURE analysis between the Malay (MAS) and HGDP-Cambodian populations. **(A)** Both MAS and HGDP-Cambodian are clustered close to each other in PC1 and PC2. **(B)** In ADMIXTURE analysis, the ancestral components for the MAS and HGDP-Cambodian are indistinguishable.

et al., 2016); 5) the native from Sarawak (the Iban) showed a closer genetic affinity to Indonesia than the mainland SEA (Simonson et al., 2011).

When dissecting the genetic architecture of the modern Malays from Peninsular Malaysia, population substructure was observed, suggesting plausible different origins from neighboring regions of SEA (Hatin et al., 2011). Finer investigation suggested that the modern Malays substructure is clustered to Northern and Southern Peninsular, correlated with the geographical latitude of the respective sampling locations (Hoh et al., 2015). The ancestors of the modern Malays diverged from the East Asian ~25 kya; but subsequently admixed with the other Austronesian populations ~1,700 years ago (Wu et al., 2019). This finding is in part consistent with our earlier study, which identified four major admixed ancestral components in the Malay populations occurred ~1,500–750 years ago: Austronesian, Proto Malay, East Asian, and South Asian (Deng et al., 2015). The slight violation of the admixture time between our study as opposed to Wu et al. (2019) is likely due to the different experimental platforms acquired. The collective findings suggest that geographical isolation and independent admixture have significantly shaped the genetic architectures and the diversity of the present days modern Malays. Intriguingly, the Malays revealed a close genetic affinity to the Cambodians (from the Human Genome Diversity Project dataset) (Deng et al., 2015; Liu et al., 2015) (**Figure 3**). The underlying reason to this link is uncertain, thus warrants further investigations.

The notion of multiple prehistorical migration waves to SEA was further supported by ancient genomes analyses (Lipson et al., 2018), revealing that the early farmers from Indochina (Vietnam) carried an admixed genetic component of East Asian (southern Chinese agriculturalist) and deeply diverged eastern Eurasian Austroasiatic-speaking hunter-gatherer ancestry, with similar ancestry south Indonesia, thus supporting an initial expansion of Austroasiatic languages.

Whilst the peopling history of Austronesian is seen clearer, the genetic affinity between the Orang Asli Negrito and ancestors of the Austroasiatic language-speaking populations remains controversial. Growing body of evidence confirm that the Orang Asli Negrito once shared common origins with the East Asian populations (Deng et al., 2014; Aghakhanian et al., 2015; Liu et al., 2015; Fu et al., 2018; Yew et al., 2018a). However, on a finer scale, they are distinct from the rest of the East Asian populations (**Figure 3**). The inference of divergence and admixture time using autosomal SNPs supports the argument that Negrito as the descendants of the earliest inhabitant of SEA (Deng et al., 2014; Aghakhanian et al., 2015; Yew et al., 2018a). Although multiple archaeological and mtDNA evidence suggested deep ancestry lineage 50–60 kya in the Orang Asli Negrito, their inferred time of divergence from Eurasian was younger than anticipated (~31–50 kya) (Jinam et al., 2017; Yew et al., 2018a). However the findings were consistent with the divergent time for Andamanese Negrito (~50 kya) (Mondal et al., 2016) and the northern Philippines Negrito (~46 kya) (Larena et al., 2021b). Intriguingly, despites exhibiting similar phenotypic characteristics, the Negritos from Malaysia, the Philippines and the Andaman were found to be distantly related. On the other hand, an ancient link between the Negrito from Peninsular Malaysia and the Andamanese Negrito was observed (Aghakhanian et al., 2015), but the proportion was too weak to confirm it significance. Our recent investigation successfully revealed a specific basal Asia ancestry component exclusively shared by the Negrito populations from Asia, dated at least 50 kya (Deng et al., 2021). Collectively, it is conceivable to speculate that the ancestors of these Negrito populations could have split prior arriving to the Sundaland, subsequently entered the mainland- and islands- of SEA independently, eventually remained isolated and formed genetic architecture unique from each other.

Intriguingly, the Orang Asli Negrito (Jehai sub-tribe) showed a shared genetic drift with ancient genomes from Hoabinhian ancestry (McColl et al., 2018), suggesting that they are genetically closer to the ancestors of Hoabinhian hunter-gatherers who occupied northern parts of Peninsular Malaysia during the late Pleistocene (Bellwood, 2007). What puzzles us, however, is that the divergence between Eurasian and Australian aborigines (Malaspinas et al., 2016) predates the divergence between Eurasian and Orang Asli Negrito (Yew et al., 2018a), thus complicates the peopling history in Peninsular Malaysia.

Recent investigation postulates that pre-Neolithic ancestors of today's Austroasiatic language-speaking populations were widespread in South Asia and SEA, and the populations from these regions were the result of multiple migrations of East Asian farmers during the Neolithic period (Tagore et al., 2021). The authors predicted that the present-day Austroasiatic populations from India and Malaysia shared a common ancestor until about 10.5 kya. Post-separation, they had a disparate genetic history. Around 7 kya, with the agriculture expansion, there was an ancestry shift in SEA. While the study has reiterated the impact of Austroasiatic populations in Peninsular Malaysia, how and when did the ancestors of Austroasiatic and Orang Asli Negrito introgressed was not addressed.

We acknowledge however, some of these findings should be interepreted with caution, because many of these studies were carried out with a rather small sample size, which may have introduced sampling biasness hence compromised conclusion.

Archaic Genomes Analysis

Whilst some old questions are addressed, peopling history in SEA is complicated by the genomic introgression between modern human and archaic hominin. Denisovan genome introgression was found significantly higher in the Papuan and the Philippine Negrito (Mamanwa); but not in the native populations from Peninsular Malaysia Borneo, and the Andamanese Negrito (Reich et al., 2011; Jinam et al., 2017; Yew et al., 2018a). Therefore it was speculated that the introgression could have occurred in the common ancestors of Australian aborigines and Papuans and the Philippines Negrito, before the divergence of with these populations and the Orang Asli Negrito; and that the similar proportion of introgression between SEA and East Asian is likely the ancestral component (Yew et al., 2018a). The recent report revealing that the Philippine Ayta carry the known highest level of Denisovan ancestry in the world – 30–40% greater than that of Papuans – therefore supports the view on an independent admixture event into the Philippine Negritos from Denisovans.

The Philippine region is thus likely inhabited by multiple archaic groups prior to the arrival of modern humans (Larena et al., 2021a), which again prolongs the debate on the links between the Orang Asli Negritos and the Negritos from the Philippines, Australia, and the Papuans. Considering that Denisovan introgression in the Philippine Negrito and Papuans could have occurred independent of the Orang Asli Negrito, it again supports the postulation that these phenotypically Negrito-like populations may have occupied the SEA landmass via different migration waves.

What is more intriguing, is that genome analysis of a human remained dated ~7 kya surprisingly expressed a substantial East Asian ancestry component, with a mixture of Denisovan gene flow, indicating an East Asian ancestry present in Wallacea much earlier than Austronesian expansion. This discovery has overthrown our common understanding of SEA human peopling history, thence shows that the peopling of SEA was much more complex than has previously been appreciated (Carlhoff et al., 2021), and the natives from Peninsular Malaysia and Borneo may be an important key to this answer.

The cumulative evidence presented above thenceforth implies "multi-layer" migration to SEA, and Peninsular Malaysia and Borneo to be more specific (Mörseburg et al., 2016), instead of the simplified "two-layer" migration model.

For the ease of reference, a summary of the published genotyping and sequencing data for the natives of Peninsular Malaysia and Borneo is tabulated in **Supplementary Table S2**.

What Is Known, and What Is Unknown?

With the advent of genomic technology, many long-standing questions were addressed with renewed vigor, some possibly with refined insights. Based on the collective findings to date, the following section summarizes, what we know about the peopling and migration history of the native populations from Malaysia:

(i) The ancestors of SEA populations migrated out of Africa via "Southern Route" along the coastal line to East Asia (Hill et al., 2006; HUGO Pan-Asian SNP Consortium et al., 2009).

(ii) The Negrito from Peninsular Malaysia once shared a common ancestry with East Asian populations (HUGO Pan-Asian SNP Consortium et al., 2009; Aghakhanian et al., 2015; Liu et al., 2015; Jinam et al., 2017; Lipson et al., 2018).

(iii) Both mtDNA molecular clocking and inference of divergence time using autosomal DNA support the notion that the ancestors of Peninsular Malaysia Negrito may be the earliest inhabitant of SEA at least 50 kya (Hill et al., 2006; Yew et al., 2018a; Deng et al., 2021).

(iv) The native populations from Peninsular Malaysia and Borneo are genetically distinct, each with a unique population history (Yew et al., 2018b).

(v) Austronesian expansion occurred at least in the SEA region, southwards to the Philippines, towards the Northern Borneo (Hill et al., 2007; Lipson et al., 2014; Yew et al., 2018b).

(vi) There are at least four prehistorical gene flows events that occurred in Peninsular Malaysia and Borneo thus concurs the "multi-layer" migration model. The first migration was likely to had derived from the earliest ancestors of the Negrito-like populations; subsequently the Austroasiatic migration southwards replacing the descendants of the early inhabitants. The third migration being the Austronesian expansion, and the last likely being the East Asian or islands of SEA (Hill et al., 2006; Jinam et al., 2012; Lipson et al., 2014; Deng et al., 2015; Soares et al., 2016; Mccoll et al., 2018; Lipson et al., 2018).

(vii) The native populations from both Peninsular Malaysia and Borneo received minimal gene flow from the archaic Denisovan hominin (Reich et al., 2011; Yew et al., 2018a).

Based on our current understandings along with earlier proposed models (Lipson et al., 2014; Norhalifah et al., 2016), we redraw a hypothetical modern human migration routes into Peninsular Malaysia (**Figure 4**).

What remained to be answered, we think, are:

(i) The genetic link between the Negrito-like populations from SEA, including Negrito from Peninsular Malaysia, the Philippines, the Andaman, Papuan and Australian aborigines, and possibly the African pygmies, remain inconclusive. Whilst it has been shown that these populations shared an ancient basal Asian ancestry component (Deng et al., 2021), they have undergone a long period of isolation from each other and could have experienced extensive local adaptation and admixture from respective neighboring populations therefore posting a challenge to uncover the precise date of migration of these populations. Analysis of natural selection may be able to at least in part address this question (Liu et al., 2015; Zhang et al., 2021). Several studies have been carried out in the SEA populations including the natives from Peninsular Malaysia and Borneo (Liu et al., 2015; Liu et al., 2017; Hoh et al., 2020). However finer genotyping, phenotypic and environmental characteristics, along with a sample size on the indigenous populations are required to warrant a convincing conclusion.

(ii) The pre-historical migrations of the Negrito into this landmass remain fragmentary. If the Orang Asli Negrito were postulated to be the direct descendants of the modern human out of Africa and that the divergence time between Australia aborigines and Eurasia occurred 65 kya (Malaspinas et al., 2016; Clarkson et al., 2017), we would then anticipate the divergence time between Orang Asli Negrito and Eurasia to have overlapped with the Australian aborigines if not earlier. However, our inferred time of divergence was younger than the Australia aborigines-Eurasia divergence time. Comprehensive investigations on a high coverage of SEA Negrito populations along with their neighboring populations are required before a conclusion can be made.

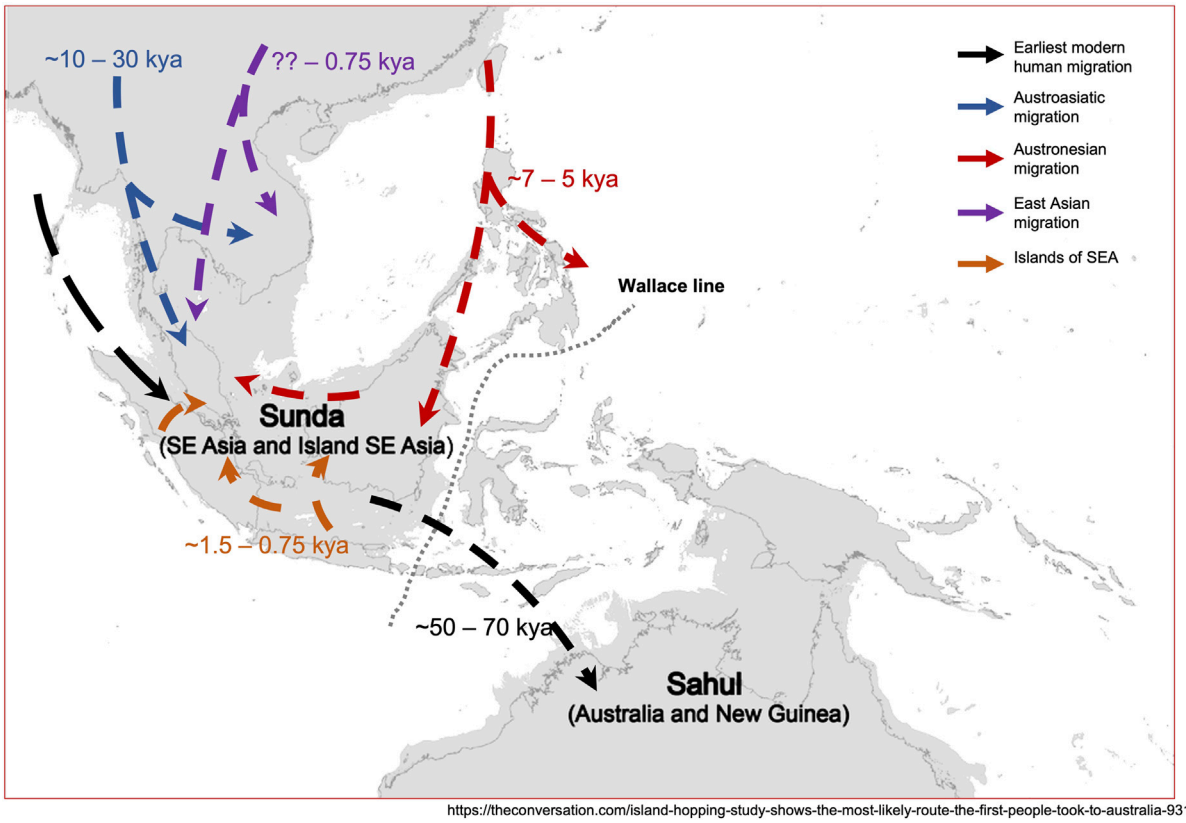

FIGURE 4 | A hypothetical model of migration history of the native populations from Malaysia. The map was adapted and modified from https://theconversation.com/island-hopping-study-shows-the-most-likely-route-the-first-people-took-to-australia-93120, with permission.

(iii) What are the relationships between the Orang Asli Negrito and Senoi populations? Different postulations have been made on the origins of Senoi. Some related them to the Veddoids; others argued that they were the descendants of admixture between Negrito and Austronesian (Tan, 2001; Oppenheimer, 2011). However, our preliminary PCA shows that the Senoi subtribes are much more heterogenous than the Negrito (**Figure 2**), suggesting that there is no simple package of elements to this conclusion. The inference of divergence time earlier proposed that they may be the descendants of the swiddening Austroasiatic agriculturists who later occupied Peninsular Malaysia (Yew et al., 2018a). Dissecting the genetic structure of the Senoi, and subsequently the connections between the Negrito and Senoi would be challenging owing to their close genetic affinity.

(iv) Reason(s) of the absence of Negrito-like populations in Borneo. Archaeological evidence suggests that Borneo was first inhabited by the ancestor of SEA modern human who were believed to be the direct descendants from the out of Africa exodus (Reyes-Centeno et al., 2015; Curnoe et al., 2016). It was postulated that these populations were then replaced by the Austronesians (Bellwood, 2007). If it was true, then how and when did it happen?

(v) It is intriguing to observe a close genetic affinity between the modern Malays and Cambodian population (but on the Proto-Malay Jakun). Further investigation should be carried out to explain the links between the modern Malay, Cambodian and Yunnan Migration Hypothesis; or to dispute this hypothesis.

(vi) Soares et al. (2016) argued that gene flow could have happened from the islands of SEA to Taiwan and subsequently back-migration again occurred. This argument does not agree with the common understanding of "Out-of-Taiwan" expansion as supported by linguistic and archaeological evidence. They subsequently claimed that there were two minor gene flows from Taiwan rather than a massive migration waves. In addition, recent study claimed no strong support for a predominant out-of-Taiwan dispersal of rice (Larena et al., 2021b), suggesting the "Out-of-Taiwan" model may be more complex than expected. More thorough investigations using both uniparental and high-density autosomal markers involving comprehensive native

populations from SEA and Polynesian are required to conclude the Austronesian migration model.

Challenges of Studying the Population Genetics of Natives in Malaysia

The population genetics study in Malaysia is not without challenge. Owing to their long period of isolation and high inbreeding rate, population genetic studies of the native populations – specifically referring to the Orang Asli and the Bornean natives – are often restricted to limited number of unrelated samples. This hinders the statistical power thence affects the strength of the evidence to support or dispute a hypothesis, resulting in a rather limited impact publication. In addition, although the cost for genome-wide analysis has reduced dramatically, it is still unaffordable for many low- and middle-income countries. Limited research funding available for such less prioritized science has always been a challenge to carry out high-throughput genome sequencing initiatives.

Nonetheless, the more pressing challenge is the ethical and integrity concerns that have been raised over the years, particularly issues on accessing the samples and data. A thorough process of ethics approval has been established especially for the genetic study of native populations in SEA countries, such as Malaysia, the Philippines and Indonesia. Efforts are spent in engaging and building trusts with the native populations, and collecting biological materials and data. Often such a process takes years before seeing success. Therefore, the contribution from the local scientists merits scientific acknowledgments. Unfortunately, we see occasionally, scientific publications or reports on the marginalized populations from SEA without appropriately acknowledging the local scientists. Such practice has raised some arguments over the years. We, therefore, call for close collaborations between investigators locally and abroad to warrant the advancement of science in mutual respect and ethical manner.

THE IMPLICATIONS OF MIGRATION AND PEOPLING HISTORY OF SEA POPULATIONS ON HUMAN HEALTH AND DISEASES

The complex demographic history of modern humans out-of-Africa has created different genetic architectures across populations from different continents, hence disease susceptibility. With genomic data from diverse populations and ancestries, we can compare these information over time and geography to better understand the origins and evolution of both individual genetic variants and human populations.

From global perspective, studying the genomics of the SEA marginalised populations, in this particular case, the natives from Peninsular Malaysia and Borneo, complements the catalogue of genetic variations that is known to be strikingly bias to European populations. Particularly, it allows characterizations of rare, and population (or ancestry) specific genetic variations (Bustamante et al., 2011; Hindorff et al., 2018). A classic success story of studying the genomic of natives is the Greenlandic Inuit populations, which found strong signals of adaptation to diet rich in protein and fatty acids (Fumagalli et al., 2015; Zhou et al., 2019). Another example is the Madagascar population (Pierron et al., 2018). Analysing the admixture and local ancestry of this population revealed a strong selection signal underlying Duffy blood group gene (*FY*) against *Plasmodium vivax* infection, and the selection signal was derived from the Asian ancestry.

From regional perspective, many tropical diseases that are uncommon in many developed countries, for instance Dengue and nasopharyngeal carcinoma. In addition, owing to different demographic history, climate, diet and lifestyles, pathophysiology for many complex diseases, for instance, cardiovascular diseases and metabolic syndromes, in the SEA populations may vary from the developed countries, and manifest in different intermediate phenotypes. Therefore, population genomic studies of the native will be utmost pertinent to the disease mapping efforts in SEA.

CONCLUSION

The population genetic study of the native populations in Malaysia has begun since the 19th century with the intention of racial classification. Postulations and arguments on human migration and peopling history especially to the SEA region were merely based on anthropology, archaeological, and linguistics evidence. With the advancement of genomics and life science technology, the picture of the peopling and migration history of the natives from Malaysia, and their attributions with other neighboring populations has become clearer. The natives from Malaysia were originated from multiple waves of migration events. It began with the direct descendants of the exodus of out-of-Africa, followed by the early Austroasiatic primitive farmers from Indochina that occupied the Peninsular Malaysia region and interacted with the existing inhabitants; subsequently with multiple waves of migrations of Austronesian populations from Southern China (and Taiwan) and other islands of SEA. However, we acknowledge that the migration model is far more complex than anticipated, and there are doubts yet to be addressed. We reiterate here, the importance of unveiling migration history and genetic diversity of the indigenous populations, not only to address the fundamental question of the origin of modern humans, but to complement the catalog for human genome variation, and to provide a stepping stone towards comprehending disease evolution and physiology (Fan et al., 2016; Hindorff et al., 2018; Wu et al., 2019). Increased attention to diversity will eventually increase the accuracy, utility and acceptability of using genomic information for clinical care (Benton et al., 2021).

AUTHOR CONTRIBUTIONS

B-PH conceptualized the concept of the article. B-PH, LD, and SX drafted the article together. All authors have read, commented and approved the draft.

FUNDING

Part of this study was funded by the Ministry of Higher Education of Malaysia (Fundamental Research Grant Scheme, project code: FRG0449-STG-1/2016; and FRGS/1/2015/ST03/UCSI/01/1). H-BP acknowledges the Chinese Academy of Sciences President's International Fellowship Initiatives (2017) (2017VBA0008) awarded to him. SX also gratefully acknowledges the support of the United Kingdom Royal Society-Newton Advanced Fellowship (NAF\R1\191094) and the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01). Authors declare that there funders have no role in any part of the study.

ACKNOWLEDGMENTS

The authors thank Kasih Norman for granting the permission to use the figure obtained from the website: https://theconversation.com/island-hopping-study-shows-the-most-likely-route-the-first-people-took-to-australia-93120.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.767018/full#supplementary-material

REFERENCES

Aghakhanian, F., Yunus, Y., Naidu, R., Jinam, T., Manica, A., Hoh, B. P., et al. (2015). Unravelling the Genetic History of Negritos and Indigenous Populations of Southeast Asia. *Genome Biol. Evol.* 7, 1206–1215. doi:10.1093/gbe/evv065

Anderson, J. (1824). *Political and Commercial Considerations Relative to the Malayan Peninsula and the British Settlements in the Straits of Malacca*. Kuala Lumpur, Malaysia: Malaysian Branch of the Royal Asiatic Society. Printed under the authority of Government by William Cox.

Annandale, N., and Robinson, H. C. (1902). Some Preliminary Results of an Expedition to the Malay Peninsula. *The J. Anthropological Inst. Great Britain Ireland* 32, 407–417. doi:10.2307/2842828

Arenas, M., Gorostiza, A., Baquero, J. M., Campoy, E., Branco, C., Rangel-Villalobos, H., et al. (2020). The Early Peopling of the Philippines Based on mtDNA. *Sci. Rep.* 10, 1–9. doi:10.1038/s41598-020-61793-7

Baer, A. S. (1999). *Health, Disease, and Survival : A Biomedical and Genetic Analysis of the Orang Asli of Malaysia*. 1st ed (Subang Jaya, Malaysia: Center for Orang Asli Concerns).

Ballinger, S. W., Schurr, T. G., Torroni, A., Gan, Y. Y., Hodge, J. A., Hassan, K., et al. (1992). Southeast Asian Mitochondrial DNA Analysis Reveals Genetic Continuity. *Genetics* 130, 139–152. doi:10.1093/genetics/130.1.139

Barker, G., Barton, H., Bird, M., Daly, P., Datan, I., Dykes, A., et al. (2007). The "Human Revolution" in lowland Tropical Southeast Asia: the Antiquity and Behavior of Anatomically Modern Humans at Niah Cave (Sarawak, Borneo). *J. Hum. Evol.* 52, 243–261. doi:10.1016/j.jhevol.2006.08.011

Bellwood, P., Gamble, C., Le Blanc, S. a., Pluciennik, M., Richards, M., and Terrell, J. E. (2007). First Farmers: the Origins of Agricultural Societies, by Peter Bellwood. Malden (MA): Blackwell, 2005; ISBN 0-631-20565-9, hardback £60; ISBN 0-631-20566-7 paperback £17.99, xix+360 pp., 59 figs., 3 tables. *Cambridge Archaeol. J.* 17, 87. doi:10.1017/S0959774307000078

Bellwood, P. (2007). *Prehistory of the Indo-Malaysian Archipelago*. ANU E Press.

Benton, M. L., Abraham, A., LaBella, A. L., Abbot, P., Rokas, A., and Capra, J. A. (2021). The Influence of Evolutionary History on Human Health and Disease. *Nat. Rev. Genet.* 22, 269–283. doi:10.1038/s41576-020-00305-9

Bustamante, C. D., Burchard, E. G., and De la Vega, F. M. (2011). Genomics for the World. *Nature* 475, 163–165. doi:10.1038/475163a

Carlhoff, S., Duli, A., Nagele, K., Nur, M., Skow, L., Sumantri, I., et al. (2021). Genome of a Middle Holocene hunter-gatherer from Wallacea. *Nature* 596, 543. doi:10.1038/s41586-021-03823-6

Cavalli-Sforza, L. L., Menozzi, P., and Piazza, A. (1994). *The History and Geography of Human Genes*. New Jersey: Princeton University Press. doi:10.1515/9780691187266

Cavalli-sforza, L. L. (2005). The Human Genome Diversity Project: Past, Present and Future. *Nat. Rev. Genet.* 6, 333–340. doi:10.1038/nrg1579

Clarkson, C., Jacobs, Z., Marwick, B., Fullagar, R., Wallis, L., Smith, M., et al. (2017). Human Occupation of Northern Australia by 65,000 Years Ago. *Nature* 547, 306–310. doi:10.1038/nature22968

Combrink, H. J. B., Soderberg, C., Boutin, M. E., Boutin, A. Y., Wise, M. R., and Zook, M. (2008). *Indigenous Groups of Sabah: An Annotated Bibliography of Linguistic and Anthropological Sources*. 2nd editio. Sabah: SIL International.

Crawfurd, J. (1820). *History of the Indian Archipelago: Containing an Account of the Manners, Arts, Languages, Religions, Institutions, and Commerce of its Inhabitants*. Edinburgh: Archibald Constable and Co.

Curnoe, D., Datan, I., Taçon, P. S. C., Leh Moi Ung, C., and Sauffi, M. S. (2016). Deep Skull from Niah Cave and the Pleistocene Peopling of Southeast Asia. *Front. Ecol. Evol.* 4, 75. doi:10.3389/fevo.2016.00075

Demeter, F., Shackelford, L. L., Bacon, A. M., Duringer, P., Westaway, K., Sayavongkhamdy, T., et al. (2012). Anatomically Modern Human in Southeast Asia (Laos) by 46 Ka. *Proc. Natl. Acad. Sci. U. S. A.* 109, 14375–14380. doi:10.1073/pnas.1208104109

Deng, L., Hoh, B.-P., Lu, D., Saw, W.-Y., Twee-Hee Ong, R., Kasturiratne, A., et al. (2015). Dissecting the Genetic Structure and Admixture of Four Geographical Malay Populations. *Sci. Rep.* 5, 14375. doi:10.1038/srep14375

Deng, L., Hoh, B. P., Lu, D., Fu, R., Phipps, M. E., Li, S., et al. (2014). The Population Genomic Landscape of Human Genetic Structure, Admixture History and Local Adaptation in Peninsular Malaysia. *Hum. Genet.* 133, 1169–1185. doi:10.1007/s00439-014-1459-8

Deng, L., Pan, Y., Wang, Y., Chen, H., Yuan, K., Chen, S., et al. (2021). Genetic Connections and Convergent Evolution of Tropical Indigenous Peoples in Asia. *Mol. Biol. Evol.*, 1–33. doi:10.1093/molbev/msab361

Endicott, K. (2016). *Malaysia's Original People: Past, Present and Future of the Orang Asli*. Editor K. Endicott (NUS Press). Available at: http://www.jstor.org/stable/j.ctv1qv35n.

Enfield, N. J. (2005). Areal Linguistics and mainland Southeast Asia. *Annu. Rev. Anthropol.* 34, 181–206. doi:10.1146/annurev.anthro.34.081804.120406

Fan, S., Hansen, M. E. B., Lo, Y., and Tishkoff, S. A. (2016). Going Global by Adapting Local: a Review of Recent Human Adaptation. *Science* 354, 54–59. doi:10.1126/science.aaf5098

Fix, A. G., and Lie-Injo, L. E. (1975). Genetic Microdifferentiation in the Semai Senoi of Malaysia. *Am. J. Phys. Anthropol.* 43, 47–55. doi:10.1002/ajpa.1330430108

Fix, A. G. (1995). Malayan Paleosociology: Implications for Patterns of Genetic Variation Among the Orang Asli. *Am. Anthropol.* 97, 313–323. doi:10.1525/aa.1995.97.2.02a00090

Fu, R., Mokhtar, S. S., Phipps, M. E., Hoh, B.-P., Xu, S., Shuhada, S., et al. (2018). A Genome-wide Characterization of Copy Number Variations in Native

Populations of Peninsular Malaysia. *Eur. J. Hum. Genet.* 26, 247–257. doi:10.1038/s41431-018-0120-8

Fumagalli, M., Moltke, I., Grarup, N., Racimo, F., Bjerregaard, P., Jørgensen, M. E., et al. (2015). Greenlandic Inuit Show Genetic Signatures of Diet and Climate Adaptation. *Science* 349, 1343–1347. doi:10.1126/science.aab2319

Graydon, J. J., Simmons, R. T., Semple, N. M., Clapham, L. J., and Wallece, E. H. (1952). Blood Genetics of Various Populations in Borneo. *Med. J. Aust.* 1, 694–702. doi:10.5694/j.1326-5377.1952.tb84092.x

Green, R. (1949). "Anthropological Blood Grouping Among the 'Sakai," in *Bulletin of the Raffles Museum, Series B* (Singapore: Printed at the Government Printing Office), 130–132.

Hatin, W. I., Nur-Shafawati, A. R., Zahri, M. K., Xu, S., Jin, L., Tan, S. G., et al. (2011). Population Genetic Structure of Peninsular Malaysia Malay Sub-ethnic Groups. *PLoS One* 6, 2–6. doi:10.1371/journal.pone.0018312

Hill, C., Soares, P., Mormina, M., Macaulay, V., Clarke, D., Blumbach, P. B., et al. (2007). A Mitochondrial Stratigraphy for Island Southeast Asia. *Am. J. Hum. Genet.* 80, 29–43. doi:10.1086/510412

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

Hindorff, L. A., Bonham, V. L., Brody, L. C., Ginoza, M. E. C., Hutter, C. M., Manolio, T. A., et al. (2018). Prioritizing Diversity in Human Genomics Research. *Nat. Rev. Genet.* 19, 175–185. doi:10.1038/nrg.2017.89

Hoh, B.-P., Zhang, X., Deng, L., Yuan, K., Yew, C.-W., Saw, W.-Y., et al. (2020). Shared Signature of Recent Positive Selection on the TSBP1 – BTNL2 – HLA-DRA Genes in Five Native Populations from North Borneo. *Genome Biol. Evol.* 12, 2245–2257. doi:10.1093/gbe/evaa207

Hoh, B. P., Deng, L., Julia-Ashazila, M. J., Zuraihan, Z., Nur-Hasnah, M., Nur-Shafawati, A. R., et al. (2015). Fine-scale Population Structure of Malays in Peninsular Malaysia and Singapore and Implications for Association Studies. *Hum. Genomics* 9, 16. doi:10.1186/s40246-015-0039-x

Hudjashov, G., Karafet, T. M., Lawson, D. J., Downey, S., Savina, O., Sudoyo, H., et al. (2017). Complex Patterns of Admixture across the Indonesian Archipelago. *Mol. Biol. Evol.* 34, 2439–2452. doi:10.1093/molbev/msx196

HUGO Pan-Asian SNP ConsortiumAbdulla, 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

Ibrahim, A., Siddique, S., and Hussain, Y. (1985). *Readings on Islam in Southeast Asia*. Singapore: Institute of Southeast Asian Studies. Available at: https://books.google.com.my/books?id=BeDKqPTeHnUC.

Jin, L., Seielstad, M., and Xiao, C. (2001). *Genetic, Linguistic and Archeological Perspectives on Human Diversity in Southeast Asia*. Singapore: World Scientific.

Jinam, T. A., Hong, L., Phipps, M. E., Stoneking, M., Ameen, M., Edo, J., et al. (2012). Evolutionary History of Continental Southeast Asians : "Early Train" Hypothesis Based on Genetic Analysis of Mitochondrial and Autosomal DNA Data. *Mol. Biol. Evol.* 29, 3513–3527. doi:10.1093/molbev/mss169

Jinam, T. A., Phipps, M. E., Aghakhanian, F., Majumder, P. P., Datar, F., Stoneking, M., et al. (2017). Discerning the Origins of the Negritos, First Sundal and People: Deep Divergence and Archaic Admixture. *Genome Biol. Evol.* 9, 2013–2022. doi:10.1093/gbe/evx118

Larena, M., McKenna, J., Sanchez-Quinto, F., Bernhardsson, C., Ebeo, C., Reyes, R., et al. (2021a). Philippine Ayta Possess the Highest Level of Denisovan Ancestry in the World. *Curr. Biol.* 31, 1–12. doi:10.1016/j.cub.2021.07.022

Larena, M., Sanchez-Quinto, F., Sjödin, P., McKenna, J., Ebeo, C., Reyes, R., et al. (2021b). Multiple Migrations to the Philippines during the Last 50,000 Years. *Proc. Natl. Acad. Sci. U. S. A.* 118, 1–9. doi:10.1073/pnas.2026132118

Lie-Injo, L. L. (1969). Distribution of Genetic Red Cell Defects in South-East Asia. *Trans. R. Soc. Trop. Med. Hyg.* 63, 664–674. doi:10.1093/nq/s7-v.116.219c10.1016/0035-9203(69)90188-610.1016/0035-9203(69)90188-6

Lie-Injo, L. E., and Chin, J. (1964). Abnormal Haemoglobin and Glucose-6-Phosphate Dehydrogenase Deficiency in Malayan Aborigines. *Nature* 204, 291–292.

Lipson, M., Loh, P.-R., Patterson, N., Moorjani, P., Ko, Y.-C., Stoneking, M., et al. (2014). Reconstructing Austronesian Population History in Island Southeast Asia. *Nat. Commun.* 5, 4689. doi:10.1038/ncomms5689

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

Liu, X., Lu, D., Saw, W.-Y., Shaw, P. J., Wangkumhang, P., Ngamphiw, C., et al. (2017). Characterising Private and Shared Signatures of Positive Selection in 37 Asian Populations. *Eur. J. Hum. Genet.* 25, 1–10. doi:10.1038/ejhg.2016.181

Liu, X., Yunus, Y., Lu, D., Aghakhanian, F., Saw, W. Y., Deng, L., et al. (2015). Differential Positive Selection of Malaria Resistance Genes in Three Indigenous Populations of Peninsular Malaysia. *Hum. Genet.* 134, 375–392. doi:10.1007/s00439-014-1525-2

Luan Eng, L. I. (1965). Hereditary Ovalocytosis and Haemoglobin E-Ovalocytosis in Malayan Aborigines [25]. *Nature* 208, 1329. doi:10.1038/2081329a0

Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., et al. (2005). Single, Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes. *Science* 308, 1034–1036. doi:10.1126/science.1109792

Malaspinas, A.-S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., et al. (2016). A Genomic History of Aboriginal Australia. *Nature* 538, 207–214. doi:10.1038/nature18299

Matsumura, H., Oxenham, M. F., Yukio, D., Domett, K., Nguyen, K. T., Nguyen, L. C., et al. (2008). Morphometric Affinity of the Late Neolithic Human Remains from Man Bac, Ninh Binh Province, Vietnam: Key Skeletons with Which to Debate the 'two Layer' Hypothesis. *Antropol. Sci.* 116, 135–148. doi:10.1537/ase.070405

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

Mijares, A. S., Détroit, F., Piper, P., Grün, R., Bellwood, P., Aubert, M., et al. (2010). New Evidence for a 67,000-Year-Old Human Presence at Callao Cave, Luzon, Philippines. *J. Hum. Evol.* 59, 123–132. doi:10.1016/j.jhevol.2010.04.008

Mondal, M., Casals, F., Xu, T., Dall'Olio, G. M., Pybus, M., Netea, M. G., et al. (2016). Genomic Analysis of Andamanese Provides Insights into Ancient Human Migration into Asia and Adaptation. *Nat. Genet.* 48, 1066. doi:10.1038/ng.3621

Mörseburg, A., Pagani, L., Ricaut, F.-X., Yngvadottir, B., Harney, E., Castillo, C., et al. (2016). Multi-layered Population Structure in Island Southeast Asians. *Eur. J. Hum. Genet.* 24, 1605. doi:10.1038/ejhg.2016.60

Nicolaisen, I. (1976). Form and Function of Punan Bah Ethno-Historical Tradition. *Sarawak Mus. J* 24, 63–95.

Norhalifah, H. K., Syaza, F. H., Chambers, G. K., and Edinur, H. A. (2016). The Genetic History of Peninsular Malaysia. *Gene* 586, 129–135. doi:10.1016/j.gene.2016.04.008

Oppenheimer, S. (2011). "MtDNA Variation and Southward Holocene Human Dispersals within mainland Southeast Asia," in *Dynamics Of Human Diversity*. Editor N. Enfield (Canberra: Pacific Linguistics), 81–108.

Pierron, D., Heiske, M., Razafindrazaka, H., Pereda-Loth, V., Sanchez, J., Alva, O., et al. (2018). Strong Selection during the Last Millennium for African Ancestry in the Admixed Population of Madagascar. *Nat. Commun.* 9, 1–9. doi:10.1038/s41467-018-03342-5

Polunin, I., and Sneath, P. H. A. (1953). Studies of Blood Groups in South-East Asia. *J. R. Anthropol. Inst. Gt. Britain Irel.* 83, 215–251. Available at: http://www.jstor.org/stable/2844033. doi:10.2307/2844033

Polunin, I. (1953). The Medical Natural History of Malayan Aborigines. *Med. J. Malaya* 8 (2), 114A–174A. concl.

Reich, D., Patterson, N., Kircher, M., Delfin, F., Nandineni, M. R., Pugach, I., et al. (2011). Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania. *Am. J. Hum. Genet.* 89, 516–528. doi:10.1016/j.ajhg.2011.09.005

Reyes-Centeno, H., Hubbe, M., Hanihara, T., Stringer, C., and Harvati, K. (2015). Testing Modern Human Out-Of-Africa Dispersal Models and Implications for Modern Human Origins. *J. Hum. Evol.* 87, 95–106. doi:10.1016/j.jhevol.2015.06.008

Saha, N., Mak, J. W., Tay, J. S., Liu, Y., Tan, J. A., Low, P. S., et al. (1995). Population Genetic Study Among the Orange Asli (Semai Senoi) of Malaysia: Malayan Aborigines. *Hum. Biol.* 67, 37–57.

Simonson, T. S., Xing, J., Barrett, R., Jerah, E., Loa, P., Zhang, Y., et al. (2011). Ancestry of the Iban Is Predominantly Southeast Asian: Genetic Evidence from Autosomal, Mitochondrial, and Y Chromosomes. *PLoS One* 6, e16338. doi:10.1371/journal.pone.0016338

Skeat, W. W., and Blagden, C. O. (1906). *Pagan Races of the Malay Peninsula Volume II*. London: Macmillan and Co., limited.

Soares, P. A., Trejaut, J. A., Rito, T., Cavadas, B., Hill, C., Eng, K. K., et al. (2016). Resolving the Ancestry of Austronesian-Speaking Populations. *Hum. Genet.* 135, 309–326. doi:10.1007/s00439-015-1620-z

Steinberg, A. G., and Eng, L. I. (1972). Immunoglobulin G Allotypes in Malayan Aborigines. *Hum. Hered.* 22, 254–258. doi:10.1159/000152495

Tagore, D., Aghakhanian, F., Naidu, R., Phipps, M. E., and Basu, A. (2021). Insights into the Demographic History of Asia from Common Ancestry and Admixture in the Genomic Landscape of Present-Day Austroasiatic Speakers. *BMC Biol.* 61, 238. doi:10.1186/s12915-021-00981-x

Tan, S. G., Gan, Y. Y., and Asuan, K. (1982). Transferrin C Subtyping in Malaysians and in Indonesians from North Sumatra. *Hum. Genet.* 60, 369–370. doi:10.1007/BF00569221

Tan, S. G. (1978). Genetic Markers in the Aborigines of Peninsular Malaysia and the Indigenous Races of Sabah, Sarawak, and Brunei. Blood Groups in the Three Major Races of Malaysia and Singapore: A Compilation of Data. *Pertanika* 1, 86–96.

Tan, S. G. (1979). Genetic Relationship between Kadazans and Fifteen Other Southeast Asian Races. *Pertanika* 2, 28–33.

Tan, S. G. (2001). "Genetic Relationships Among Sixteen Ethnic Groups from Malaysia and Southeast Asia," in *Genetic, Linguistic and Archaeological Perspectivces on Human Diversity in Southeast Asia*. Editors L. Jin, M. Sielstad, and C. Xiao (Singapore: World Scientific), 83–92. doi:10.1142/9789812810847_0007

Tan, S. G., and Teng, Y. S. (1978). Saliva Acid Phosphatases and Amylase in Senoi and Aboriginal Malays and Superoxide Dismutase in Various Racial Groups of Peninsular Malaysia. *Jpn. J. Hum. Genet.* 23, 133–138. doi:10.1007/BF02001794

Teo, Y. Y., Sim, X., Ong, R. T. H., Tan, A. K. S., Chen, J., Tantoso, E., et al. (2009). Singapore Genome Variation Project: A Haplotype Map of Three Southeast Asian Populations. *Genome Res.* 19, 2154–2162. doi:10.1101/gr.095000.109

The 1000 Genomes Project Consortium (2010). A Map of Human Genome Variation from Population Scale Sequencing. *Nature* 467, 1061–1073. doi:10.1038/nature09534.A

The International HapMap Project (2003). The International HapMap Project. *Nature* 426, 789–796. doi:10.1038/nature02168

Wang, W. S. (2001). "Human Diversity and Language Diversity," in *Genetic, Linguistic And Archaeological Perspectivces On Human Diversity In Southeast Asia*. Editors J. Li, S. Mark, and C. J. Xiao (Singapore: World Scientific), 172. doi:10.1142/9789812810847_0002

Wilkinson, R. (1926). *The Aboriginal Tribes (Papers on Malay Subjects)*. Kuala Lumpur: Government of the Federated Malay States, Kuala Lumpur.

Williams, T. R. (1965). *The Dusun: A North Borneo Society*. New York: Holt, Rinehart and Wnston.

Wong, L.-P., Ong, R. T.-H., Poh, W.-T., Liu, X., Chen, P., Li, R., et al. (2013). Deep Whole-Genome Sequencing of 100 Southeast Asian Malays. *Am. J. Hum. Genet.* 92, 52–66. doi:10.1016/j.ajhg.2012.12.005

Wu, D., Dou, J., Chai, X., Bellis, C., Wilm, A., Shih, C. C., et al. (2019). Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. *Cell* 179, 736–749. doi:10.1016/j.cell.2019.09.019

Yew, C. W., Lu, D., Wong, L., Twee-Hee Ong, R., Lu, Y., Wang, X., et al. (2018a). Genomic Structure of the Native Inhabitants of Peninsular Malaysia and North Borneo Suggests Complex Human Population History in Southeast Asia. *Hum. Genet.* 137, 161–173. doi:10.1007/s00439-018-1869-0

Yew, C. W., Minsong, A., Tiek, S., Lau, Y., Pugh-kitingan, J., Ransangan, J., et al. (2018b). Genetic Relatedness of Indigenous Ethnic Groups in Northern Borneo to Neighboring Populations from Southeast Asia , as Inferred from Genome-wide SNP Data. *Ann. Hum. Genet.* 82, 216. doi:10.1111/ahg.12246

Zhang, X., Qi, L., Hui, Z., Shilei, Z., Jiahui, H., Sovannary, T., et al. (2021). The Distinct Morphological Phenotypes of Southeast Asian Aborigines Are Shaped by Novel Mechanisms for Adaptation to Tropical Rainforests. *Natl. Sci. Rev.*, nwab072. doi:10.1093/nsr/nwab072

Zhou, S., Xie, P., Quoibion, A., Ambalavanan, A., Dionne-Laporte, A., Spiegelman, D., et al. (2019). Genetic Architecture and Adaptations of Nunavik Inuit. *Proc. Natl. Acad. Sci. U. S. A.* 116, 16012–16017. doi:10.1073/pnas.1810388116

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 17 February 2022
doi: 10.3389/fgene.2022.815285

Fine-Scale Population Admixture Landscape of Tai–Kadai-Speaking Maonan in Southwest China Inferred From Genome-Wide SNP Data

Jing Chen[1†], Guanglin He[2,3,4,5†], Zheng Ren[1], Qiyan Wang[1], Yubo Liu[1], Hongling Zhang[1], Meiqing Yang[1], Han Zhang[1], Jingyan Ji[1], Jing Zhao[2,3,4], Jianxin Guo[2,3,4], Jinwen Chen[2,3,4], Kongyang Zhu[2,3,4], Xiaomin Yang[2,3,4], Rui Wang[2,3,4], Hao Ma[2,3,4], Le Tao[2,3,4], Yilan Liu[2,3,4], Qu Shen[2,3,4], Wenjiao Yang[2,3,4], Chuan-Chao Wang[2,3,4] and Jiang Huang[1]**

[1]Department of Forensic Medicine, Guizhou Medical University, Guiyang, China, [2]State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen, China, [3]Department of Anthropology and Ethnology, School of Sociology and Anthropology, Institute of Anthropology, National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China, [4]State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen, China, [5]Institute Of Rare Diseases, West China Hospital of Sichuan University, Chengdu, China

OPEN ACCESS

Edited by:
Pavel Flegontov,
Harvard University, United States

Reviewed by:
Pittayawat Pittayaporn,
Chulalongkorn University, Thailand
Chao Ning,
Peking University, China

***Correspondence:**
Chuan-Chao Wang
wang@xmu.edu.cn
Jiang Huang
mmm_hj@126.com

[†]These authors have contributed equally to this work

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

Received: 15 November 2021
Accepted: 27 January 2022
Published: 17 February 2022

Citation:
Chen J, He G, Ren Z, Wang Q, Liu Y, Zhang H, Yang M, Zhang H, Ji J, Zhao J, Guo J, Chen J, Zhu K, Yang X, Wang R, Ma H, Tao L, Liu Y, Shen Q, Yang W, Wang C-C and Huang J (2022) Fine-Scale Population Admixture Landscape of Tai–Kadai-Speaking Maonan in Southwest China Inferred From Genome-Wide SNP Data. Front. Genet. 13:815285. doi: 10.3389/fgene.2022.815285

Guizhou Province harbors extensive ethnolinguistic and cultural diversity with Sino-Tibetan-, Hmong–Mien-, and Tai–Kadai-speaking populations. However, previous genetic analyses mainly focused on the genetic admixture history of the former two linguistic groups. The admixture history of Tai–Kadai-speaking populations in Guizhou needed to be characterized further. Thus, we genotyped genome-wide SNP data from 41 Tai–Kadai-speaking Maonan people and made a comprehensive population genetic analysis to explore their genetic origin and admixture history based on the pattern of the sharing alleles and haplotypes. We found a genetic affinity among geographically different Tai–Kadai-speaking populations, especially for Guizhou Maonan people and reference Maonan from Guangxi. Furthermore, formal tests based on the f_3/f_4-statistics further identified an adjacent connection between Maonan and geographically adjacent Hmong–Mien and Sino-Tibetan people, which was consistent with their historically documented shared material culture (Zhang et al., iScience, 2020, 23, 101032). Fitted qpAdm-based two-way admixture models with ancestral sources from northern and southern East Asians demonstrated that Maonan people were an admixed population with primary ancestry related to Guangxi historical people and a minor proportion of ancestry from Northeast Asians, consistent with their linguistically supported southern China origin. Here, we presented the landscape of genetic structure and diversity of Maonan people and a simple demographic model for their evolutionary process. Further whole-genome-sequence–based projects can be presented with more detailed information about the population history and adaptative history of the Guizhou Maonan people.

Keywords: fine-scale genetic structure, Tai–Kadai-speaking Maonan, admixture history, ethnolinguistic diversity, east Asian

INTRODUCTION

Guizhou Province is localized in the Yunnan–Guizhou Plateau and has a mountainous environment. This region has rich archaeologically attested cultures, historically documented ethnic groups, and languages. Populations belonging to different language families, including Tai–Kadai, Hmong–Mien, and Sino-Tibetan families, permanently resided here. Archaeological findings associated with the Chengtoushan, Daxi, and Shijiahe sites supported that southern China was the original birthplace of rice agriculture and the homelands of many languages. Further direct evidence from both archeology and language supported that the prosperity of rice farming led to the formation of the ancestral populations of present-day Tai–Kadai-, Hmong–Mien-, Austroasiatic-, and Austronesian-speaking populations and their used languages (Diamond and Bellwood, 2003; Zhao, 2011; Stevens and Fuller, 2017). In addition, linguistic evidence not only demonstrated that there was a common origin between Tai–Kadai and Austronesian language but also revealed that the Tai–Kadai language shared more language components (language borrowing) with surrounding Hmong–Mien and Sino-Tibetan families (Edmondson, 1988; Edmondson, 1997; Lu, 2008), which was also confirmed *via* the genome-wide SNP data (He et al., 2020). Importantly, a recent published genetic analysis based on the whole-genome SNP data from Southeast Asians also revealed the complex divergence processes, which showed that Austroasiatic people diverged from mainland Chinese populations approximately 15 thousand years ago (kya), Austronesian people diverged from mainland Sinitic-speaking Han and Tai–Kadai-speaking Dai around 10 kya, and Cordilleran people split from indigenous Taiwanese people at eight kya (Larena et al., 2021). The divergence time between the ancestors of Hmong-Mien and Tai-Kadai people keep unknown, but they experienced massive interaction with each other and with the southward Northeast Asians (Huang et al., 2020; Yang et al., 2020; Wang et al., 2021c; Wang C.C. et al., 2021). The initial landscape of the population history of southern China has been characterized *via* evidence from genetics and linguistics, whose detailed genetic history and admixture process with its neighbors need to be further explored based on the high-density genome-wide SNP data, especially for some geographically and ethnolinguistically diverse and specific indigenous populations.

The linguistic survey has found that the Tai–Kadai language was widely distributed in Southeast Asia, including Zhuang-Dai, Dong-Shui, Li, and Ge-Yang language sub-branches (Edmondson, 1988; Liang and Zhang, 1996; Edmondson, 1997; Kutanan et al., 2019; Liu et al., 2020; Kutanan et al., 2021). Linguistic findings showed that Maonan is a subgroup of the Dong-Shui language, mainly distributed in Guizhou, Guangdong, and Guangxi provinces (Lu, 2008). Historians hold the opinion that Tai–Kadai people in South China are one of the indigenous population with a long history (Wang, 2004; Huang, 2016; Zhang, 2016), which is also evidenced *via* the linguistic documents and ancient DNA evidence (Edmondson, 1988; Liang and Zhang, 1996; Lipson et al., 2018; McColl et al., 2018; Liu et al., 2020; Kutanan et al., 2021). Direct documents from historic materials showed that the Maonan ethnic group is related to the modern Southern Chinese indigenous ethnic groups such as Bouyei, Mulam, and Gelao (Wang, 2004; Huang, 2016).However, the historical records for the origin of Maonan and the records of local Chronicles inscriptions and genealogies are unknown. From an archeological perspective (Ma et al., 2020), the presumed ancestral populations of modern Tai–Kadai and Hmong–Mien speakers are probably related to Daxi culture and Qujialing culture around the Yunnan–Guizhou Plateau. Therefore, the genetic investigation should be comprehensively carried out in areas with high ethnic and linguistic diversity to explore the genetic connection among modern ethnolinguistically different populations and the genetic interactions between Guizhou indigenous people with ancient northern and southern East Asians (Ning et al., 2020; Yang et al., 2020; Mao et al., 2021; Wang et al., 2021c; Wang C. C. et al., 2021).

Considering the importance of a comprehensive and deep genetic survey of South China, previous genetic studies based on forensic genetic markers have shed light on the basic genetic profile and demographic history among Tai–Kadai speakers from southern China in the past two decades. From the perspective of uniparentally inherited Y chromosome haplogroup and mitochondrial haplogroup, Chen et al. made a preliminary exploration focused on the forensic parameters and genetic structure of the Tai–Kadai-speaking populations in Guizhou and the population genetic relationship based on the short tandem repeat (STR) on the autosome and X/Y-chromosomes (Chen et al., 2018c). Further genetic studies had focused on the population admixture history and genetic diversity of Tai–Kadai-speaking Gelao and Bouyei by insertion/deletion polymorphisms (InDels) and ancestry-informative single-nucleotide polymorphism (AISNPs) (He et al., 2019a; He et al., 2019b; He et al., 2021b). The obtained research results showed a significant genetic interaction between Tai–Kadai- and Hmong–Mien-speaking populations (He et al., 2019a; He et al., 2019b; He et al., 2021b). However, these studies were conducted based on low-density genetic markers and were mainly focused on exploring forensic characteristics. The low resolution of low-density markers was restricted to provide a fine-scale population genetic structure that can show the detailed information for the population admixture, evolutionary, and adaptive history. Recent population genomic history characterization of ethnolinguistically diverse people in Guizhou, including Han, Chuanqing, Gejia, Dongjia, Xijia, Mongolian, Manchu, Bouyei, Sui, Tujia, Dong, and Gelao people, has provided new insights into their formation processes and complete landscape of genetic history (Lu et al., 2020; Chen et al., 2021; Bin et al., 2021; He et al., 2021a; Wang et al., 2021a; Wang et al., 2021b). However, the admixture history of another important Guizhou indigenous Maonan still remains unknown.

Maonan people used the Maonan language belonging to the Dongshui branch in the Tai–Kadai language family (Edmondson, 1988, 1997; Lu, 2008), which was widely distributed in Guizhou and Guangxi provinces. Historians supported that the Maonan people were one of the major descendants of ancient indigeneous tribes in coastal southern China, and that they were especially associated with the hanging coffin burial custom (Zhang et al.,

2020). Based on the ancient DNA from the mitochondrial chromosome from historic hanging coffin people in Fujian, Guangxi, ancient people related to Tai–Kadai populations migrated westward from Fujian Wuyi Mount in the historical times and then across Southwest China to Southeast Asia (Zhang et al., 2020). Historical materials based on the hanging coffin customs also showed similar patterns of population migrations. Recent population genetic studies have included the Maonan people in Guangxi as the reference populations in the ancient DNA study. However, their fine-scale population history and their genetic relationship with the surrounding Hmong–Mien- and Sino-Tibetan-speaking populations have not been fully characterized, which are especially focused on the sharing of genome-wide haplotype data. Genetic studies of populations from the Yunnan–Guizhou Plateau regions have found enriched genetic diversity and complex mixed population genetic history (He et al., 2019a; He et al., 2019b; Zhang et al., 2019; Liu et al., 2021a; He et al., 2021b; Liu et al., 2021b; Chen et al., 2021). The detailed relationship between Maonan and modern and ancient neighboring populations needs to be characterized further. Thus, we conducted a genetic analysis based on the array-based genotyping of approximately 700 K SNPs in Tai–Kadai-speaking Maonan people in Guizhou to reconstruct its genetic diversity and evolutionary relationship with surrounding populations. Then, we merged our data with modern and ancient available East Asian data to explore their fine-scale population genetic structure and evolutionary history.

MATERIAL AND METHODS

Sample Collection and DNA Preparation

We collected saliva samples from 41 unrelated Maonan individuals in Pingtang County in Guizhou Province, Southwest China (**Supplementary Figure S1**). Participants whose parents and grandparents are indigenous people and reside in the sampling palaces at least three generations should have non-consanguineous marriage at the same ethnical group. The study was approved by the Medical Ethics Committee of Guizhou Medical University, and the recommendations provided by the revised Helsinki Declaration of 2000 were followed. All the participants signed written informed consent prior to participating in the study. We genotyped the genome-wide SNP data using the Infinium Global Screening Array, which included approximately 700 K SNPs and covered SNPs from the autosome, Y-chromosome, and mitochondrial DNA. We used a similar in-house commonly used standard to conduct quality control filter procedures (Wang et al., 2021b). Then, we merged the genome-wide data of 41 Guizhou Maonan individuals with previously published present-day and ancient East Asian and Southeast Asian populations from Human Origins (HO) and 1240 k datasets included in the Allen Ancient DNA Resource (AADR) and our recently published genome-wide SNP data based on the Illumina platform (Lu et al., 2020; Wang et al., 2020; Wang et al., 2021a; Liu et al., 2021b; Wang et al., 2021b; He et al., 2021c; Chen et al., 2021; Yao et al., 2021).

Principal Component Analysis

We carried out principal component analysis (PCA) *via* the *smartpca* program of the EIGENSOFT v.6.1.4 package (Patterson et al., 2006) based on the merged Human Origin dataset. All default parameters were used with the additional parameter of lsqproject: YES, in which ancient DNA was projected based on the genetic landscape of the modern East Asians from Hmong–Mien, Tai–Kadai, Austronesian, Austroasiatic, and Sino-Tibetan speakers.

ADMIXTURE Analysis

To prune the strong linkage disequilibrium, we first used PLINK tools (Chang et al., 2015) with the additional parameters (--indep-pairwise 200 25 0.4) to obtain the unlinked SNP data among Eurasian modern and ancient populations. Model-based clustering analysis was performed *via* ADMIXTURE (Alexander et al., 2009), and we ran ADMIXTURE with default parameters with the predefined ancestry sources or clusters ranging from K = 2 to 20. We assessed an optimal K value based on the lowest cross-validation error values using 10-fold cross-validation with different random seeds.

Admixture-f_3-Statistics and Outgroup-f_3-Statistics

We used ADMIXTOOLS (Patterson et al., 2012) to compute *f*-statistic values and estimate standard errors by a block jackknife and default parameters. We used the *qp3Pop* program of EIGENSOFT to calculate the outgroup-f_3-statistics in the form f_3(Population 1, Population 2; Mbuti) using the default parameters, and this index was used for evaluating the shared genetic drift between Population 1 and Population 2 since their separation from the outgroup population of Mbuti. Then we also used the qp3pop to perform the admixture-f_3-statistics in the form f_3(Source 1, Source 2; Targeted population) to explore the admixture signals in Maonan samples with different East Asian and Southeast Asian ancestral source candidates. The value with |Z-score|>3 denoted that Source 1 and Source 2 could generate the potential admixture signal for the target population.

f_4 Statistics

We used the *qpDstat* program in ADMIXTOOLS (Patterson et al., 2012) with default parameters to assess whether W or X harbored more ancestry related to population Y in the f_4 (W, X; Y, Outgroup), which can be used to determine the signals and directions of admixture, and the primary source of gene flow to Guizhou Maonan and other modern and ancient reference East Asians.

Pairwise qpWave and qpAdm Estimation

We used *qpWave*/*qpAdm* as implemented in the ADMIXTOOLS (Patterson et al., 2012) package with default parameters and estimated standard errors to detect the minimum number of ancestral populations, and quantitatively estimate corresponding admixture proportions. We used ancient Northeast Asian-related ancestry as the northern sources and Guangxi- or Taiwan-related ancestry as the southern sources to perform the two population

qpAdm model. 1500-year-old BaBanQinCen people are the major ancestral southern sources in our admixture models as it was reported as the direct ancestral sources of modern Tai–Kadai people (Wang et al., 2021c). BaBanQinCen was a meta-population, which comprised two individuals from the Balong site (BalongKD10 and BalongKD07), two individuals from the Banda site (BandaKD15 and BandaKD11), one individual from Qinchang (QinchangKD13 and QinchangKD14), and one individual from Cenxun (CenxunKP05). We used the Mbuti, Ust_Ishim, Kostenki14, Papuan, Australian, Mixe, MA1, Jehai, and Tianyuan as outgroups. We also conducted pairwise qpWave analysis among Tai–Kadai, Hmong–Mien, Sinitic, and ancient Guangxi people to explore their genetic homogeneity. Admixture times were estimated using ALDER with the sources from northern and southern East Asia (Loh et al., 2013).

TreeMix

We ran TreeMix version 1.13 (Pickrell and Pritchard, 2012) to infer the patterns of population splits and admixtures between our target populations and multiple ancestral populations. First, we explored the genetic relationship between Maonan and 15 Chinese populations based on the Illumina array, which was also used in the following haplotype-based analysis. Second, we constructed the TreeMix-based phylogenetic tree among 39 populations to explore the genetic relationship with more reference populations.

Haplotype-Based Fine-Scale Population Structure

We used SHAPEIT v2 (Browning and Browning, 2011) to phase the genome-wide data of Maonan and other Chinese populations in Guizhou and the neighboring regions. Then we conducted the ChromoPainter and FineSTRUCTURE analysis (Hellenthal et al., 2014) to explore the coancestry matrix. We also used R packages implemented in the FineSTRUCTURE to perform the PCA analysis and explore the phylogenetic relationship of studied individuals and populations.

Uniparental Haplogroups

Based on this Illumina array, we genotyped the lineage-informative SNPs (LISNPs) in mitochondrial DNA and Y-chromosome. The haplogroup assignment was used as the in-house manuscripts followed by our recent publications (Lu et al., 2020; Wang et al., 2020; Wang et al., 2021a; Liu et al., 2021b; Wang et al., 2021b; He et al., 2021c; Chen et al., 2021; Yao et al., 2021).

RESULTS

General Structure Inferred From ADMIXTURE and PCA

We generated genome-wide data in approximately 700,000 SNPs for 41 Maonan individuals from Guizhou Province, Southwest China. We first merged our data with modern and ancient published populations from the Human Origins dataset. Then we carried out a principal component analysis (PCA) to understand the general patterns of relatedness between Guizhou Maonan and reference populations (**Figure 1**). We observed three major genetic clines: the northern cline consisting of Mongolic- and Tungusic-speaking populations, the southern cline comprising Hmong–Mien-, Tai–Kadai-, Austroasiatic-, and Austronesian-speaking populations, and the Sino-Tibetans comprising Sinitic- and Tibeto–Burman-speaking populations, which was located at an intermediate position between the northern cline and the southern cline. We projected publicly available data of ancient individuals from China into modern PC plots. Our studied Tai–Kadai-speaking Maonan population all overlapped with modern Tai–Kadai populations. To gain further insight into the genetic architecture of Guizhou Maonan, we further focused on the genetic backgrounds of Tai–Kadai and other southern modern and ancient East Asians (**Supplementary Figure S2**). We observed that our studied group partially overlapped with previously published Austroasiatic- and Hmong–Mien-speaking populations.

We carried out the model-based ADMIXTURE clustering analysis to dissect ancestral components and genetic similarity of our studied group with geographically close ancient and present-day populations. We used cross-validation to identify an "optimal" number of clusters (K = 6) (**Figure 2**). At optimal K = 6, we observed four specific ancestral components in our studied Pingtang Maonan in Guizhou Province: the ancestry maximizing in this cluster is ubiquitous in modern Sinitic-speaking populations (dark green), with the second component maximized in Austroasiatic-speaking populations (dark blue). The remaining ancestry component was maximized in Hmong–Mien- (dark purple) and Austronesian-speaking populations (dark pink). Hmong–Mien-related ancestry component was maximized in historical GaoHuaHua individuals and found at the highest proportions in Hmong. Austronesian-related ancestry component was maximized in ancient and modern Austronesian Taiwanese with a high proportion in earlier Fujian Neolithic individuals (Taiwan_Hanben/Taiwan_Gongguan) and found at the highest proportions in Atayal. We found that our studied Pingtang Maonan is genetically like the other Tai–Kadai-speaking populations, in which both harbored similar patterns of ancestry components.

Phylogenetic Relationship Sharing Alleles and Sharing Haplotypes

The estimated pairwise Fst genetic distances showed the close genetic relationship between Maonan and geographically close Tai–Kadai-speaking populations (Sui and Maonan people in Guangxi), which is consistent with the identified genetic affinity based on the outgroup-f_3 values. Based on the Illumina Array dataset, we further explored the phylogenetic relationship with admixture among 16 Chinese populations belonging to Sinitic, Tungusic, Mongolic, and Tai–Kadai people (**Supplementary Figure S3**). We found that Pingtang Maonan clustered with the Sandu Sui people and following closed with

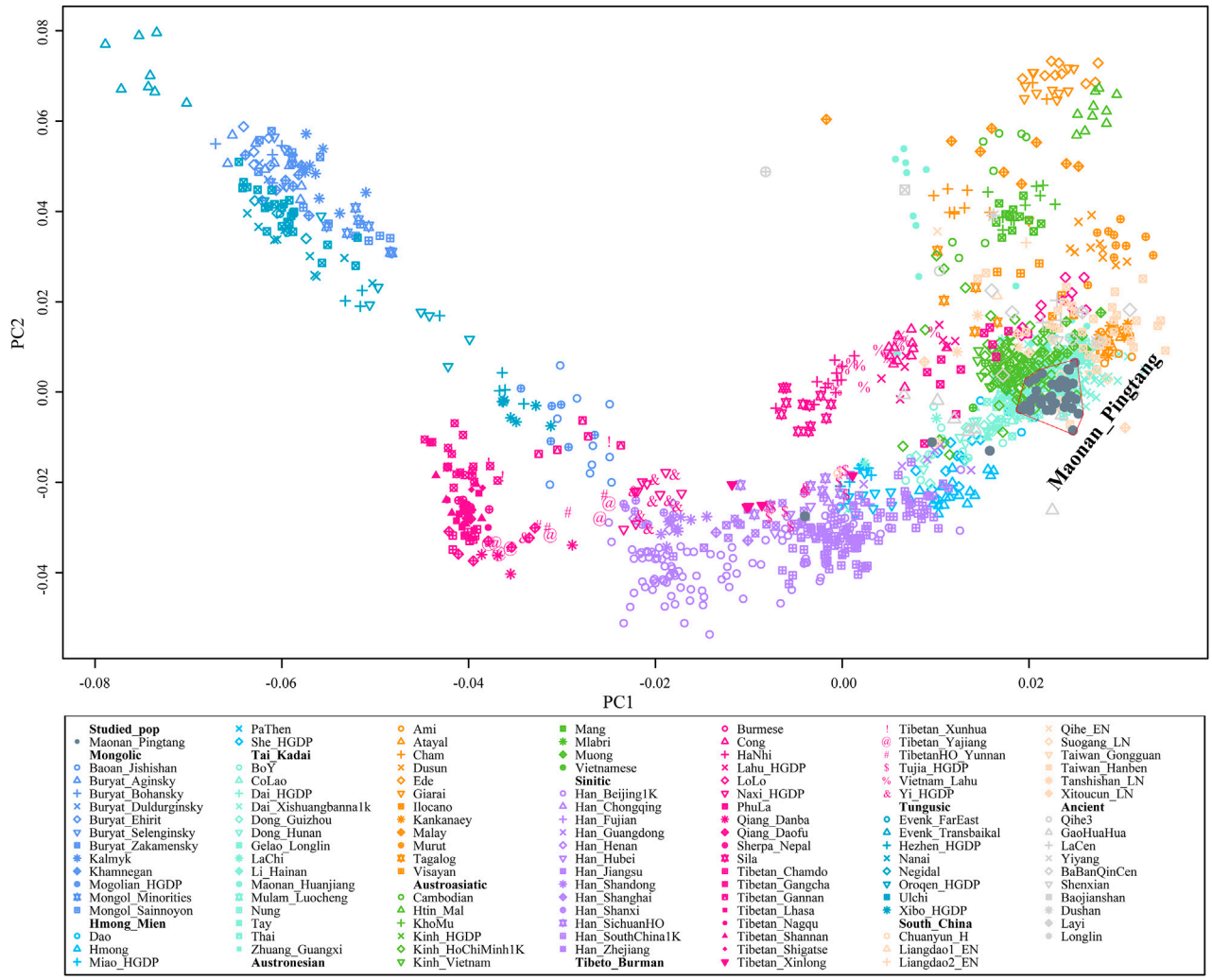

FIGURE 1 | Principal component analysis between Maonan and other modern and ancient references of East Asians.

Manchu and Mongolian people in Guizhou Province. We did not identify gene flow events into Maonan or from Maonan influx to other reference populations. In addition, to explore the genetic relationship with more reference populations in the maximum-likelihood–based TreeMix tree, we also reconstructed a tree among 39 East Asians (**Figure 3**). We found that the Northeast Asians clustered closely with each other, including Tungusic-, Mongolic-, and Sino-Tibetan-speaking populations. Southeast Asians also clustered with each other, including Tai–Kadai and Austronesian people. The genetic affinity between Maonan people and other Tai–Kadai-speaking populations was once again confirmed here, including Dai, Li, Zhuang, and Mulam, and in other Austronesian-speaking populations (Ami and Atayal). This identified genetic phylogenetic relationship further confirmed the close genetic relationship between Tai–Kadai and Austronesian people, which was consistent with recent linguistic similarities.

We further explored the fine-scale population structure between Maonan people and other East Asians based on the patterns of the sharing haplotypes. PCA based on the coancestry matrix showed that Han Chinese clustered together and separated from Pingtang Maonan people and Guizhou Manchu and Mongolic people, which showed that the Maonan people had a different genetic history compared with Han Chinese populations (**Figure 4A–C**). Model-based ADMIXTURE results with three predefined ancestral sources further confirmed the genetic differentiation among Han Chinese, Manchu, and Maonan, who shared their unique ancestry component in **Figure 4D**. We also explored the genetic relationship among these populations based on the individual-level and population-level pairwise coincidence matrix estimated from the coancestry data (**Figure 4E–F**). We found that Maonan people clustered as a separated clade and then clustered with Manchu people. We finally explored the genetic heterozygosity and homogeneity among Tai–Kadai, Hmong–Mien, and Sinitic people using pairwise qpWave analysis. We found the statistically non-significant values (p_rank0>0.05) between Maonan and other Tai–Kadai populations, as well as the geographically

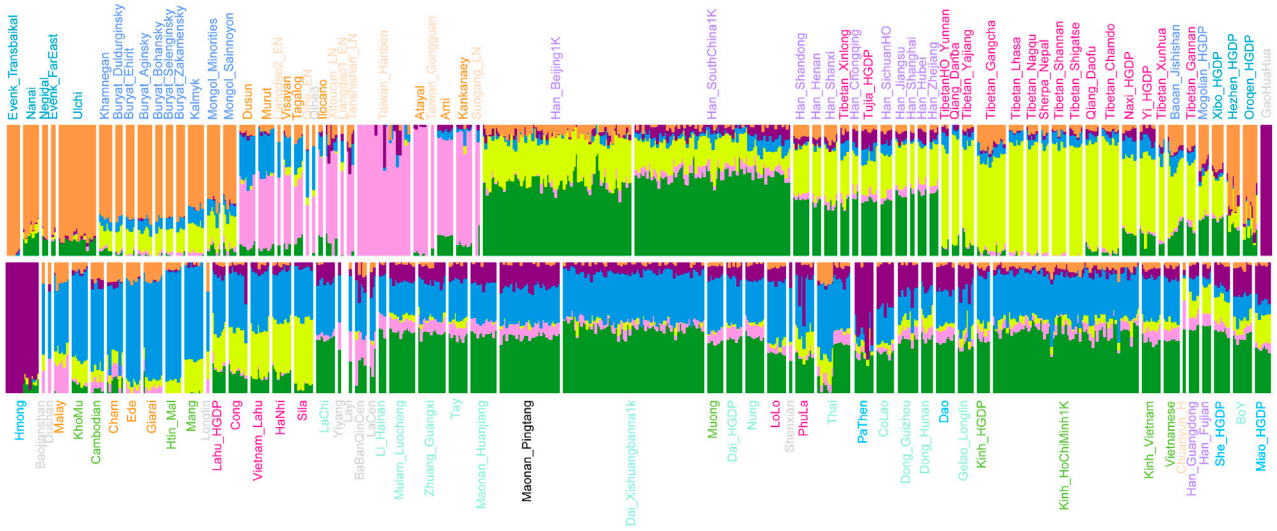

FIGURE 2 | Model-based ADMIXTURE analysis showed the admixture composition of Eurasians.

close Hmong–Mien people. These observed patterns suggested Maonan had a relatively close relationship with other Guizhou populations when a distant outgroup was used here (**Supplemntary Table S1**). Indeed, we found that Maonan, Hmong–Mien, and Tai–Kadai people shared similar patterns of the distribution of the p_rank0 values, and they clustered together and formed one clade in **Figure 5**.

Shared Genetic Drift Inferred From *F*-Statistics

To explore the relationship between the investigated population and reference populations, we measured allele sharing and admixture signals *via* outgroup-f_3 and admixture-f_3-statistics. We performed outgroup-f_3 statistics in the form of f_3(*X, Maonan_Pingtang; Mbuti*), and found that Guizhou Maonan shared more genetic drift with southern Chinese populations, especially for Hmong–Mien- (e.g., Yao, Gejia, and Dongjia) and Tai–Kadai-speaking populations (e.g., Mulam, Zhuang, and Dong) (**Supplementary Table S2A**). Then we used admixture-f_3 statistics of the form f_3(*Source 1, Source two; Maonan_Pingtang)* to model possible sources for Maonan_Pingtang people (**Supplementary Table S2B**). However, we did not observe admixture signals (Z-scores less than -3) significantly in the Maonan_Pingtang when we used different East Asian and Southeast Asian ancestral source candidates.

To further explore the differentiation between the Maonan_Pingtang and other East Asian populations, we performed the f_4(*Reference population 1, Studied population; Reference population 2, Mbuti*). The identified statistically non-significant f_4-values (absolute Z-scores less than 3) in f_4(*Maonan_Huanjiang, Maonan_Pingtang; East Asians, Mbuti*) indicated that Maonan_Pingtang and Maonan_Huanjing have a close genetic relationship (**Supplementary Table S3A**) compared with other reference populations. Focused on other Tai–Kadai-speaking populations, we observed significantly negative f_4-values in f_4(*Dai/Zhuang_Guangxi, Maonan_Pingtang; East Asians, Mbuti*), which suggested that Guizhou indigenous populations and Hmong–Mien-speaking populations shared more alleles with Maonan_Pingtang than other Tai–Kadai-speaking populations (**Supplementary Table S3B**). Among Austronesian-speaking populations, the observed significant negative f_4 statistics in the form f_4(*Atayal/Ami, Maonan_Pingtang; East Asians of Hmong–Mien- and Tai–Kadai-speaking population, Mbuti)* in **Supplementary Table S3C** showed that Pingtang Maonan shared more alleles with Hmong–Mien- and Tai–Kadai-speaking populations compared with modern Austronesian people. Further comparative ancient DNA evidence demonstrated that Maonan_Pingtang also harbored more ancestry related to Hmong–Mien- and Sinitic-speaking populations when we used Neolithic to Iron Age from Fujian and Taiwan as reference population 1 in the form f_4(*ancient southeastern East Asians, Maonan_Pingtang; East Asians, Mbuti*) (**Supplementary Table S3D**).

To directly compare the genetic relationship between Maonan- and Hmong–Mien-speaking populations, we used Hmong–Mien-speaking populations in Vietnam and Guizhou as reference population one in the form f_4(*Hmong–Mien-speaking populations, Maonan_Pingtang; East Asians, Mbuti).* We observed significantly negative f_4-values here, indicating that Maonan_Pingtang shared more alleles with Tai–Kadai- and Austronesian-speaking populations (**Supplementary Table S3E,F**) than with Hmong–Mien-speaking populations. To

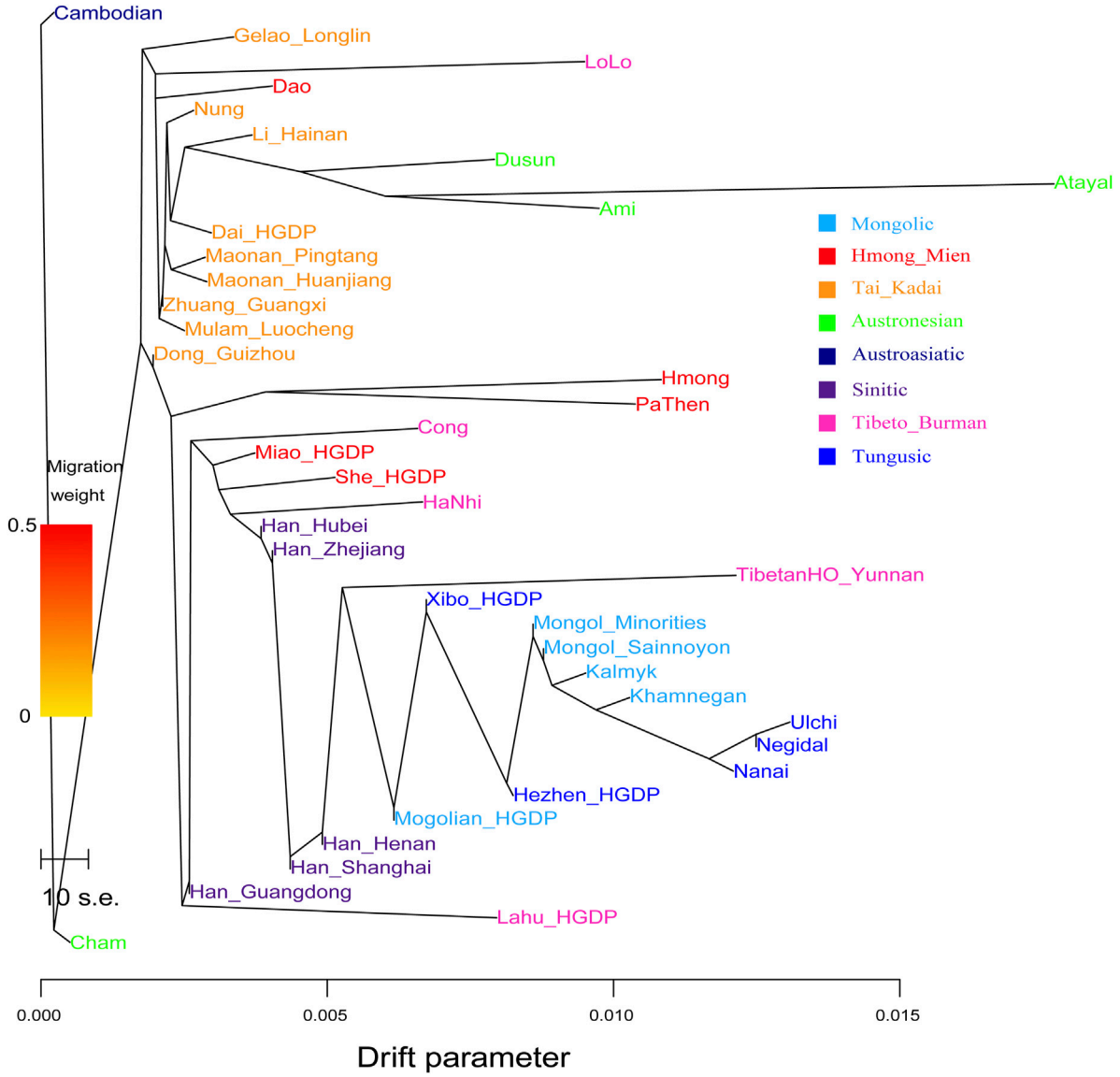

FIGURE 3 | Phylogenetic relationship among East Asians showed the close genetic relationship between Maonan and Chinese Tai–Kadai and Hmong–Mien people.

explore the genetic relationship between the studied population and the ancient populations in Guangxi, we used GaoHuahua, Longlin, Baojianshan, and BaBanQinCen as reference population 1. We observed that Guizhou Maonan shared more alleles with Hmong–Mien- and Sino-Tibetan–speaking populations than GaoHuahua, Longlin, and Baojianshan, as we observed significantly negative f_4-values in the form of f_4(*GaoHuahua/Longlin/Baojianshan, Maonan_Pingtang; East Asians, Mbuti*), which suggests studied population was influenced by gene flow from the north (**Supplementary Tables S3G–I**). We have not observed significant negative f_4-values in f_4(*BaBanQinCen, Maonan_Pingtang; East Asians, Mbuti*), indicating that Maonan_Pingtang shared more alleles with northern populations than BaBanQinCen in Guangxi. We observed BaBanQinCen shared more alleles with other ancient populations than Maonan_Pingtang *via* significant positive f_4-values in the form of f_4(*BaBanQinCen, Maonan_Pingtang; ancient East Asians, Mbuti*) (**Supplementary Table S3J**). We observed significant negative f_4-values in f_4(*Austroasiatic-speaking populations, Maonan_Pingtang; East Asians, Mbuti*), which showed that Sinitic-, Tai–Kadai-, and Hmong–Mien-speaking populations from southern China shared more alleles with Maonan_Pingtang than Austroasiatic-speaking populations (**Supplementary Table S3K**). Significant negative values were

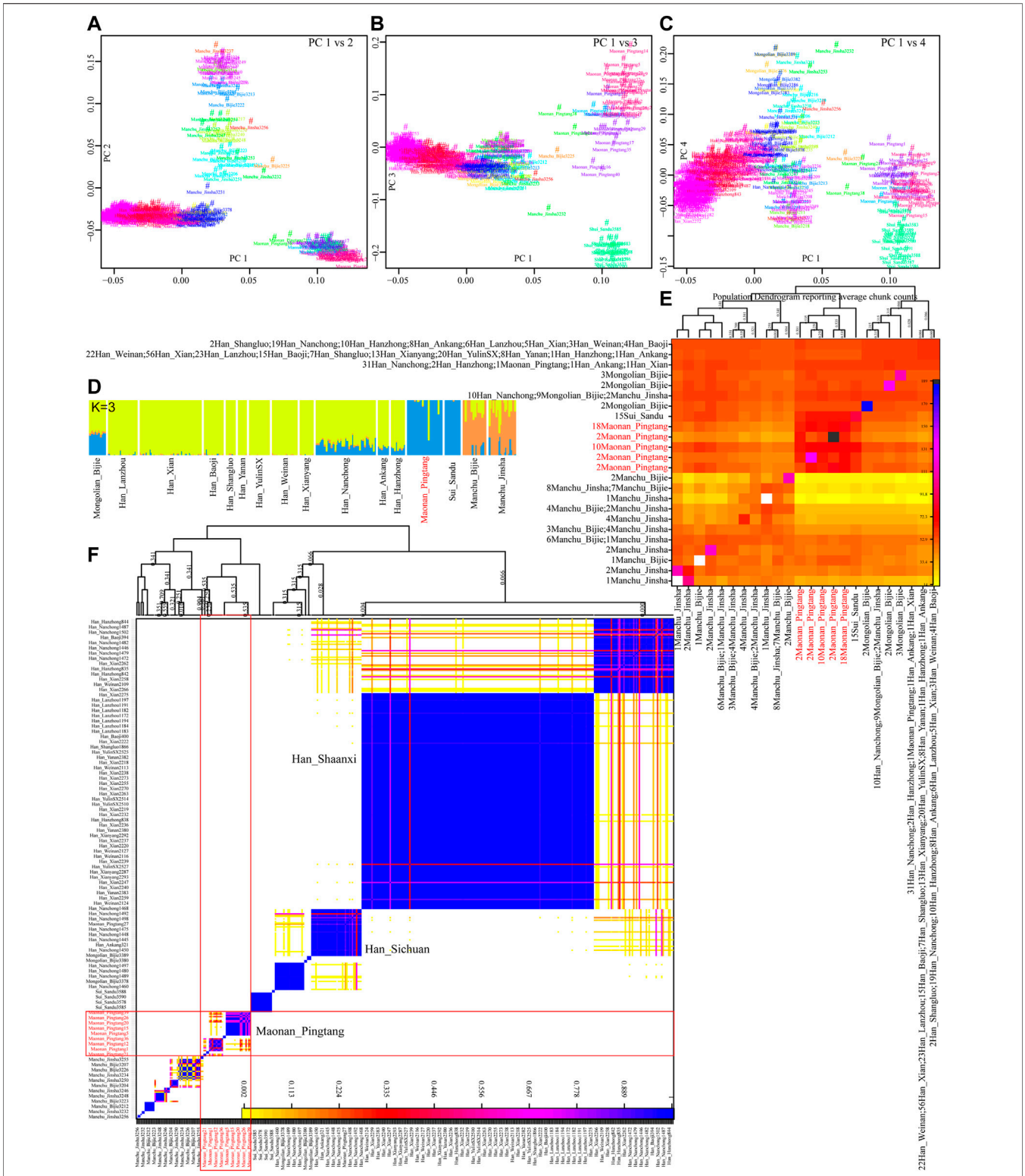

FIGURE 4 | Fine-scale genetic structure among Maonan and other Chinese populations. **(A–C)** Principal component analysis based on the coancestry matrix. **(D)** Model-based ADMIXTURE results showed the three-ancestry component among the used Chinese populations. **(E,F)** Clustering patterns of Chinese populations or individuals based on the pairwise coincidence matrix inferred from the coancestry matrix.

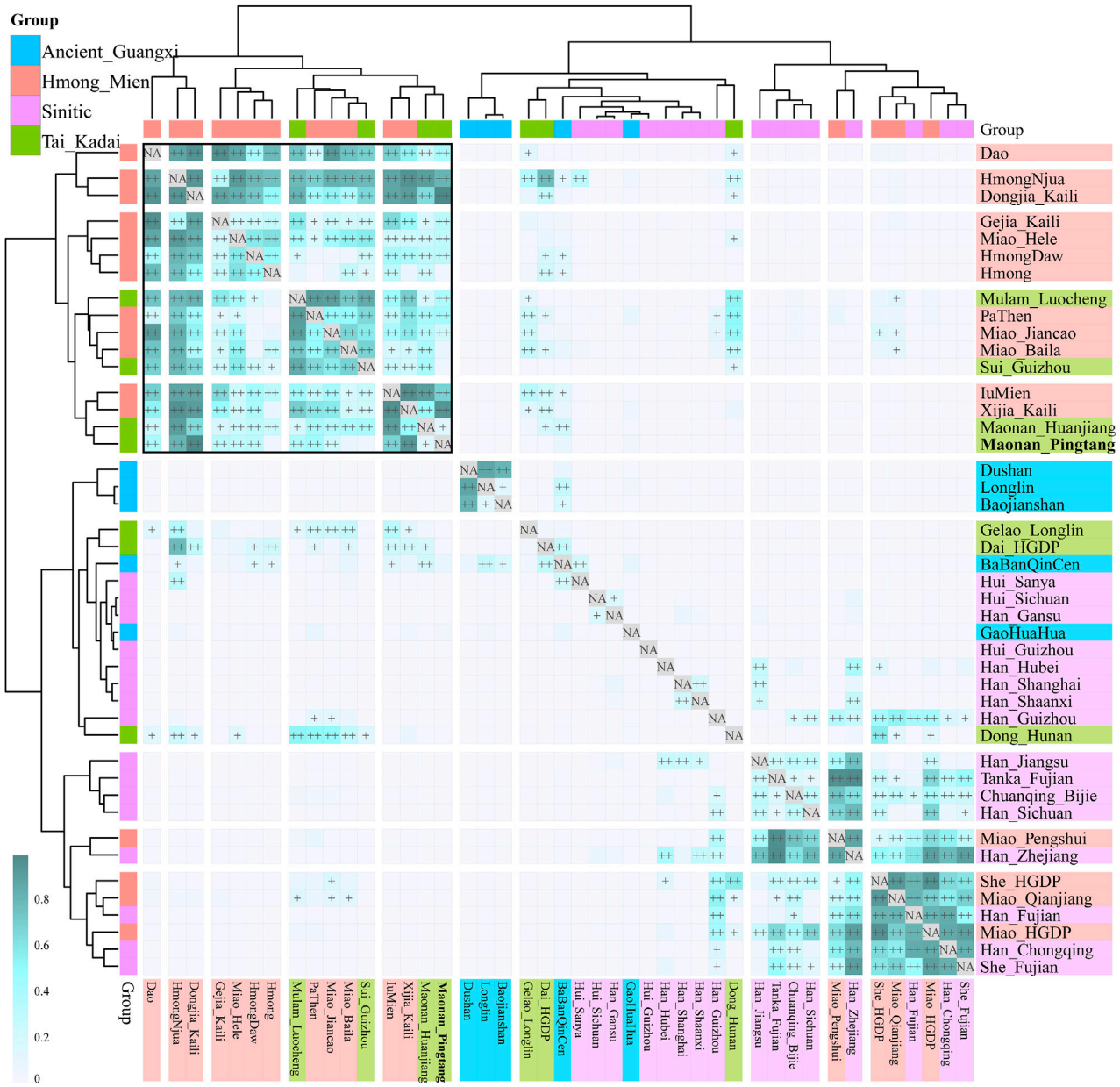

FIGURE 5 | Pairwise qpWave results showed the genetic homogeneity between Maonan and Tai–Kadai and Hmong–Mien people.

observed in the f_4(*southern Tibeto-Burman-speaking population, Maonan_Pingtang; East Asians, Mbuti*), it was shown that Tai–Kadai-, Austronesian-, and Hmong–Mien-speaking populations shared more alleles with studied population than southern Tibeto–Burman-speaking populations (**Supplementary Table S3L**). We found that Maonan_Pingtang harbored more Southeast Asian–related ancestry than ancient Yellow River farmers *via* significant negative f_4-values in f_4(*ancient Yellow River farmers, Maonan_Pingtang; Southeast Asians, Mbuti*) (**Supplementary Table S3M**).

Admixture Modeling and Estimates of Admixture Times

To explore the genomic formation of Guizhou Maonan people, we applied *qpWave*/*qpAdm* methods to model the minimum

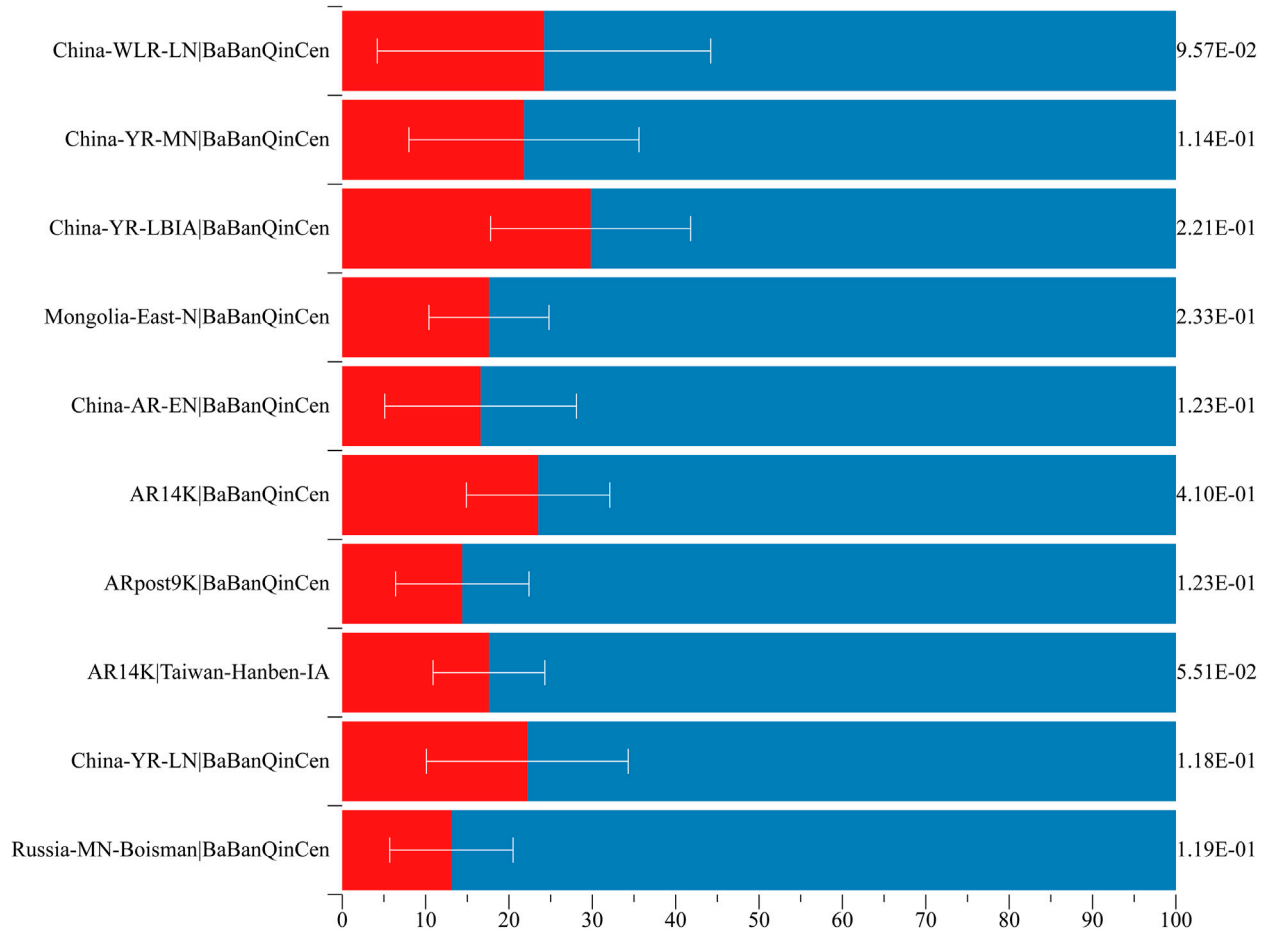

FIGURE 6 | Two-way admixture models showed that both Northeast Asians and Southeast Asians contributed to the formation of Maonan people.

number of ancestry sources and evaluated the corresponding ancestry coefficients. We used ancient northern populations (Russia_MN_Boisman, China_YR_LN, AR14K, Chokhopani, ARpost9K, AR14K, China_AR_EN, Mongolia_East_N, China_YR_LBIA, China_YR_MN, and China_WLR_LN) as the northern ancestral sources and ancient southern populations (Taiwan_Hanben_IA, Guangxi_BaBanQinCen) as the southern ancestral sources to estimate the admixture proportions. Here, BaBanQinCen was the representative ancestral source in Guangxi Province, where many Tai–Kadai people lived. When we used ancient Yellow River farmers as the northern source, Southeast Asian ancestry related to the BaBanQinCen in Maonan people spanned from 70.2 to 86.9% (**Figure 6**, **Supplementary Table S4**). Date estimates with southern and northern East Asians further revealed that these identified north-to-south admixture events occurred in different historical times (**Supplementary Table S5**), which is consistent with genetically attested and historically documented southward population movements (Yang et al., 2020; Wang et al., 2021c; Wang C. C. et al., 2021).

Uniparental Admixture History of Maonan People

Southern China was the originated center of many dominant uniparental lineages of southern Chinese indigenous populations and Southeast Asians (Ke et al., 2001; Wen et al., 2004; He et al., 2020; Sun et al., 2021). Sun et al. provided a higher-resolution phylogeny of O1a-M119 and found this founding lineage widely existed in modern Austronesian-, Tai–Kadai-speaking populations, and southern Han Chinese, suggesting that O1a-M119 lineage was the common lineage of these populations, and the diversified sub-lineages of O1a-M119 were their unique downstream paternal lineages (Sun et al., 2021). Other population genetic studies of Southeast Asians also identified other paternal founding lineages, including O1b1a1a-M95 and O2-M122 (Wen et al., 2004). Similarly, the complex landscape of

maternal lineages in South China was also identified in the previous control-region or whole mitochondrial sequencing projects, including D4, B4, and M7 (Li et al., 2019; Mengge et al., 2020). Here, we also identified that northern and southern paternal and maternal lineages contributed to the uniparental gene pool of Guizhou Maonan people (**Supplementary Table S6**). We observed eight paternal Y-chromosome lineages (O1a1a2a1, O1b1a1a1a1a1a1b, O2a2b1a1a5a2, D1a1a2a2~, O1b1a1a1b, O1b1a1a1a1a1a1a1, O2a1c2, and O1b1a1a1a1a1a1a2-Z24131) with the frequencies ranging from 0.0435 to 0.6957 among 23 males, in which haplogroup of O1b1a1a1a1a1a1a2-Z24131 was observed in 16 individuals. Our results showed that Southeast Asian-dominant paternal lineage of O1b1a1a1a1a1a1a2 was the founding lineage of Tai–Kadai-speaking Maonan people. O1b1a1a* (O-M95*) was previously evidenced to contribute much to the paternal gene pool of populations from South China, Thailand, and Laos (Kutanan et al., 2019). O1b1a1a1a* was also evidenced has experienced significant population expansions in the Neolithic period in Tai–Kadai and Austroasiatic populations. Thus, combined with the identified unique population structure of Maonan people based on the autosomal SNPs and this identified specific lineage of O1b1a1a1a1a1a1a2 with high frequency, Maonan people could be treated as the most representative inland Tai–Kadai-speaking population of southern ancient indigenous populations with the relatively minor genetic influence of southward Han Chinese expansion. Our results also suggested that the genetically documented marriage pattern was consistent with that of the culturally documented customers of patrilocality, which is also evidenced *via* the marriage pattern among Tai–Kadai people in Southeast Asians (Kutanan et al., 2019). In addition, we assigned all 41 maternally inherited mtDNA lineages into 25 terminal lineages with frequencies ranging from 0.0244 to 0.0732 (B4a1e, B5a1c1, F1a, F1c, F3b, and B4), which were observed frequently in both northern and southern East Asians.

DISCUSSION

Guizhou Province is rich in ethnolinguistic and cultural diversity (Wang et al., 2021a). Previous genetic studies have investigated the general landscape of genetic variations of Guizhou populations based on the autosomal, X/Y-chromosomal short tandem repeats (STRs), and ancestry-informative SNPs (Chen et al., 2018a; Chen et al., 2018b; Chen et al., 2018c; Chen et al., 2018d; Chen et al., 2019a; Chen et al., 2019b; He et al., 2021b). Archaeologically attested Daxi, Qujialing, and Shijiahe people were occupied in what is now Guizhou Province, and the present Guizhou region was occupied by Sino-Tibetan–speaking (Han, Yi, and others), Tai–Kadai-speaking (Gelao, Bouyei, Dong et al.), and Hmong–Mien-speaking populations (Miao, She, and Yao). Recent ancient DNA studies from the Yellow River Basin in Northeast Asia, Fujian, and Guangxi provinces from Southeast Asia also found that the bi-directional north-to-south population movements have shaped the genetic landscape of East Asians (Ning et al., 2020; Yang et al., 2020; Wang et al., 2021c; Mao et al., 2021; Wang C. C. et al., 2021). Demographic modeling of central Chinese populations, including Han and Tujia, also showed the genetic influences from both the northern millet farmers and southern rice farmers (He et al., 2021a). The Tai–Kadai language family and their corresponding people widely existed in Guizhou and surrounding regions; however, the fine-scale genetic structure of this linguistically specific population still needs to be further explored.

We genotyped the genome-wide SNP data from 41 Tai–Kadai-speaking Maonan people and explored their genetic origin, admixture history, and phylogenetic relationship with surrounding populations. Descriptive analysis based on the PCA and ADMIXTURE analyses showed that the Guizhou Maonan people and Guangxi Maonan people had the closest genetic relationship and shared the most genetic affinity, suggesting their common origin and admixture history. The genetic affinity within the population among Tai–Kadai-speaking populations was also evidenced by the Dong and Bouyei based on the Affymetrix-based array data (Wang et al., 2021a). In addition, population genetic analysis based on the forensic genetic markers (STR and Indels) also revealed the genetic affinity between geographically different Tai–Kadai-speaking populations rather than populations from other language families (Chen et al., 2018c; He et al., 2019a; He et al., 2019b). Interestingly, we also identified the genetic affinity between Maonan and Hmong–Mien-speaking populations in Guizhou Province among the non-Tai–Kadai-speaking populations based on the observed non-statistically significant f_4-statistics in the form f_4(*Maonan, Hmong–Mien-speaking Gejia/Dongjia/Xijia; other reference Asians populations, Mbuti*) and statistically negative f_4-values in f_4 (*Asian reference populations, Maonan; Hmong–Mien, Mbuti).* Our results were consistent with previous reports based on the forensic markers. He et al. explored the genetic diversity and forensic features of Guizhou Tai–Kadai-speaking Gelao people and identified the population interplay between Gelao and neighboring Hmong–Mien-speaking populations (He et al., 2019b). Our identified extensive genetic admixture between Hmong–Mien and Tai–Kadai people suggested that there was no clear genetic barrier between geographically close but linguistically different ethnic groups, which suggested that they have experienced extensive population interaction although initially they were of independent origin.

Linguistic interaction between Tai–Kadai and Sino-Tibetan languages was widely documented (Edmondson, 1988; Edmondson, 1997; Lu, 2008). Population interaction between Maonan- and Sino-Tibetan-speaking populations was also identified in our genetic study, which is consistent with other recently published population genetic investigations. He et al. recently explored the fine-scale genetic structure of four Guizhou Han populations and found their extensive admixture with Guizhou indigenes (Wang et al., 2021b). In addition, Wang et al. studied the genetic admixture of Guizhou culturally unique Hui people and found their connection between indigenous Han people (Wang et al., 2020). Other genome-wide SNP-based genetic analyses focused on Guizhou officially unrecognized Chuanqing people also found their genetic affinity with geographically close Han populations (Lu et al., 2020). Additionally, we found a close genetic relationship between Maonan and Guizhou Hans in the f_3/f_4-statistics and the TreeMix-based maximum-likelihood–based phylogenetic tree, which suggested their recent admixture process. The shared ancestry between Guizhou Maonan and

Han people in ADMIXTURE and their qualitative indices was consistent with the shared cultural background between present-day Guizhou Han and Maonan people.

More and more ancient genomes in the surrounding regions of Guizhou Province were reported recently, especially important ancestry sources of the possible ancestor of Tai–Kadai-speaking populations (BaBanQinCen) and a possible ancestor of Hmong–Mien-speaking populations (GaoHuaHua). BaBanQinCen was one meta-population from four archeological sites that lived in Guangxi Province 2000 years ago, and GaoHuaHua was also a meta-population from three archaeological sites that lived in Guangxi around 1,500 years ago (Wang et al., 2021c). Here, we found that Maonan people shared the most genetic affinity with ancient Guangxi historic BaBanQinCen, which was recently genetically attested as the direct ancestor of Guizhou Tai–Kadai people (Wang et al., 2021c). Outgroup-f_3-statistics and shared ancestry inferred from f_4-statistics further confirmed the closest genetic connection between Maonan and BaBanQinCen compared with other Guizhou historic people (GaoHuaHua, Layi et al.) and prehistoric Longlin, Dushan, and Baojianshan people. Among all reported ancient Northeast Asians, including the inland and coastal Neolithic Northeast Asians from Shandong, Henan, Shaanxi, Gansu, Inner Mongolia in the Yellow River Basin, and Neolithic Siberian, we found a close genetic connection between Maonan and Bronze Age to Iron Age people from Henan Province, which suggested Maonan people might have obtained gene influxes from them. Indeed, we obtained statistically negative f_4-values in the f_4(*Guangxi historic people, Maonan; Yellow River millet farmers, Mbuti),* which directly evidenced that compared with the genetically attested ancestors of Tai–Kadai-speaking populations from the Guangxi region, Maonan people shared additional genetic materials from Northeast Asians. This identified admixture process was further confirmed *via* the qpAdm-based two-way admixture models with one source from Guangxi historic people and the other sources from Northeast Asians. We also found that two-way admixture models of Hanben-Northeast Asians can also well-fit the genetic composition of studied Maonan, which suggests the genetic influence from the southeastern coastal Fujian ancient people in the gene pool of inland Tai–Kadai-speaking Maonan people. As we know, Tai–Kadai-speaking populations were widely distributed in South China, including Guizhou, Guangxi, and Hainan. Previous genetic analysis from Hainan Province also reported a relatively isolated genetic structure of the Hainan Tai–Kadai-speaking Hlai people (He et al., 2020). Thus, dense sampling of Tai–Kadai-speaking populations and obtaining their whole-genome sequencing data would help to characterize the complete genetic admixture landscape of Chinese Tai–Kadai-speaking populations.

CONCLUSION

We reported the first batch of genome-wide SNP data of Tai–Kadai-speaking Maonan people from Guizhou Province and comprehensively explored genetic structure, origin, and admixture processes based on the descriptive analyses (PCA and ADMIXTURE) and qualitative measures (*f*-statistics, *qpAdm*). Results from PCA and ADMIXTURE showed a close genetic relationship between Maonan and other geographically different Tai–Kadai-speaking populations, especially for the closest relationship between Guizhou Maonan and Guangxi Maonan. No-admixture signatures were identified *via* admixture-f_3 statistics showed the unique genetic structure of Maonan people compared with geographically close Han people. Further analysis based on the outgroup-f_3 statistics and f_4-based analysis showed a close relationship between Maonan and Guizhou Sino-Tibetan and Hmong-speaking populations, as well as a close connection between Guangxi historic people and Guizhou Tai–Kadai-speaking populations, suggesting their admixture history with the sources from surrounding regions. The well-fitted two-way admixture models with ancient northern and southern East Asians demonstrated that Tai–Kadai-speaking populations derived primary ancestry related to 1500-year-old Guangxi BaBanQinCen people and additional genes from Northeast Asia.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://zenodo.org/record/5701604, 10.5281/zenodo.5701604.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Medical Ethics Committee of Guizhou Medical University and Xiamen University (Approval Number: XDYX2019009). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

C-CW and JH designed this study. JC, GH, and C-CW wrote and revised the manuscript. ZR, QW, YL, HZ, MY, HZ, JJ, and JH collected the samples. QW, ZR, HZ, JJ, YL, MY, JC, and JH conducted the experiment. GH, JZ, JG, JC, KZ, XY, RW, HM, LT, YL, QS, WY, and C-CW analyzed the data. All authors reviewed the manuscript.

FUNDING

This work was funded by the Guizhou Scientific Support Project, Qian Science Support (2021) General 448; Guizhou Province Education Department, Characteristic Region Project, Qian Education KY No. (2021) 065; Guizhou "Hundred" High-level Innovative Talent Project, Qian Science Platform Talents (2020) 6,012; Guizhou Scientific Support Project, Qian Science Support

(2020) 4Y057; Guizhou Science Project, Qian Science Foundation (2020) 1Y353; Guizhou Scientific Support Project, Qian Science Support (2019) 2,825; the Guizhou Scientific Cultivation Project, Qian Science Platform Talent (2018) 5779-X; and the Guizhou Engineering Technology Research Center Project, Qian High-Tech of Development and Reform Commission NO. (2016) 1,345; the National Natural Science Foundation of China (NSFC 31801040), 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, 0310/X2106027); Nanqiang Outstanding Young Talents Program of Xiamen University (X2123302); the Major Project of National Social Science Foundation of China granted to C-CW (21 & ZD285) and Xiaohua Deng (20 & ZD248); the European Research Council (ERC) grant to Dan Xu (ERC-2019-ADG-883700-TRAM). We thank S. Fang and Z. Xu from Information and Network Center of Xiamen University for their help with high-performance computing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.815285/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

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. (2011). A Fast, Powerful Method for Detecting Identity by Descent. *Am. J. Hum. Genet.* 88, 173–182. doi:10.1016/j.ajhg.2011.01.010

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. (2021). 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, P., Han, Y., He, G., Luo, H., Gao, T., Song, F., et al. (2018a). Genetic Diversity and Phylogenetic Study of the Chinese Gelao Ethnic Minority via 23 Y-STR Loci. *Int. J. Leg. Med* 132, 1093–1096. doi:10.1007/s00414-017-1743-y

Chen, P., He, G., Xing, H., Gao, H., Wang, M., Zhao, M., et al. (2019a). Forensic Characteristics and Phylogenetic Analysis of 23 Y-STR Loci in the Miao Population from Guizhou Province, Southwest China. *Ann. Hum. Biol.* 46, 84–87. doi:10.1080/03014460.2019.1583374

Chen, P., He, G., Zou, X., Wang, M., Jia, F., Bai, H., et al. (2018b). Forensic Characterization and Genetic Polymorphisms of 19 X-Chromosomal STRs in 1344 Han Chinese Individuals and Comprehensive Population Relationship Analyses Among 20 Chinese Groups. *PLoS One* 13, e0204286. doi:10.1371/journal.pone.0204286

Chen, P., He, G., Zou, X., Wang, M., Luo, H., Yu, L., et al. (2018c). Genetic Structure and Polymorphisms of Gelao Ethnicity Residing in Southwest china Revealed by X-Chromosomal Genetic Markers. *Sci. Rep.* 8, 14585. doi:10.1038/s41598-018-32945-7

Chen, P., He, G., Zou, X., Zhang, X., Li, J., Wang, Z., et al. (2018d). Genetic Diversities and Phylogenetic Analyses of Three Chinese Main Ethnic Groups in Southwest China: A Y-Chromosomal STR Study. *Sci. Rep.* 8, 15339. doi:10.1038/s41598-018-33751-x

Chen, P., Luo, L., Gao, H., Wu, J., Wang, Y., He, G., et al. (2019b). Forensic Performance of 30 InDels Included in the Investigator DIPplex System in Miao Population and Comprehensive Genetic Relationship in China. *Int. J. Leg. Med* 133, 1389–1392. doi:10.1007/s00414-019-02057-6

Diamond, J., and Bellwood, P. (2003). Farmers and Their Languages: the First Expansions. *Science* 300, 597–603. doi:10.1126/science.1078208

Edmondson, J. A. (1988). *Comparative Kadai : Linguistic Studies beyond Tai*. Dallas: Summer Institute of Linguistics , University of Texas at Arlington.

Edmondson, J. A. (1997). *Comparative Kadai : The Tai branch*. Dallas: Summer Institute of Linguistics , University of Texas at Arlington.

He, G., 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., Liu, J., Wang, M., Zou, X., Ming, T., Zhu, S., et al. (2021b). Massively Parallel Sequencing of 165 Ancestry-Informative SNPs and Forensic Biogeographical Ancestry Inference in Three Southern Chinese Sinitic/Tai-Kadai Populations. *Forensic Sci. Int. Genet.* 52, 102475. doi:10.1016/j.fsigen.2021.102475

He, G., Ren, Z., Guo, J., Zhang, F., Zou, X., Zhang, H., et al. (2019a). Population Genetics, Diversity and Forensic Characteristics of Tai-Kadai-Speaking Bouyei Revealed by Insertion/deletions Markers. *Mol. Genet. Genomics* 294, 1343–1357. doi:10.1007/s00438-019-01584-6

He, G., Wang, M., Zou, X., Chen, P., Wang, Z., Liu, Y., et al. (2021c). Peopling History of the Tibetan Plateau and Multiple Waves of Admixture of Tibetans Inferred from Both Ancient and Modern Genome-wide Data. *Front. Genet.* 12, 725243. doi:10.3389/fgene.2021.725243

He, G., Wang, Z., Guo, J., Wang, M., Zou, X., Tang, R., et al. (2020). Inferring the Population History of Tai-Kadai-Speaking People and Southernmost Han Chinese on Hainan Island by Genome-wide Array Genotyping. *Eur. J. Hum. Genet.* 28, 1111–1123. doi:10.1038/s41431-020-0599-7

He, G., Wang, Z., Zou, X., Wang, M., Liu, J., Wang, S., et al. (2019b). 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

Huang, J. (2016). An Analysis of the Origin of Rice-growing Culture in China. *Local Cult. Res.* 4, 40–57. (In Chinese).

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. *bioRxiv*. doi:10.1101/2020.11.08.373225

Ke, Y., Su, B., Song, X., Lu, D., Chen, L., Li, H., et al. (2001). African Origin of Modern Humans in East Asia: a Tale of 12,000 Y Chromosomes. *Science* 292, 1151–1153. doi:10.1126/science.1060011

Kutanan, W., Kampuansai, J., Srikummool, M., Brunelli, A., Ghirotto, S., Arias, L., et al. (2019). Contrasting Paternal and Maternal Genetic Histories of Thai and Lao Populations. *Mol. Biol. Evol.* 36, 1490–1506. doi:10.1093/molbev/msz083

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

Li, Y.-C., Ye, W.-J., Jiang, C.-G., Zeng, Z., Tian, J.-Y., Yang, L.-Q., et al. (2019). River Valleys Shaped the Maternal Genetic Landscape of Han Chinese. *Mol. Biol. Evol.* 36, 1643–1652. doi:10.1093/molbev/msz072

Liang, M., and Zhang, J. (1996). *An Introduction to the Tai-Kadai Language Family (Chinese)* Beijing, China: Social Sciences Publishing House.

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, Y., Xie, J., Wang, M., Liu, C., Zhu, J., Zou, X., et al. (2021a). 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

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

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

Lu, J., Zhang, H., Ren, Z., Wang, Q., Liu, Y., Li, Y., et al. (2020). Genome-wide Analysis of Unrecognised Ethnic Group Chuanqing People Revealing a Close Affinity with Southern Han Chinese. *Ann. Hum. Biol.* 47, 465–471. doi:10.1080/03014460.2020.1782470

Lu, T. Q. (2008). *A Grammar of Maonan*. Boca Raton, United States: Universal Publishers.

Ma, T., Rolett, B. V., Zheng, Z., and Zong, Y. (2020). Holocene Coastal Evolution Preceded the Expansion of Paddy Field rice Farming. *Proc. Natl. Acad. Sci. USA* 117, 24138–24143. doi:10.1073/pnas.1919217117

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

Mengge, W., Guanglin, H., Yongdong, S., Shouyu, W., Xing, Z., Jing, L., et al. (2020). Massively Parallel Sequencing of Mitogenome Sequences Reveals the Forensic Features and Maternal Diversity of Tai-Kadai-Speaking Hlai Islanders. *Forensic Sci. Int. Genet.* 47, 102303. doi:10.1016/j.fsigen.2020.102303

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

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

Stevens, C. J., and Fuller, D. Q. (2017). The Spread of Agriculture in Eastern Asia. *Lang. Dyn. Change* 7, 152–186. doi:10.1163/22105832-00702001

Sun, J., Li, Y. X., Ma, P. C., Yan, S., Cheng, H. Z., Fan, Z. Q., et al. (2021). Shared Paternal Ancestry of Han, Tai-Kadai -speaking, and Austronesian-speaking Populations as Revealed by the High Resolution Phylogeny of O1a-M119 and Distribution of its Sub-lineages within China. *Am. J. Phys. Anthropol.* 174, 686–700. doi:10.1002/ajpa.24240

Wang, C. C., Yeh, H.-Y., Popov, A. N., Zhang, H.-Q., Matsumura, H., Sirak, K., et al. (2021). 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. (2021a). Reconstructing The Genetic Admixture History of Tai-Kadai and Sinitic People: Insights From Genome-Wide SNP Data From South China. *J. Syst. Evol.*. doi:10.1111/jse.12825

Wang, M., Yuan, D., Zou, X., Wang, Z., Yeh, H.-Y., Liu, J., et al. (2021b). 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, T., Wang, W., Xie, G., Li, Z., Fan, X., Yang, Q., et al. (2021c). 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

Wang, W. G. (2004). A Comprehensive Study of the History of Baiyue Ethnic Group. *J. yunnan Univ.* (In Chinese).

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

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

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, J. (2016). An Exploration of the Bronze Drums Culture and the Origin of the Belief System of Southern China. *Arts Explorat* 30 (4), 68–80. (In Chinese). doi:10.13574/j.cnki.artsexp.2016.04.008

Zhang, X., Li, C., Zhou, Y., Huang, J., Yu, T., Liu, X., et al. (2020). A Matrilineal Genetic Perspective of Hanging Coffin Custom in Southern China and Northern Thailand. *iScience* 23, 101032. doi:10.1016/j.isci.2020.101032

Zhao, Z. (2011). New Archaeobotanic Data for the Study of the Origins of Agriculture in China. *Curr. Anthropol.* 52, S295–S306. doi:10.1086/659308

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 17 February 2022
doi: 10.3389/fgene.2021.801275

Development and Performance Evaluation of a Novel Ancestry Informative DIP Panel for Continental Origin Inference

Yongsong Zhou[1,2], *Xiaoye Jin*[3], *Buling Wu*[1]* and *Bofeng Zhu*[1,2,4,5]*

[1]Shenzhen Stomatology Hospital (Pingshan), Southern Medical University, Shenzhen, China, [2]Guangzhou Key Laboratory of Forensic Multi-Omics for Precision Identification, School of Forensic Medicine, Southern Medical University, Guangzhou, China, [3]School of Forensic Medicine, Guizhou Medical University, Guiyang, China, [4]Key Laboratory of Shaanxi Province for Craniofacial Precision Medicine Research, College of Stomatology, Xi'an Jiaotong University, Xi'an, China, [5]Clinical Research Center of Shaanxi Province for Dental and Maxillofacial Diseases, College of Stomatology, Xi'an Jiaotong University, Xi'an, China

Ancestry informative markers (AIMs) are useful to infer individual biogeographical ancestry and to estimate admixture proportions of admixed populations or individuals. Although a growing number of AIM panels for forensic ancestry origin analyses were developed, they may not efficiently infer the ancestry origins of most populations in China. In this study, a set of 52 ancestry informative deletion/insertion polymorphisms (AIDIPs) were selected with the aim of effectively differentiate continental and partial Chinese populations. All of the selected markers were successfully incorporated into a single multiplex PCR panel, which could be conveniently and efficiently detected on capillary electrophoresis platforms. Genetic distributions of the same 50 AIDIPs in different continental populations revealed that most loci showed high genetic differentiations between East Asian populations and other continental populations. Population genetic analyses of different continental populations indicated that these 50 AIDIPs could clearly discriminate East Asian, European, and African populations. In addition, the 52 AIDIPs also exhibited relatively high cumulative discrimination power in the Eastern Han population, which could be used as a supplementary tool for forensic investigation. Furthermore, the Eastern Han population showed close genetic relationships with East Asian populations and high ancestral components from East Asian populations. In the future, we need to investigate genetic distributions of these 52 AIDIPs in Chinese Han populations in different regions and other ethnic groups, and further evaluate the power of these loci to differentiate different Chinese populations.

Keywords: ancestry informative marker, deletion/insertion polymorphism, AIDIP, forensic ancestry analysis, Eastern Han

OPEN ACCESS

Edited by:
Chuan-Chao Wang,
Xiamen University, China

Reviewed by:
Zheng Wang,
Sichuan University, China
Pengyu Chen,
Affiliated Hospital of Zunyi Medical College, China

***Correspondence:**
Buling Wu
bulingwu@smu.edu.cn
Bofeng Zhu
zhubofeng7372@126.com

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

Received: *25 October 2021*
Accepted: *01 December 2021*
Published: *17 February 2022*

Citation:
Zhou Y, Jin X, Wu B and Zhu B (2022) Development and Performance Evaluation of a Novel Ancestry Informative DIP Panel for Continental Origin Inference. Front. Genet. 12:801275. doi: 10.3389/fgene.2021.801275

INTRODUCTION

Ancestry informative markers (AIMs) refer to genetic variations that exhibit high allelic frequency divergences between different ancestral populations (Phillips et al., 2007). AIMs are useful to infer individual biogeographical ancestry and to estimate admixture proportions of admixed populations or individuals. In the last decade, as a new supplementary test, forensic ancestry information analysis

provides much valuable information for forensic investigative applications and other forensic fields (Phillips, 2015; Phillips and de la Puente, 2021). Most recently, a growing number of AIM panels to estimate ancestry origin of continental and sub-continental populations (Santos et al., 2016a; Wei et al., 2016; Carvalho Gontijo et al., 2020; Xavier et al., 2020) or to distinguish population structure of Asian or Chinese populations (Sun et al., 2016; Jin et al., 2019; Qu et al., 2019) were developed by forensic researchers from abroad and in China, respectively. However, the capacity of these panels to effectively infer the ancestry origins of other populations in China may not be competent enough. Furthermore, large-scale and representative population genetic data are the key element of forensic assay development and application. Unfortunately, AIM reference population data in most Chinese populations are still undeveloped to date, which limit population-specific marker selections to some extent. Accordingly, we need to investigate genetic distributions of more AIMs in Chinese populations. These data can not only enrich the genetic information resources of Chinese population, but also facilitate the screening of population specific molecular markers.

Deletion/insertion polymorphisms (DIPs) are one type of genetic variations that arise from random deletion or insertion of DNA fragments (Weber et al., 2002). This kind of polymorphism exhibits unique characteristics as AIMs: (i) wide distributions in the human genome; (ii) with low mutation rates; (iii) the frequencies of alleles varies greatly between populations; and (iv) can be easily detected by multiplex PCR and capillary electrophoresis platform (Santos et al., 2010; Li et al., 2012; LaRue et al., 2014). In recent years, DIPs received a large amount of attention from forensic geneticists. A set of DIP panels for various forensic purposes have been constructed. For example, Chen et al. developed a multiplex panel of autosomal DIPs for forensic identity testing (Chen et al., 2019); Lan et al. presented a multiplex system of 39 ancestry informative DIPs (AIDIPs) for forensic ancestry origins of three different continental populations (Lan et al., 2019); Chen et al. also constructed a novel multiplex system that could detect 38 X-chromosome DIPs to assist in individual identification and paternity testing (Chen et al., 2021). Collectively, the DIPs showed great application values in forensic research.

In this study, we firstly selected 52 AIDIPs for ancestry origin predictions of different continental populations based on the 1,000 Genome Project (Genomes Project et al., 2015) and previous studies (Mills et al., 2006; Pereira et al., 2009; Santos et al., 2010; Pereira et al., 2012b). Secondly, we evaluated the efficiencies of these AIDIPs for dissecting continental population structure. At the same time, a multiplex panel of these 52 AIDIPs was developed on the basis of capillary electrophoresis platform. Next, genetic distributions and forensic statistical parameters of these 52 AIDIPs in Eastern Han population were assessed. Finally, ancestral components of Eastern Han population were explored in comparison with continental populations.

MATERIALS AND METHODS

AIDIPs Selection and Development of the Multiplex Panel

We aim to construct a multiplex PCR assay of 52 AIDIPs based on the capillary electrophoresis platform for forensic individual biogeographic ancestry inference and population genetic structure and background analyses. A batch of 52 AIDIPs located on autosomal chromosome were selected; they showed high allele frequency divergences among European, East Asian, and African populations, which were confirmed by previous studies (Mills et al., 2006; Pereira et al., 2009; Santos et al., 2010; Pereira et al., 2012b). AIDIPs selection criteria were consistent with Lan et al. (2019). We screened 52 biallelic DIP genetic markers that performed the following requirements: (i) all DIP markers were selected from the autosomes; (ii) variable size of deletion/insertion fragments ranged from 2 to 20 bp; (iii) allele frequency differentials ≥0.2 between at least two continental populations; and (iv) no departures from Hardy–Weinberg equilibrium (HWE) in any continental population. The detailed genomic information and reference sequences of these selected AIDIP loci were obtained from dbSNP (http://www.ncbi.nlm.nih.gov/SNP/). The primer design and multiplex assay construction of 52 AIDIP loci proceeded based on the workflows described by Pereira et al. (2012a; 2012b). Primers were designed by the Primer Premier 5. 0 software according to the following two main principles: the T_m value was close to 65°C and the amplicon sizes varied from 60 to 250 bp. Potential primer dimers and hairpin structures were evaluated by AutoDimerv1 software. Subsequently, all markers were assigned and labeled by four different fluorescent dyes (FAM, HEX, TAMRA, and ROX), respectively, and all of the primers were synthesized by Sangon Biotech (Sangon Biotech Co., Ltd., Shanghai, China).

Multiplex Amplification and AIDIP Genotyping

The PCR reaction of the 52 AIDIPs was finally optimized to amplify in a single tube with the 25-μl reaction volume. The reaction system composed of 5 μl of primers, 10 μl of reaction mix (AGCU Biotech Co., Ltd., Wuxi, China), 1 μl of template DNA, 1 μl of polymerase (5 U/μl, Takara Biomedical Technology Co., Ltd., Beijing, China), and 8 μl of sdH_2O. The PCR cycling conditions were as follows: 95°C for 5 min; 30 cycles of 94°C for 15 s, 60°C for 50 s, and 62°C for 55 s; and a final extension at 70°C for 20 min.

For capillary electrophoresis, 1 μl of amplification products were added to 12.5 μl of loading mixtures, which consisted of 12 μl of deionized Hi-Di® formamide (Thermo Fisher Scientific, Waltham, MA United States) and 0.5 μl of AGCU SIZ-500 internal size standard (AGCU Biotech). Detection and separation for 52 AIDIPs were performed on 3500xL Genetic Analyzers (Thermo Fisher Scientific) under default injection conditions. The raw data were genotyped with GeneMapper® ID-X v1.5 software (Thermo Fisher Scientific).

Ethics Statement, Population Sample Collection, and Genomic DNA Extraction

Buccal samples were obtained from the volunteers with written informed consents for the above-mentioned research purposes, approved by the Ethics Committee of Xi'an Jiaotong University, China (No. XJTULAC 2013). Buccal samples stored on FTA™ cards (GE Healthcare, Buckinghamshire, United Kingdom) were collected from 345 unrelated healthy Han individuals who lived in eastern China for more than three generations, including 200 individuals living in Wuxi city, Jiangsu Province and 145 individuals living in Hangzhou city, Zhejiang Province, China. The genomic DNA was extracted and quantified using Chelex® 100 resin-based method (Phillips et al., 2012) and the Applied Biosystems® 7,500 Real-Time PCR System (Thermo Fisher Scientific), respectively. Genomic DNA was diluted to 1 ng per microliter with Tris-EDTA buffer and stored at −20°C for later use.

Statistical Analysis

Firstly, we assessed allelic frequency distributions of the same 50 AIDIPs in different continental populations. A heatmap of deletion allelic frequencies of 50 AIDIPs in African, American, East Asian, European, and South Asian populations was plotted by the pheatmap package v1.0.12 of *R* software v4.1.0. Pairwise fixation index (F_{ST}) and informativeness (*In*) values of 50 AIDIPs among continental populations were calculated by Arlequin software v3.5.1.2 (Excoffier et al., 2007) and Infocalc software version 1.1 (Rosenberg et al., 2003), respectively. Then, the F_{ST} and *In* values were graphically displayed by the TBtools software v1.09861 (Chen et al., 2020) and ggplot2 package version 3.3.0 of *R* software, respectively. Population-specific divergences (PSDs) of 50 AIDIPs in each continental population were estimated by a previous report (Phillips, 2015). Next, the performance of these AIDIPs for inferring ancestry origins of continental populations was evaluated by the following methods. Principal component analysis (PCA) of five continental populations was conducted using the Plink software version 1.9 (Chang et al., 2015), and then a scatter plot of these population levels was drawn by the ggplot2 package. Genetic structure of these continental populations was explored by the Admixture software version 1.3 at $K = 2–7$ (Alexander et al., 2009). Thorough analyses of different continental populations were performed by the *Snipper* online tool v2.5 (http://mathgene.usc.es/snipper/) based on 50 AIDIPs.

For the Eastern Han population, allelic frequencies, forensic statistical parameters, HWE tests, and linkage disequilibrium analyses of the 52 AIDIPs were estimated by the STRAF online tool v1.0.5 (Gouy and Zieger, 2017). PCA of Eastern Han and continental populations was also conducted by Plink and ggplot2 packages. Genetic structure of Eastern Han population was assessed by the Admixture software. Different continental populations were viewed as training sets and the Eastern Han population was viewed as the testing samples, and then ancestry origin analyses of Eastern Han population were assessed by the *Snipper*.

RESULTS AND DISCUSSION

Development of a Novel AIDIPs Multiplex Assay

An informative and applicable AIDIPs multiplex assay was developed for simultaneous genotyping of 52 AIDIP loci on the basis of capillary electrophoresis platform. The 52 AIDIP loci were laid out in blue (FAM), green (HEX), yellow (TAMRA), and red (ROX) dye channels according to dye color and expected amplicon size (**Table 1**). The size of amplicons varied from 63 bp at the deletion alleles of loci rs3092383, rs10549914, and rs11576045 to 246 bp at the insertion allele of rs3028297 locus. Generally, full profiles were obtained when various amounts of template DNA (0.2–10 ng) were added, during the testing of the Eastern Han population. However, the optimal concentration of template DNA for this multiplex assay is 0.5–5 ng in a 25-μl PCR final volume. When the amounts of inputted DNA were above 5 ng or below 0.5 ng, the intra-locus and/or intra-color imbalance were randomly observed. As illustrated in **Figure 1**, a complete genotyping profile was obtained when 500 pg of Control DNA 9948 (Promega Corporation, Madison, WI, United States) was added into a 25-μl reaction volume. Compared with AIDIP panels previously reported (Santos et al., 2010; Pereira et al., 2012b; Sun et al., 2016; Lan et al., 2019), the assay developed in this study involved a higher number of AIDIP loci in a single PCR reaction system. More AIDIPs might be more beneficial to discriminate Chinese populations than these reported panels, which remained to be investigated further.

Genetic Distributions of Selected AIDIPs in Different Continental Populations

Although 52 AIDIP loci were selected and successfully incorporated into the novel assay for ancestry origin inference, the population data of rs3033053 and rs1305047 loci were not available in the 1,000 Genome Project. Thus, genetic data of the same 50 AIDIPs were assessed in five different continental populations. To visually display the analytical results, the distributions of deletion allele frequencies of the 50 AIDIPs are shown by a heatmap. As shown in **Figure 2**, the allelic frequencies for the vast majority of these selected AIDIPs varied greatly among different populations. For example, rs67205569, rs10668859, rs149676649, rs3839049, rs3217613, and rs3216128 loci displayed relatively high frequencies in the East Asian populations. It is important to note that 46 AIDIP loci showed almost completely opposite allelic frequency distributions between East Asian and European populations with the exception of rs25630, rs138123572, rs1160852, and rs2307998 loci. It seems to imply that these loci were of considerable potency to distinguish East Asian populations from European populations. Furthermore, we also found that rs25630, rs3217613, rs138123572, rs1160852, and rs2307998 loci exhibited significant allele frequency differences between African and non-African populations: American, European,

TABLE 1 | General information of the 52 AIDIP loci. The numbers 1 and 2 in the "Genotype of 9948" column represent deletion and insertion of nucleotides, respectively.

ID number	Internal code	rs number	Chromosome	Position (GRCh38)	Alleles described in dbSNP	Genotype of 9948	Range of amplicon size (bp)	Fluorescent labels
1	B1	rs3092383	Chr20	46848769	-/AACA	1,2	60–69	FAM
2	B2	rs140864	Chr1	35926061	-/TTC	2	74–82	FAM
3	B3	rs3033053	Chr14	42085293	-/TCAGCAG	2	85–95	FAM
4	B4	rs72375069	Chr3	27386331	-/AATT	2	95.6–102	FAM
5	B5	rs140498743	Chr3	139513672	-/TGTC	1,2	103–109	FAM
6	B6	rs67205569	Chr10	93181810	-/TTGAC	2	110–119	FAM
7	B7	rs74499778	Chr11	130071487	-/AGCT	2	124–130	FAM
8	B8	rs139220746	Chr2	199340972	-/TATC	1	131–137.5	FAM
9	B9	rs10668859	Chr19	266759	-/GAAAG	1,2	139–147	FAM
10	B10	rs140847	Chr9	12617325	-/CGTT	2	162–168	FAM
11	B11	rs16711	Chr17	20179106	-/TTTCTTCCTA	1,2	169–181	FAM
12	B12	rs149676649	Chr5	28495279	-/GATT	2	181.5–188	FAM
13	B13	rs57237250	Chr6	109941799	-/GAGT	1,2	191–198	FAM
14	B14	rs2308163	Chr14	57583363	-/TGAT	2	198.5–213	FAM
15	B15	rs16438	Chr20	25297829	-/CCCAC/CCCCA	1	223–231	FAM
16	B16	rs3028297	Chr9	104604012	-/GCTAA/CTAA	1,2	240.72–250	FAM
17	G1	rs10549914	Chr17	5425659	-/TTTA	2	62–68.5	HEX
18	G2	rs10581451	Chr8	72942426	-/TGAG	2	70–78	HEX
19	G3	rs67934853	Chr2	74716761	-/TAAC	1,2	86–92.5	HEX
20	G4	rs3831920	Chr1	1292285	-/CTCA	2	95–102	HEX
21	G5	rs1160852	Chr6	137024720	-/TT	2	108–113	HEX
22	G6	rs25630	Chr6	14734110	-/AG	1	117–123	HEX
23	G7	rs2307998	Chr5	7814232	-/GGA	2	127–135	HEX
24	G8	rs138123572	Chr15	72493894	-/TGAC	2	142–151	HEX
25	G9	rs3839049	Chr2	26254260	-/ACT	2	154–160	HEX
26	G10	rs2307840	Chr1	35633488	-/GT	1	168–174	HEX
27	G11	rs35779249	Chr13	43390341	-/TAA	1	178–185	HEX
28	G12	rs1305047	Chr17	16181674	-/CACA	1,2	186–193	HEX
29	G13	rs66693708	Chr12	77004626	-/TAAG	2	195–202	HEX
30	G14	rs2308036	Chr15	64914812	-/CC	1	218–224	HEX
31	G15	rs3074939	Chr21	42002321	-/CAGT	1	225–232	HEX
32	Y1	rs11576045	Chr12	111361720	-/ACA	1	60–69	TAMRA
33	Y2	rs3217613	Chr15	84932992	-/ATA	2	80–85	TAMRA
34	Y3	rs3059936	Chr11	112701065	-/AT	1,2	86–90	TAMRA
35	Y4	rs3840274	Chr4	68494125	-/CTCA	2	95–102	TAMRA
36	Y5	rs3840614	Chr7	78029712	-/TTC	1,2	119–124	TAMRA
37	Y6	rs3033100	Chr4	140872558	-/CAG	1,2	136–145	TAMRA
38	Y7	rs3838001	Chr20	63684046	-/CAA	2	162–168	TAMRA
39	Y8	rs3049003	Chr7	6660366	-/AT	1,2	172–178	TAMRA
40	Y9	rs3051160	Chr10	100927312	-/TG	1	182–188	TAMRA
41	Y10	rs5796380	Chr12	10124046	-/AAG	1	203–209	TAMRA
42	R1	rs5783058	Chr10	8742655	-/TGTT	1	66–74	ROX
43	R2	rs3216128	Chr21	42559334	-/AGA	2	84–90	ROX
44	R3	rs5822884	Chr18	5980141	-/TAGT	1	103–110	ROX
45	R4	rs3053514	Chr21	28691710	-/TAC	1	113–118	ROX
46	R5	rs3840019	Chr15	65752777	-/AATT	2	121–128	ROX
47	R6	rs1610951	Chr5	109664135	-/CCAA	2	144–151	ROX
48	R7	rs1305057	Chr5	57319423	-/TGTTTCA	1,2	165–175	ROX
49	R8	rs3073179	Chr11	18237493	-/AT	1,2	176–181	ROX
50	R9	rs2307727	Chr2	135675653	-/TT	1	197–203	ROX
51	R10	rs5891726	Chr8	59361584	-/TACT	1	213–220	ROX
52	R11	rs5824539	Chr18	44391691	-/TA	2	221–224	ROX

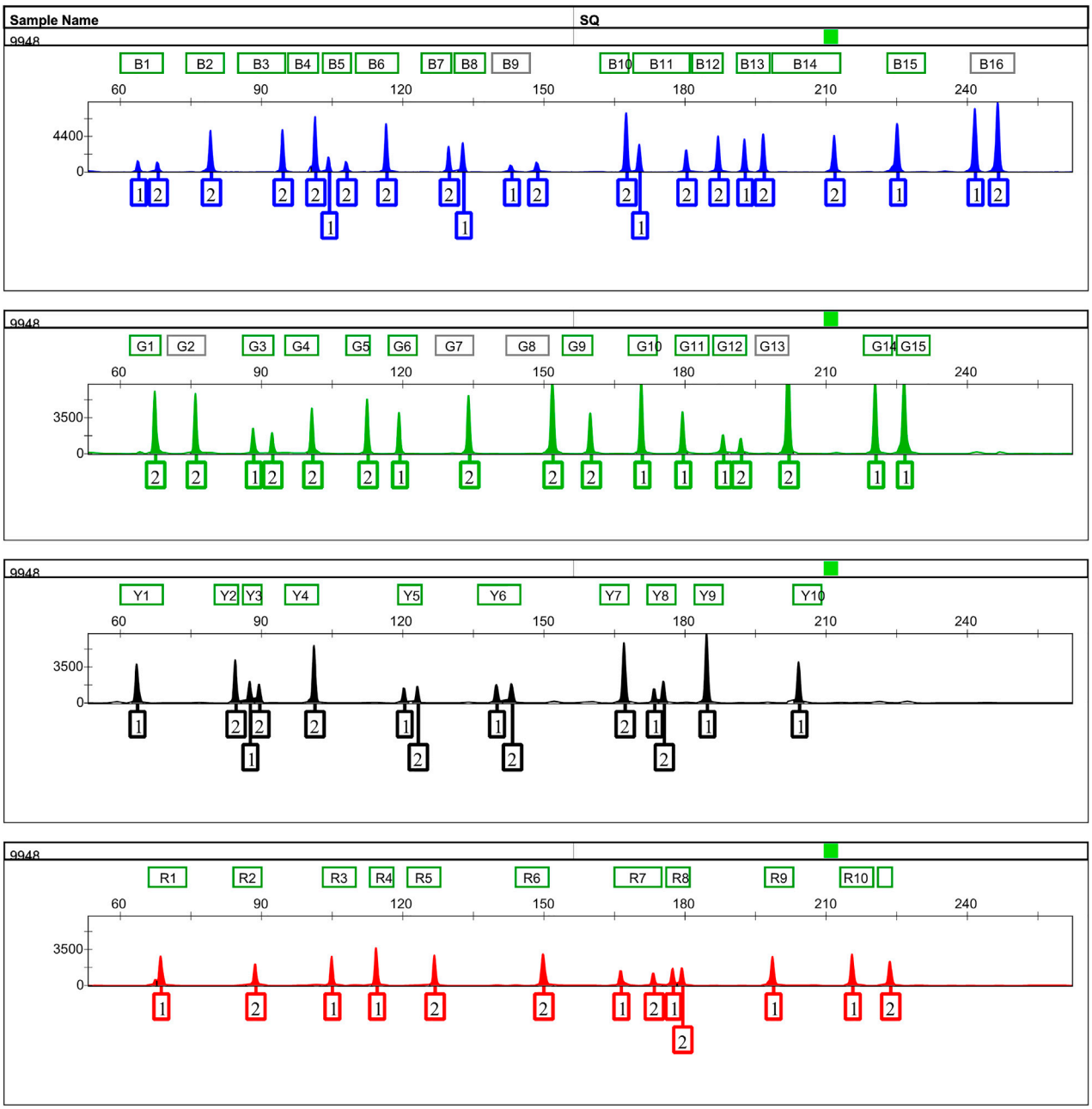

FIGURE 1 | Representative 500 pg of Control DNA 9948 profile amplified with the 52-AIDIP panel for 25-μl reaction volumes. Five hundred picograms of Control DNA 9948 was amplified with the developed panel for 30 cycles. One microliter of PCR product added into 12.5 μl of loading mixtures (12 μl Hi-Di formamide +0.5 μl SIZ-500 Size Standard) was electrophoresed on a 3500xL Genetic Analyzer using the default injection conditions.

South Asian, and East Asian populations. However, the differences of allele distributions between American and European/South Asian populations were relatively small for most loci.

To reveal genetic divergences of these AIDIPs among different continental populations better, pairwise F_{ST} values were also calculated, as shown in **Figure 3**. Results revealed that most loci showed relatively high F_{ST} values between East Asian and other continental populations, especially between East Asians and Europeans, whereas most loci showed low F_{ST} values between American and European/South Asian populations. *In* is commonly used to evaluate the ancestral information of genetic markers in different populations (Phillips, 2015). Hence, the pairwise *In* values of these 50 AIDIPs were also estimated. Similar to F_{ST} values, most loci showed high *In* values between East Asian and other continental populations (**Supplementary Figure 1**). Shriver et al. stated that the developed AIM panel should possess balance differentiation efficiencies among each population, which could

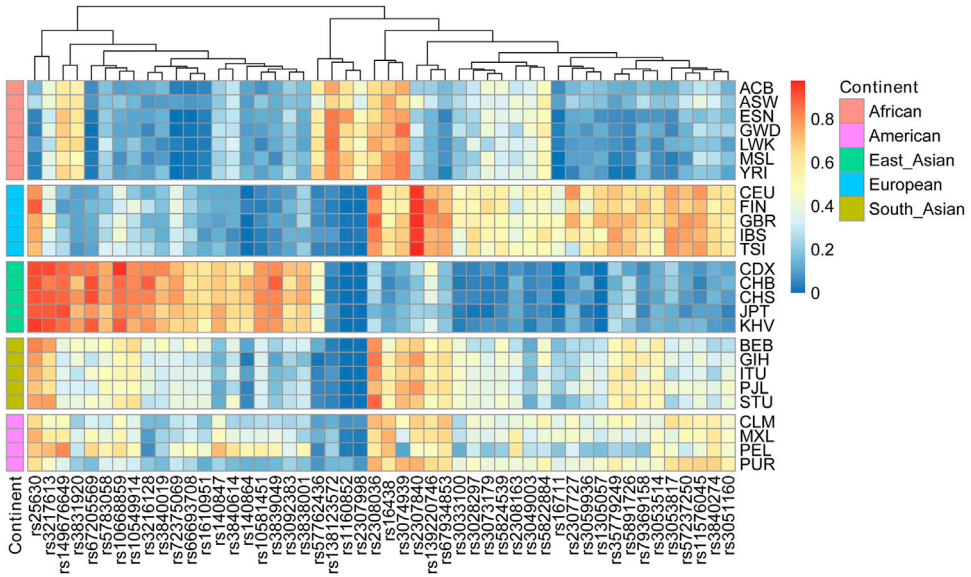

FIGURE 2 | Heatmap of deletion allelic frequencies of 50 ancestry informative DIPs in different continental populations.

FIGURE 3 | Pairwise F_{ST} values of five continental populations for 50 ancestry informative DIPs.

bring little bias into ancestral components of admixed individuals (Shriver et al., 2004). Therefore, we assessed the cumulative PSD values of 50 AIDIPs in the five continental populations. Results demonstrated that these 50 loci showed the highest cumulative PSD values in the East Asian population, followed by African, European, South Asian, and American populations (**Supplementary Figure 2**). In this study, to infer ancestry origins of East Asian populations more accurately, we selected AIDIP loci that showed high genetic variations between East Asian and other continental populations, resulting in higher cumulative PSD values in East Asian populations. In addition, we found that 50 AIDIP loci also showed relatively high cumulative PSD values in European and African populations.

FIGURE 4 | Ancestry origin analyses of different continental populations based on 50 ancestry informative DIPs. **(A)** PCA of different continental populations at PC1, PC2, and PC3. **(B)** Population genetic structure analyses of different continental populations at K = 2–7.

Nonetheless, relatively low cumulative PSD values of these 50 loci in South Asian and American populations suggested that they might not be suitable for ancestry origin analyses of these two intercontinental populations.

Ancestry Resolutions of the Developed AIDIP Panel for Continental Populations

Here, the PCA was primarily conducted on the basis of the same 50 AIDIPs to evaluate the capacity of the developed AIDIP assay to differentiate continental populations. Results of the PCA analysis for the five continental populations are shown in **Figure 4A**. At PC1, African, European, and East Asian individuals formed three population clusters, respectively, and they could be clearly separated from each other. At PC3, some South Asian and American individuals could be differentiated from other continental populations. Subsequently, the genetic structure of these continental populations was also explored. The results with K ranging from two to seven are presented in **Figure 4B**. At K = 2, five East Asian populations exhibited high blue components and could be discriminated from other populations. As K becomes 3, African, European, and East Asian populations showed their distinct ancestral components, respectively. Moreover, American and South Asian populations showed similar ancestral component distributions. When K increased to 4, South Asian populations could be separated from other populations. No more significant changes in population structure were observed from the bar plot when the K values were greater than 4. These results demonstrated that the novel AIDIP panel could clearly differentiate African, European, and East Asian populations. The capacity of this assay to differentiate continental populations is similar to those of previously reported panels (Santos et al., 2010; Pereira et al., 2012b; Lan et al., 2019). Nevertheless, unlike the weaker capacity of the 46-AIM-InDels panel to differentiate the East Asian population (Pereira et al., 2012b), the current AIDIP panel revealed an excellent characteristic to estimate the ancestry information of East Asians.

The *Snipper* online tool was developed to infer ancestry origins of populations by the Bayesian method (Santos et al., 2016b). Therefore, we further evaluated ancestry resolutions of 50 AIDIPs for continental populations by the *Snipper*. Results indicated that most individuals from African, European, East Asian, and South Asian populations could be classified into correct continental origins, whereas some individuals from American populations were classified into European and South Asian populations (**Supplementary Figure 3**). Admixed genetic background of American populations went against their ancestry origin

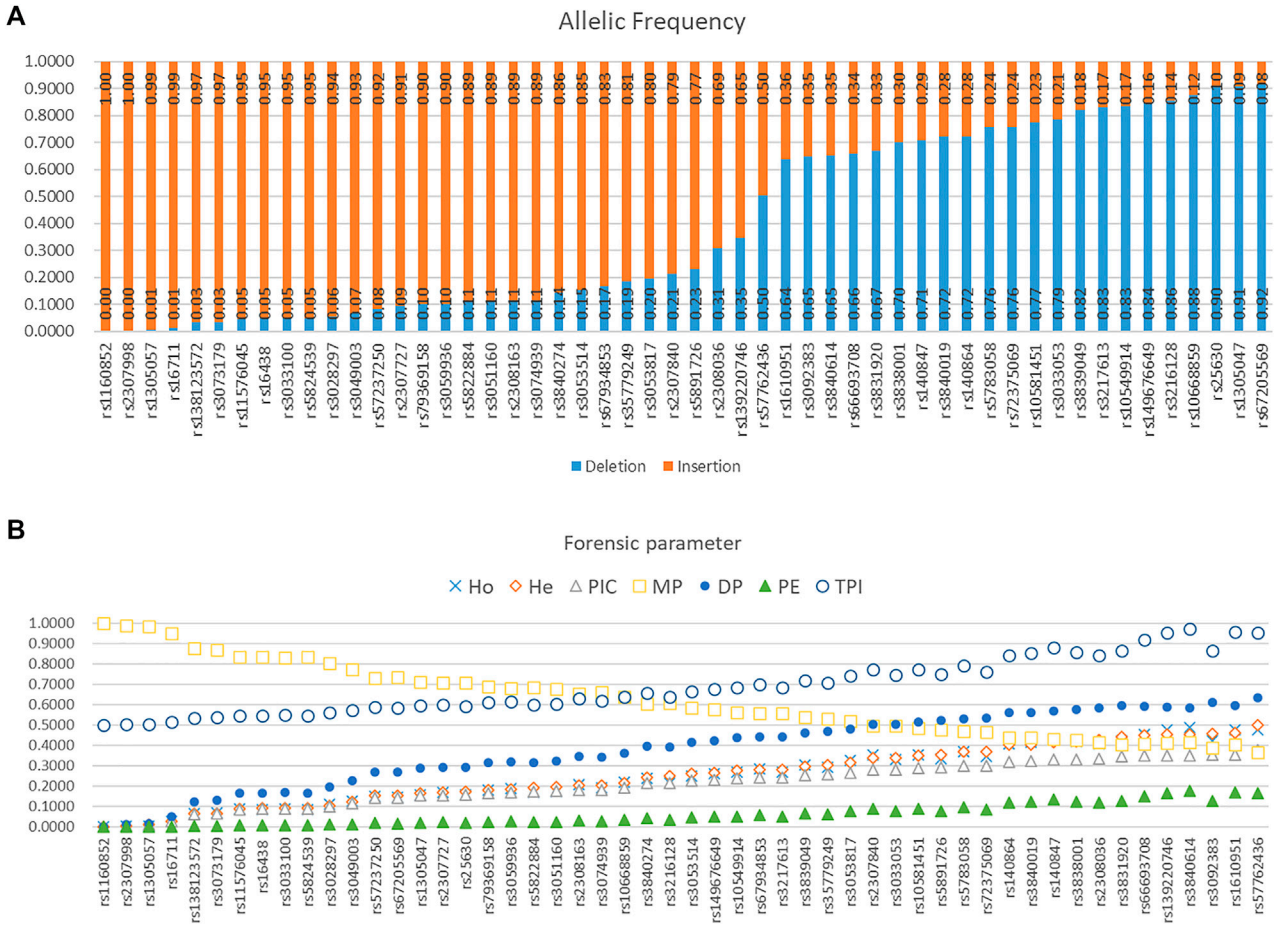

FIGURE 5 | Allelic frequencies and forensic parameters of 52 ancestry informative DIPs in an Eastern Han population. **(A)** Allelic frequency. **(B)** Forensic parameter.

inferences (Genomes Project et al., 2015). In addition, relatively few American-specific genetic markers in the extant panel might also lead to this result. Even so, obtained results revealed that these 50 AIDIPs could be utilized to differentiate African, European, and East Asian populations well.

Allelic Frequencies and Forensic Statistical Parameters of 52 AIDIPs in Eastern Han Population

HWE tests of 52 AIDIPs in the Eastern Han population are given in **Supplementary Table 1**. No loci were observed to deviate from HWE after applying Bonferroni correction ($p = 0.05/52 = 0.00096$). Linkage disequilibrium (LD) analyses of pairwise AIDIPs in the Eastern Han population are listed in **Supplementary Table 2**. The results showed that a significant association between rs2307840 and rs140864 loci was revealed, even after applying Bonferroni correction ($p = 0.05/1,326 = 0.000037$). LD between the two loci in the Eastern Han population may be caused by genetic linkage because both of them are located on chromosome one and just 292,573 bp apart from each other. The locus rs140864 exhibited a little more excellent characteristic of forensic statistical parameters, therefore, it was preferentially selected for further data analysis in Eastern Han population. Furthermore, to better understand the associations among loci in different population groups, further evaluations containing more populations and lager sample sizes need to be investigated.

Allelic frequencies of 52 AIDIPs in the Eastern Han population are presented in **Figure 5A** and **Supplementary Table 1**. Deletion allelic frequencies of these loci ranged from 0.0000 to 0.9159. We also calculated forensic parameters of these loci in Eastern Han population, as given in **Figure 5B** and **Supplementary Table 1**. Mean observed heterozygosity (Ho), expected heterozygosity (He), polymorphism information content (PIC), match probability (MP), discrimination power (DP), power of exclusion (PE), and typical paternity index (TPI) of 52 AIDIPs in Eastern Han population were

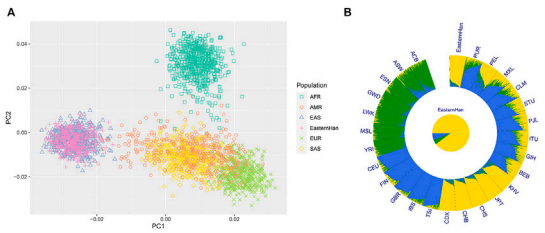

FIGURE 6 | Ancestry component analyses of the studied Eastern Han populations in comparisons with other continental populations. **(A)** PCA of Eastern Han and continental populations at PC1 and PC2. **(B)** Population genetic structure analyses of Eastern Han and continental populations at $K = 3$.

0.2491, 0.2527, 0.2107, 0.6215, 0.3785, 0.0594, and 0.6910, respectively. Cumulative DP and PE of these loci in the Eastern Han population were 0.999 999 999 9977646 and 0.9619, respectively. As expected, these loci exhibited relatively low genetic diversities in the Eastern Han population. Even so, relatively high cumulative DP indicated that these loci could be viewed as a supplementary tool for forensic identity testing in the Eastern Han population.

Ancestry Component Dissections of Eastern Han Populations by 50 AIDIPs

Based on the raw data of 50 AIDIPs, PCA of Eastern Han and continental populations was conducted. We found that the studied Eastern Han individuals were predominately superimposed on the East Asian individual cluster located on the right part of the plot (**Figure 6A**). Ancestral components of the Eastern Han populations were also assessed in comparisons to five continental populations, as presented in **Figure 6B**. The studied Eastern Han population displayed high ancestral components from East Asian populations. Subsequently, we treated five continental populations as training set and Eastern Han population as unknown population and explored the power of these AIDIPs to infer ancestral origins of Eastern Han population by the *Snipper*. The obtained results revealed that all Eastern Han individuals could be categorized into East Asian population, implying that these AIDIPs could perform ancestry origin analyses of Eastern Han population well. Besides, these results also reflected that the studied Eastern Han population had intimate genetic relationships with East Asian populations.

Lang et al. assessed genetic structure of Eastern Han population by 27 Y-STRs and 143 Y-SNPs and found that the Han populations showed closer genetic affinities with East Asian populations than South Asian populations. Furthermore, they also pointed out that genetic differentiations between Southern Han and Northern Han populations were observed (Lang et al., 2019). Lu et al. investigated genetic distributions of 17 autosomal STRs in an Eastern Han population (Jiangsu Han) and they found that the Han population showed low genetic divergences with Hubei Han populations (Lu et al., 2019). Chiang et al. conducted a comprehensive analysis of genetic variations in Chinese Han populations and found an east–west differentiation among Han populations except for a known south–north cline (Chiang et al., 2018). Moreover, Li et al. exploited the genetic landscape of Chinese Han populations based on the mitochondria DNA and revealed that genetic divergences among Han populations residing in different river systems existed (Li et al., 2019). On this basis, we speculated that genetic substructure potentially existed among different Han populations in China. Consequently, we intend to investigate genetic polymorphism distributions of selected 52 AIDIPs in Han populations from different regions. Those studies can not only depict the genetic architecture of different Han Chinese populations, but also contribute to screen region-specific genetic markers. Moreover, due to the large allele frequency differences between European and East Asian populations of these AIDIP loci selected in the present study, next we intend to explore the capacity of this novel assay to infer the ancestral origins of groups with admixed Eurasian ancestry in China.

CONCLUSION

In summary, we developed a multiplex PCR panel for ancestry origin predictions of different continental populations that contained 52 AIDIP loci. Most loci out of these 52 AIDIPs showed high genetic divergences between East Asian and non-East Asian populations. We also demonstrated that this AIDIP panel could be employed for inferring biogeographical origins of continental populations, specifically for East Asian, African, and European populations. In addition, these 52 AIDIP loci also showed relatively high application values for forensic identity testing in the Eastern Han population. For ancestral component analysis of the Eastern Han population, the novel panel could accurately estimate its close genetic affinities and high ancestral components with East Asian populations. In the future, we need to assess genetic distributions of the 52 AIDIPs in other populations from different regions to unveil genetic portraits of these populations. Only in this way could the performance of the developed panel to infer sub-populations and estimate inter-ethnic admixture proportions be completely understood.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available to maintain the participants privacy. Requests to access the datasets should be directed to the corresponding author, BZ.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Xi'an Jiaotong University, China. The patients/participants provided their written informed consents to participate in this study.

AUTHOR CONTRIBUTIONS

YZ, Investigation, Sample collection, Methodology, Data curation, and Manuscript preparation and revision; XJ, Data curation, Formal analysis, Visualization, and Manuscript preparation and revision; BW, Conceptualization, Supervision, and Manuscript review and editing; BZ, Conceptualization, Supervision, Resources, Funding acquisition, Project administration, and Manuscript review and editing. All listed authors have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

FUNDING

This project was supported by the National Natural Science Foundation of China (No. 81772031), GDUPS (2017).

ACKNOWLEDGMENTS

The authors are grateful to all the volunteers that took part in the present study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.801275/full#supplementary-material

Supplementary Figure 1 | Pairwise *In* values of five continental populations for 50 ancestry informative DIPs.

Supplementary Figure 2 | Cumulative PSD values of 50 ancestry informative DIPs in different continental populations.

Supplementary Figure 3 | Ancestry origin analyses of different continental populations by the *Snipper* 2.5 online tool.

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast Model-Based Estimation of Ancestry in Unrelated Individuals. *Genome Res.* 19 (9), 1655–1664. doi:10.1101/gr.094052.109

Carvalho Gontijo, C., Porras-Hurtado, L. G., Freire-Aradas, A., Fondevila, M., Santos, C., Salas, A., et al. (2020). PIMA: A Population Informative Multiplex for the Americas. *Forensic Sci. Int. Genet.* 44, 102200. doi:10.1016/j.fsigen.2019.102200

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, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Molecular Plant* 13 (8), 1194–1202. doi:10.1016/j.molp.2020.06.009

Chen, L., Du, W., Wu, W., Yu, A., Pan, X., Feng, P., et al. (2019). Developmental Validation of a Novel Six-Dye Typing System with 47 A-InDels and 2 Y-InDels. *Forensic Sci. Int. Genet.* 40, 64–73. doi:10.1016/j.fsigen.2019.02.009

Chen, L., Pan, X., Wang, Y., Du, W., Wu, W., Tang, Z., et al. (2021). Development and Validation of a Forensic Multiplex System with 38 X-InDel Loci. *Front. Genet.* 12, 670482. doi:10.3389/fgene.2021.670482

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 (11), 2736–2750. doi:10.1093/molbev/msy170

Excoffier, L., Laval, G., and Schneider, S. (2007). Arlequin (Version 3.0): an Integrated Software Package for Population Genetics Data Analysis. *Evol. Bioinform Online* 1, 47–50.

Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A Global Reference for Human Genetic Variation. *Nature* 526 (7571), 68–74. doi:10.1038/nature15393

Gouy, A., and Zieger, M. (2017). STRAF-A Convenient Online Tool for STR Data Evaluation in Forensic Genetics. *Forensic Sci. Int. Genet.* 30, 148–151. doi:10.1016/j.fsigen.2017.07.007

Jin, X.-Y., Wei, Y.-Y., Lan, Q., Cui, W., Chen, C., Guo, Y.-X., et al. (2019). A Set of Novel SNP Loci for Differentiating continental Populations and Three Chinese Populations. *PeerJ* 7, e6508. doi:10.7717/peerj.6508

Lan, Q., Shen, C., Jin, X., Guo, Y., Xie, T., Chen, C., et al. (2019). Distinguishing Three Distinct Biogeographic Regions with an In-house Developed 39-AIM-InDel Panel and Further Admixture Proportion Estimation for Uyghurs. *Electrophoresis* 40 (11), 1525–1534. doi:10.1002/elps.201800448

Lang, M., Liu, H., Song, F., Qiao, X., Ye, Y., Ren, H., et al. (2019). Forensic Characteristics and Genetic Analysis of Both 27 Y-STRs and 143 Y-SNPs in Eastern Han Chinese Population. *Forensic Sci. Int. Genet.* 42, e13–e20. doi:10.1016/j.fsigen.2019.07.011

LaRue, B. L., Lagacé, R., Chang, C.-W., Holt, A., Hennessy, L., Ge, J., et al. (2014). Characterization of 114 Insertion/deletion (INDEL) Polymorphisms, and Selection for a Global INDEL Panel for Human Identification. *Leg. Med.* 16 (1), 26–32. doi:10.1016/j.legalmed.2013.10.006

Li, C., Zhang, S., Li, L., Chen, J., Liu, Y., and Zhao, S. (2012). Selection of 29 Highly Informative InDel Markers for Human Identification and Paternity Analysis in Chinese Han Population by the SNPlex Genotyping System. *Mol. Biol. Rep.* 39 (3), 3143–3152. doi:10.1007/s11033-011-1080-z

Li, Y.-C., Ye, W.-J., Jiang, C.-G., Zeng, Z., Tian, J.-Y., Yang, L.-Q., et al. (2019). River Valleys Shaped the Maternal Genetic Landscape of Han Chinese. *Mol. Biol. Evol.* 36 (8), 1643–1652. doi:10.1093/molbev/msz072

Lu, Y., Sun, H.-j., Zhou, J.-c., and Wu, X. (2019). Genetic Polymorphisms, Forensic Efficiency and Phylogenetic Analysis of 17 Autosomal STR Loci in the Han Population of Wuxi, Eastern China. *Ann. Hum. Biol.* 46 (7-8), 601–605. doi:10.1080/03014460.2019.1693628

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., et al. (2006). An Initial Map of Insertion and Deletion (INDEL) Variation in the Human Genome. *Genome Res.* 16 (9), 1182–1190. doi:10.1101/gr.4565806

Pereira, R., Pereira, V., Gomes, I., Tomas, C., Morling, N., Amorim, A., et al. (2012a). A Method for the Analysis of 32 X Chromosome Insertion Deletion Polymorphisms in a Single PCR. *Int. J. Leg. Med.* 126 (1), 97–105. doi:10.1007/s00414-011-0593-2

Pereira, R., Phillips, C., Alves, C., Amorim, A., Carracedo, Á., and Gusmão, L. (2009). A New Multiplex for Human Identification Using Insertion/deletion Polymorphisms. *Electrophoresis* 30 (21), 3682–3690. doi:10.1002/elps.200900274

Pereira, R., Phillips, C., Pinto, N., Santos, C., Santos, S. E. B. d., Amorim, A., et al. (2012b). Straightforward Inference of Ancestry and Admixture Proportions through Ancestry-Informative Insertion Deletion Multiplexing. *PLoS One* 7 (1), e29684. doi:10.1371/journal.pone.0029684

Phillips, C., and de la Puente, M. (2021). The Analysis of Ancestry with Small-Scale Forensic Panels of Genetic Markers. *Emerg. Top. Life Sci.* 5 (3), 443–453. doi:10.1042/ETLS20200327

Phillips, C. (2015). Forensic Genetic Analysis of Bio-Geographical Ancestry. *Forensic Sci. Int. Genet.* 18, 49–65. doi:10.1016/j.fsigen.2015.05.012

Phillips, C., Salas, A., Sánchez, J. J., Fondevila, M., Gómez-Tato, A., Álvarez-Dios, J., et al. (2007). Inferring Ancestral Origin Using a Single Multiplex Assay of Ancestry-Informative Marker SNPs. *Forensic Sci. Int. Genet.* 1 (3-4), 273–280. doi:10.1016/j.fsigen.2007.06.008

Phillips, K., McCallum, N., and Welch, L. (2012). A Comparison of Methods for Forensic DNA Extraction: Chelex-100 and the Qiagen DNA Investigator Kit (Manual and Automated). *Forensic Sci. Int. Genet.* 6 (2), 282–285. doi:10.1016/j.fsigen.2011.04.018

Qu, S., Zhu, J., Wang, Y., Yin, L., Lv, M., Wang, L., et al. (2019). Establishing a Second-Tier Panel of 18 Ancestry Informative Markers to Improve Ancestry

Distinctions Among Asian Populations. *Forensic Sci. Int. Genet.* 41, 159–167. doi:10.1016/j.fsigen.2019.05.001

Rosenberg, N. A., Li, L. M., Ward, R., and Pritchard, J. K. (2003). Informativeness of Genetic Markers for Inference of Ancestry*. *Am. J. Hum. Genet.* 73 (6), 1402–1422. doi:10.1086/380416

Santos, C., Phillips, C., Fondevila, M., Daniel, R., van Oorschot, R. A. H., Burchard, E. G., et al. (2016a). Pacifiplex : an Ancestry-Informative SNP Panel Centred on Australia and the Pacific Region. *Forensic Sci. Int. Genet.* 20, 71–80. doi:10.1016/j.fsigen.2015.10.003

Santos, C., Phillips, C., Gomez-Tato, A., Alvarez-Dios, J., Carracedo, Á., and Lareu, M. V. (2016b). Inference of Ancestry in Forensic Analysis II: Analysis of Genetic Data. *Methods Mol. Biol.* 1420, 255–285. doi:10.1007/978-1-4939-3597-0_19

Santos, N. P. C., Ribeiro-Rodrigues, E. M., Ribeiro-Dos-Santos, Â. K. C., Pereira, R., Gusmão, L., Amorim, A., et al. (2010). Assessing Individual Interethnic Admixture and Population Substructure Using a 48-Insertion-Deletion (INSEL) Ancestry-Informative Marker (AIM) Panel. *Hum. Mutat.* 31 (2), 184–190. doi:10.1002/humu.21159

Shriver, M. D., Kennedy, G. C., Parra, E. J., Lawson, H. A., Sonpar, V., Huang, J., et al. (2004). The Genomic Distribution of Population Substructure in Four Populations Using 8,525 Autosomal SNPs. *Hum. Genomics* 1 (4), 274–286. doi:10.1186/1479-7364-1-4-274

Sun, K., Ye, Y., Luo, T., and Hou, Y. (2016). Multi-InDel Analysis for Ancestry Inference of Sub-populations in China. *Sci. Rep.* 6, 39797. doi:10.1038/srep39797

Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C., and Marth, G. (2002). Human Diallelic Insertion/deletion Polymorphisms. *Am. J. Hum. Genet.* 71 (4), 854–862. doi:10.1086/342727

Wei, Y.-L., Wei, L., Zhao, L., Sun, Q.-F., Jiang, L., Zhang, T., et al. (2016). A Single-Tube 27-plex SNP Assay for Estimating Individual Ancestry and Admixture from Three Continents. *Int. J. Leg. Med.* 130 (1), 27–37. doi:10.1007/s00414-015-1183-5

Xavier, C., de la Puente, M., Phillips, C., Eduardoff, M., Heidegger, A., Mosquera-Miguel, A., et al. (2020). Forensic Evaluation of the Asia Pacific Ancestry-Informative MAPlex Assay. *Forensic Sci. Int. Genet.* 48, 102344. doi:10.1016/j.fsigen.2020.102344

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.

The handling editor and the reviewer ZW declared as past co-authorship with one of the authors BZ.

frontiers in Genetics

ORIGINAL RESEARCH
published: 23 February 2022
doi: 10.3389/fgene.2022.729514

Genetic Background of Kirgiz Ethnic Group From Northwest China Revealed by Mitochondrial DNA Control Region Sequences on Massively Parallel Sequencing

Hongdan Wang[1,2†], *Man Chen*[3,4†], *Chong Chen*[1,5], *Yating Fang*[3,4], *Wei Cui*[3,4], *Fanzhang Lei*[3,4] *and Bofeng Zhu*[1,3,4,5*]

[1]Key Laboratory of Shaanxi Province for Craniofacial Precision Medicine Research, College of Stomatology, Xi'an Jiaotong University, Xi'an, China, [2]Medical Genetic Institute of Henan Province, Henan Provincial People's Hospital, Zhengzhou University People's Hospital, Zhengzhou, China, [3]Guangzhou Key Laboratory of Forensic Multi-Omics for Precision Identification, School of Forensic Medicine, Southern Medical University, Guangzhou, China, [4]Multi-Omics Innovative Research Center of Forensic Identification, Department of Forensic Genetics, School of Forensic Medicine, Southern Medical University, Guangzhou, China, [5]College of Forensic Medicine, Xi'an Jiaotong University, Xi'an, China

OPEN ACCESS

Edited by:
Wibhu Kutanan,
Khon Kaen University, Thailand

Reviewed by:
Qing-Peng Kong,
Kunming Institute of Zoology (CAS), China
Antonia Picornell,
University of the Balearic Islands, Spain

***Correspondence:**
Bofeng Zhu
zhubofeng7372@126.com

[†]These authors have contributed equally to this work

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

Received: 23 June 2021
Accepted: 24 January 2022
Published: 23 February 2022

Citation:
Wang H, Chen M, Chen C, Fang Y, Cui W, Lei F and Zhu B (2022) Genetic Background of Kirgiz Ethnic Group From Northwest China Revealed by Mitochondrial DNA Control Region Sequences on Massively Parallel Sequencing. Front. Genet. 13:729514. doi: 10.3389/fgene.2022.729514

The mitochondrial DNA (mtDNA) has been used to trace population evolution and apply to forensic identification due to the characteristics including lack of recombination, higher copy number and matrilineal inheritance comparing with nuclear genome DNA. In this study, mtDNA control region sequences of 91 Kirgiz individuals from the Northwest region of China were sequenced to identify genetic polymorphisms and gain insight into the genetic background of the Kirgiz ethnic group. MtDNA control region sequences of Kirgiz individuals presented relatively high genetic polymorphisms. The 1,122 bp sequences of mtDNA control region could differ among unrelated Kirgiz individuals, which suggested the mtDNA control region sequences have a good maternal pedigree tracing capability among different Kirgiz individuals. The neutrality test, mismatch distribution, Bayesian phylogenetic inference, Bayesian skyline analysis, and the median network analyses showed that the Kirgiz group might occurred population expansion, and the expansion could be observed at about ~53.41 kilo years ago (kya) when ancestries of modern humans began to thrive in Eurasia. The pairwise population comparisons, principal component analyses, and median network analyses were performed based on haplogroup frequencies or mtDNA control region sequences of 5,886 individuals from the Kirgiz group and the 48 reference populations all over the world. And the most homologous haplotypes were found between Kirgiz individuals and the East Asian individuals, which indicated that the Kirgiz group might have gene exchanges with the East Asian populations.

Keywords: mitochondrial DNA control region, massively parallel sequencing, Kirgiz ethnic group, genetic polymorphism, population evolution

INTRODUCTION

The Kirgiz group is one of the official minority ethnic group in China. According to the seventh census, there are about 186 thousand Kirgiz individuals in China, which are mainly located in the Northwest China. Additionally, hundreds of Kirgiz individuals are settled in Chinese Heilongjiang Province. The Kirgizs speak Kepchak, which is a subgroup of the Turkic group of Altaic language family. In China, the history of the Kirgiz group can be traced back to the period of Emperor Wu of the Western Han Dynasty (109–91 BC), and since then the Kirgizs have the appearance characteristics of both European and East Asian people according to the historical record (Gordon, 2009). Later, Kirgiz ancestors gradually expanded geographically and spatially, and experienced the "Kigu", "Pikasi", and "Bulgari" periods successively during Tang Dynasty and Qing Dynasty (Abramzon and Tabyshaliev, 1990).

In previous studies researchers Guo et al. (Guo et al., 2018), Xie et al. (Xie et al., 2020) and Zhang et al. (Zhang et al., 2021) clarified the genetic background of Kirgiz group based on insertion/deletion (InDel), including the 30 commercial InDel system and self-developed 39 ancestry informative marker (AIM) InDel system, respectively. The above mentioned researches demonstrated that Kirgiz group had the relatively close genetic distances with Kazakh and Hui groups based on autosomal InDel genetic markers. Wang et al. (2019) explored the genetic diversity of Kirgiz group based on the presence/absence polymorphisms of killer cell immunoglobulin like receptor genes. And the research indicated that the Kirgiz group represented small genetic differences with populations speaking the same family language. The genetic population studies were conducted by Guo et al., Song et al., and Chen et al., they studied the genetic diversity distributions of Kirgiz group on basis of 60 single nucleotide polymorphisms (SNPs) in mtDNA and 24 Y chromosomal short tandem repeats (Y-STRs) (Guo et al., 2020); 17 Y chromosomal SNPs (Y-SNPs) and 27 Y-STRs (Song et al., 2021); and 23 autosomal-STRs (Pengyu Chen et al., 2019), respectively. And they concluded that the Kirgiz group had the genetic admixture of East Asia and Europe after comparing with the other continental populations.

The control region, also known as the D-loop region, is a sequence of 1,122 bp on the mtDNA, including two segments at 1-576 and 16024-16569, respectively (Anderson et al., 1981). The multiple copies of mtDNA, and the resistance to degradation with the circular structure make mtDNA more suitable for forensic trace and degraded samples than nuclear DNA (Gallimore et al., 2018; Amorim et al., 2019). The high mutation rates of mtDNA sequences make that the sequences are high polymorphisms especially in the control region. And the highly polymorphic mtDNA control region sequences have the potential to distinguish unrelated individuals (Strobl et al., 2019; Ta et al., 2021). The characteristics of the mtDNA maternal inheritance can be used to track the maternal family (Bandelt and Dür, 2007; Parson et al., 2014), and can also be used to track the genetic relationships among the Kirgiz group and other reference populations.

Based on the massively parallel sequencing (MPS) platform, the influences of multiple copies of mtDNA and the homologous fragments of nuclear DNA make that the mtDNA data often require deeper sequencing depth, and well-balanced sequencing read and amplicon to ensure the obtained reliable results (Amorim et al., 2019). In addition, having specific mtDNA control region sequences among different maternal pedigrees is the prerequisite basis for forensic maternal tracing application. And the above-mentioned characteristics would be reflected in the high genetic diversities among unrelated individuals (Scally, 2016). This study was carried out to evaluate the forensic application efficiencies of the mtDNA control region sequences in the Kirgiz group, and to uncover possible historical events in the Kirgiz origin, and to disclose the genetic affinities between Kirgiz and Chinese other populations or international populations. In this study, the mtDNA control region sequences in 91 unrelated healthy individuals from Kirgiz group were sequenced to analyze the genetic structure of Kirgiz group based on the maternal genetic materials. Furthermore, the mtDNA control region sequences of 5,795 individuals from 48 previously published populations were also collected to compare genetic differences among populations and gain insight into the genetic background of Kirgiz group.

MATERIALS AND METHODS

Sample Collection

In total, bloodstains within FTA™ cards were collected from 91 unrelated healthy individuals of the Kirgiz group in the Northwest region of China. The detailed sampling locations of all 49 populations including Kirgiz group and 48 reference populations were shown on the world map in **Figure 1**. All participants were informed in details of the content and purpose of the present study, and all signed the written informed consents. This study was approved by the Ethics Committee of Xi'an Jiaotong University Health Science Center (Approval Number: XJTULAC201) and complied with the ethical principle of the World Medical Association Declaration of Helsinki (World Medical Association, 2013). According to the questionnaire survey, the selected donors of random sampling were all local residents and their immediate family members (parents, grandparents, and maternal grandparents) were all Kirgiz individuals. To protect the privacy of volunteers, all samples were numbered and anonymized during the whole experiments.

Library Preparation and Sequencing

Genome mtDNA of the 91 bloodstain samples was extracted with the PrepFiler BTA™ forensic DNA extraction kit (Thermo Fisher Scientific, MA, United States). The magnetic bead method was used for DNA extraction by the above kit. The extraction method selected in this study combined uniquely structured magnetic particles and the optimized multi-component chemistry surface, to provide efficient DNA binding capacity and high DNA recovery rate. The extracted genome DNA was quantified with the Qubit™ dsDNA HS assay kit (Thermo Fisher Scientific) at the

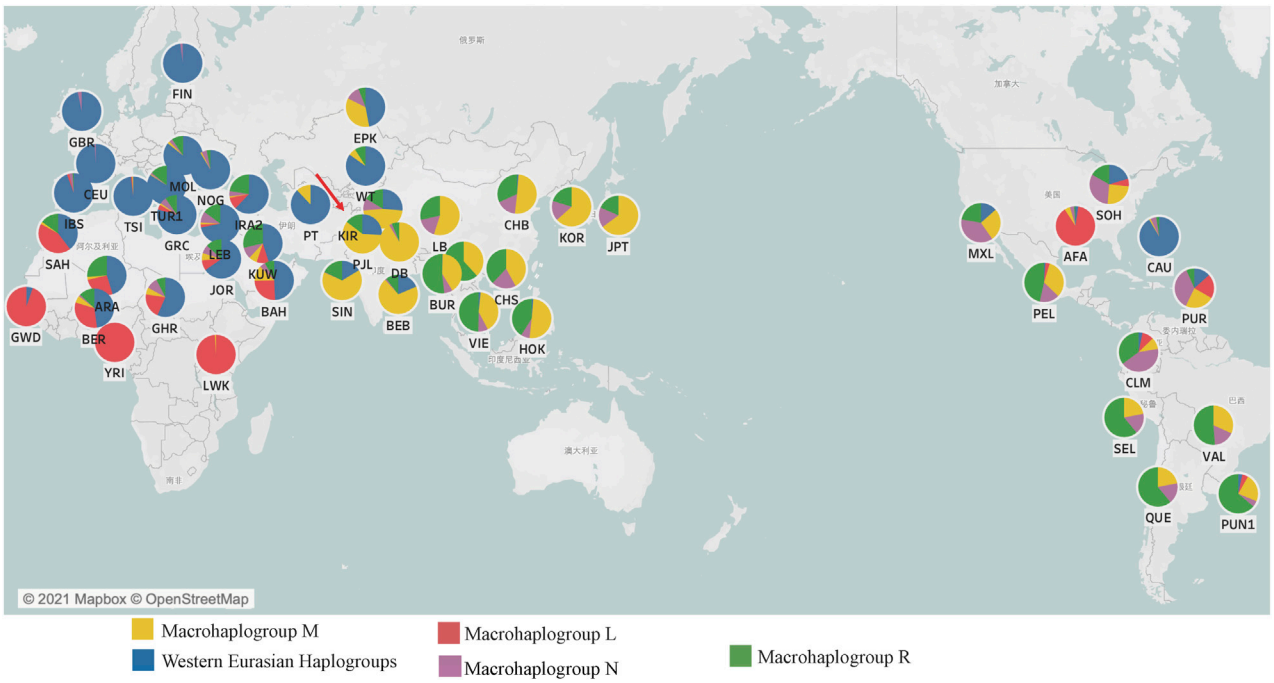

FIGURE 1 | The macrohaplogroup distributions and the geographical coordinates of the Kirgiz group and 48 reference populations around the world. In each pie chart, the red showed the Africa specific macrohaplogroup L, the yellow showed the commonly observed macrohaplogroup M in East Asian populations, the blue indicated the Western Eurasian haplogroup, the purple showed the macrohaplogroup N and the green showed the macrohaplogroup R. Within the macrohaplogroups N and R of this figure, the Western Eurasian specific haplogroups had been excluded.

Qubit™ four Fluorometer (Thermo Fisher Scientific). The mtDNA control region sequences of 91 unrelated individuals and positive control samples (007 and 9947A) were amplified using the Precision ID mtDNA Control Region Panel (Thermo Fisher Scientific). The libraries were prepared with Precision ID DL8 Kit (Thermo Fisher Scientific) on Ion Chef™ System (Thermo Fisher Scientific). The mtDNA control region was amplificated with 14 primer pairs in two primer pools. The libraries were quantified using the Ion Library TaqMan™ Quantitation Kit (Thermo Fisher Scientific) on the 7,500 Real-Time PCR System (Thermo Fisher Scientific). The sequencing templates were prepared using with the Ion S5™ Precision ID Chef & Sequencing Kit (Thermo Fisher Scientific), and were loaded to the Ion 530™ Chip (Thermo Fisher Scientific) on the Ion Chef™ System (Thermo Fisher Scientific). The sequencing runs were accomplished using the Ion S5™ Precision ID Chef & Sequencing Kit (Thermo Fisher Scientific) on Ion S5™ XL System (Thermo Fisher Scientific). All the above experimental procedures were performed strictly according to the manufacturer's recommendations.

Data Analyses

The raw sequencing data were analyzed using Converge ™ Software V2.2 (Thermo Fisher Scientific) of MPS mtDNA module with HID Genotyper 2.2 of default parameters. After aligning to the Revised Cambridge Reference Sequence (rCRS) (Andrews et al., 1999), all the variations were called for further scoring. After the mutations were verified, they were mapped to nodes in the Phylotree 17 (http://www.phylotree.org/tree/index.htm) of human mitochondrial DNA lineages, further allocated the haplotypes to the closest haplogroups, which were submitted to the EMPOP database for quality control (Parson and Dür, 2007). The haplotypes were annotated according to the recommendations of International Society of Forensic Genetics with the nomenclature rules of mtDNA typing (Parson et al., 2014). The analytical threshold at all 1,122 sequences was 20×, and the thresholds of confirmed mutation, point heteroplasmy, insertion and deletion (length heteroplasmy) were 96%, 10%, 20%, and 30% of the total reads with each variant, respectively.

Statistical Analyses

A total of 5,795 individuals from 48 reference populations distributed all over the world were chosen for further investigation of the genetic relationships among the studied Kirgiz ethnic group and the reference populations based on the mtDNA control region sequences. The sample size of each selected population was more than 30 individuals. The 48 reference populations included seven populations from Africa, five populations from North America, six populations from South America, nine populations from Europe, four populations from Central Asia, eight populations from East Asia, four populations from South Asia, two populations from Southeast Asia and four

populations from West Asia. The detailed information of the Kirgiz group and these 48 reference populations was showed in **Supplementary Table S1**.

To evaluate the sequencing performance, we calculated several parameters including the true allelic ratio, noise ratio, ratio of two sequencing directions and the mean depth of each sample, the average depth of each amplicon, and the depth ratio between two primer pools of each sample. The true allele ratios were calculated by dividing the read depth of the true allele into total read depth of each sample, and the remaining undefined reads were defined as noise. The sequencing ratios were calculated by dividing the coverage of 5′ reads into the 3' reads. The mean depth per sample and average depth of each amplicon were also calculated directly. The depth ratios were by dividing the lowest read depth into the greatest read depth of each sample.

The haplogroup frequencies based on the mtDNA control region sequences of 91 Kirgizs were calculated directly. The phylogenetic tree for 91 Kirgiz mtDNA control region sequences was performed using the online tool of HaploGrep 2 v2.0 (https://haplogrep.i-med.ac.at/app/index.html) under the Kulczynski measure. The example case of U1a2 with detailed calculation formula could be found at the website https://haplogrep.i-med.ac.at/2018/06/21/explaining-the-formula/. These mutations including 309.1C(C), 315.1C, 515-522 InDel of AC, A16182C, A16183C, 16193.1C(C), C16519T and T16519C were excluded when the phylogenetic tree was reconstructed, and the tree was showed in **Supplementary Figure S1**.

DNA Sequence Polymorphism v6 (DnaSP v6) (Librado and Rozas, 2009) was used to calculate the statistical parameters of the 91 mtDNA control region sequences, which included the number of polymorphic site (S), haplotype diversity (Hd), nucleotide diversity (Pi), and the average number of nucleotide difference (k). Analysis of molecular variation (AMOVA), neutrality tests (Tajima's D and Fu's Fs tests), and mismatch distribution (the sum of square deviation, SSD, and Harpending's raggedness index, R) were performed with mtDNA control region sequences using Arlequin 3.5.1.3, simultaneously (Excoffier and Lischer, 2010). The Bayesian phylogenetic inference and Bayesian Skyline Plot (BSP) of Kirgiz group were performed using BEAST 2.6.5 software to deduce the time when the Kirgiz group expansion occurred (Bouckaert et al., 2019). The nucleotide substitution model of TN93 and gamma site model with the substitution rate of 1.57E-8 of the single nucleotide substitution per year were set (Soares et al., 2009), and the strict clock model was selected. The pairwise divergence time for *Homo sapiens* and *Homo* neanderthalensis was chosen to calibrate the divergence time, as a median time of 0.55 (±0.054) million years ago (MYA) (Soares et al., 2009). The adaptive Monte Carlo Markov chain (MCMC) approach was used to evaluate the molecular evolution of the mtDNA control region sequences, the MCMC chain length was 1,000,000, and auto optimize displayed for 10,000 steps. Bayesian Skyline Plot (BSP) reconstruction was performed using Tracer 1.7.0 (Rambaut et al., 2018) and the phylogenetic tree was annotated by FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). The detailed geographical coordinates and the macrohaplogroup compositions of all the populations were visualized using Tableau Desktop v 2021.1. Heatmap was constructed using the 'heatmap' package of *R* statistical software based on the pairwise F_{ST} values between the Kirgiz group and 48 reference populations (https://www.r-project.org/). A neighbor-joining (NJ) tree was built using Molecular Evolutionary Genetics Analysis (Mega) v10.2.1 based on the pairwise F_{ST} values among 49 populations. And the NJ tree was displayed and annotated using the online tool iTOL v 6.1.2 (https://itol.embl.de/itol.cgi). The two-dimensional (2D) and three-dimensional (3D) principle component analyses (PCA) were constructed using the dimensionality reduction analysis module of SPSS Statistics v.19.0 based on the haplogroup frequencies generated from mtDNA control regions of the Kirgiz group and 48 reference populations. The haplogroup network calculations were performed using Network v.10.2.0.0 of the median joining module based on the mtDNA control region sequences from the Kirgiz and the 48 reference populations, finally, of which 1,403 individuals from the 48 reference populations sharing the same or similar haplogroups with the target Kirgiz individuals were selected to participate in the network relationship evaluations. The following network plots were visualized with the network publisher (https://www.fluxus-engineering.com/nwpub.htm).

RESULTS

General Performance of the Sequencing Results

All the mtDNA control region sequences from the 91 Kirgiz individuals and the positive control samples (9947A and 007) were successfully generated from four sequencing runs on the Ion S5™ XL system. The true allelic ratios and noise ratios were shown in **Figures 2A,B**. Even when outliers were taken into consideration, the true alleles of all tested samples covered more than 96% of sequencing reads, while the noise was less than 4%. The balance ratios of all the amplicons from two directions in the 91 samples were shown in **Figure 2C**. The balance parameters of most samples were around 0.5, which indicated the relatively good performance of the bi-directional sequencing. In **Figure 2D**, the blue and green bars represented the average depths of two primer pools with 14 amplicons when the control region of mtDNA was amplified using the imbrication strategy. The mean depth per individual (the first red boxplot in **Figure 2D**) was 12989 ± 10865 × (mean ± standard deviation). The mean depths of the first pool ranged from 4,598 × (P1A6) to 24535 × (P1A3) which were shown in the blue boxplot in **Figure 2D**. The mean depths of the second pool ranged from 3,632 × (P2A5) to 33856 × (P2A1), as shown in the green boxplot in **Figure 2D**. The average depth ratio of two pools (the purple boxplot in **Figure 2D**) was 68.65%.

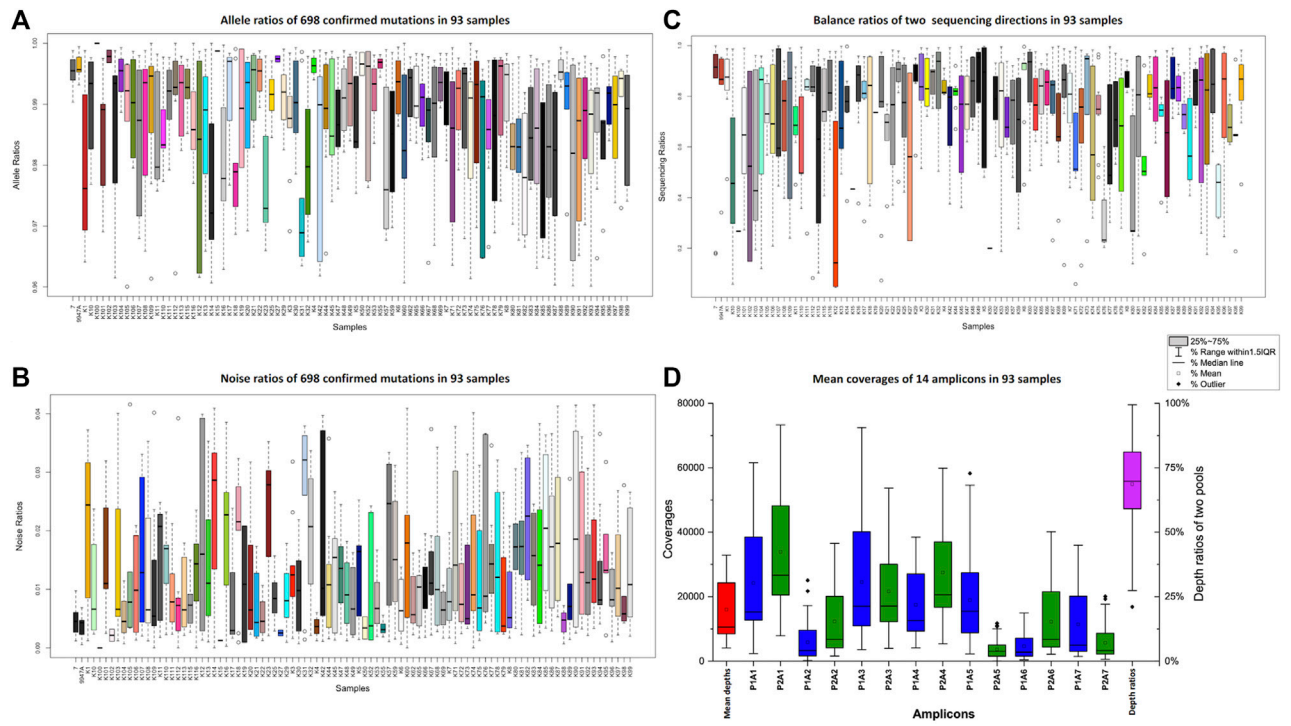

FIGURE 2 | Ratios of true alleles **(A)**, ratios of noises **(B)**, balance ratios of two different sequencing directions **(C)** and the average depths of 14 amplicons **(D)** of the 91 tested Kirgiz individuals and two positive control samples. The 007 and 9947A were showed at the first two boxplots within **Figure 2**. The series of blue and green boxplots in **Figure 2** represented the 14 amplicons, the red box plot showed the average depths of the whole panel in 91 Kirgiz individuals, and the purple one showed the balance ratios of two pools in 91 Kirgiz individuals and two positive control samples.

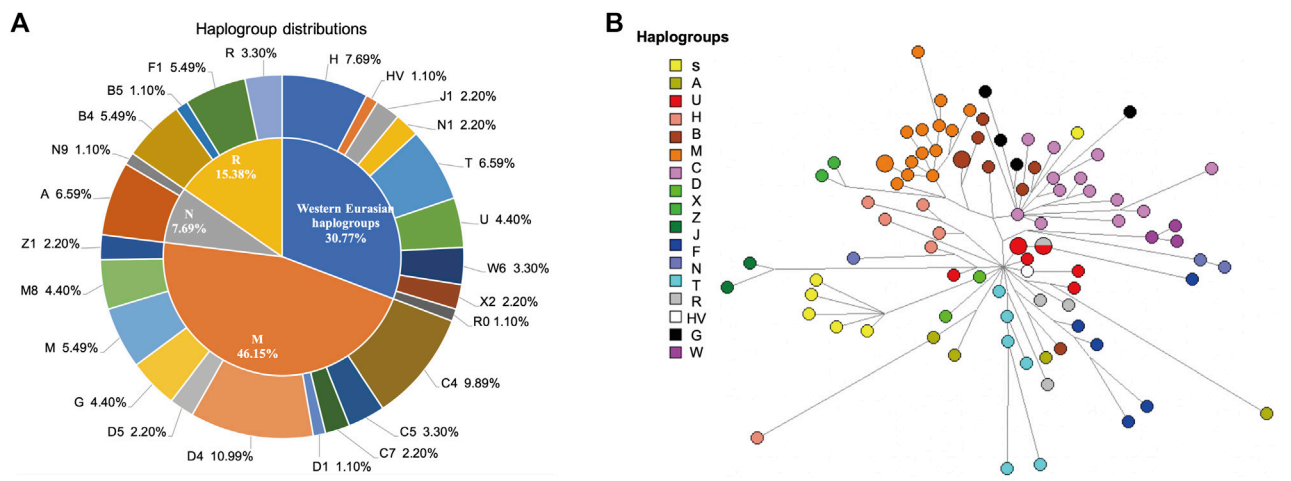

FIGURE 3 | The haplogroup distribution **(A)** and median network **(B)** of the mtDNA control region sequences of 91 Kirgiz individuals. In **(A)**, the macrohaplogroups R, M and N were commonly observed in East Asia, and the Western Eurasian haplogroups were commonly observed in Europe. In **(B)**, the median network analysis among Kirgiz individuals was conducted using the mtDNA control region sequences originated from 91 Kirgiz individuals. Different colors represented different haplogroups and the same haplogroups clustered together and separated with others.

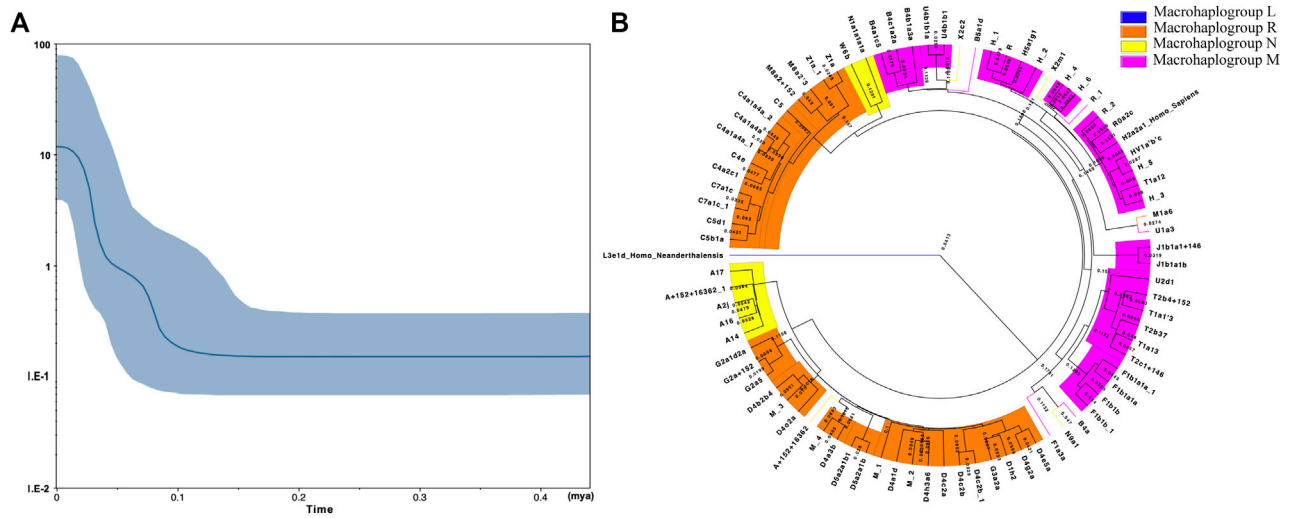

FIGURE 4 | The BSP plot **(A)** and the Bayesian phylogenetic tree **(B)** generated from the mtDNA control region sequences of 91 Kirgiz individuals and two reference sequences (*Homo sapiens* and *Homo neanderthalensis*). In **(A)**, the X - axis represented the expansion time (million years ago for each unit) while the Y - axis showed the assumed effective population size on a logscale. The blue line in the bold represented the median population size. The blue shade showed the boundary of the 95% highest posterior density. The phylogenetic tree shown in **(B)** was reconstructed on basis of the tree file which was generated from the process of expansion time evaluation with BEAST 2.6.5 by the FigTree v1.4.4.

Analyses at the Intra-Population Level

Genetic Variants of mtDNA Control Region Sequences Observed in Kirgiz Group

The phylogenetic tree of mtDNA control region mutations generated from 91 unrelated Kirgiz individuals was performed using the online tool HaploGrep v.2.0 based on the latest mtDNA tree Build 17, and the result was shown in **Supplementary Figure S1**. All samples could be distinguished between each other based on mtDNA control region sequence variations. In total, 168 variants were observed among 91 unrelated Kirgiz individuals including 52 local and three unexpected global mutations (199M, 451G, 502 del), the global mutations had never been reported previously in the EMPOP database. The details of the 168 mutations and the mutation rates were shown in **Supplementary Table S2**. The variants including 263G, 73G, 16223T, 489C and 16189C were the first five most common mutation points identified in Kirgiz group when compared with the rCRS.

Haplogroup Allocations and Forensic Parameters of mtDNA Control Region Sequences in Kirgiz Group

A total of 66 haplogroups and 81 haplotypes were identified from 91 Kirgiz unrelated individuals based on mtDNA control region sequences. The variants (including single point mutation, InDel, and heteroplasmy were marked in green, grey and yellow, respectively), haplotypes and allocated haplogroups of 91 Kirgiz individuals were shown in **Supplementary Table S3**. The mtDNA haplogroup distributions of the Kirgiz group were shown in **Figure 3A**. The haplogroups of 91 Kirgizs comprised of 69.23% East Eurasian haplogroups and 30.77% Western Eurasian haplogroups. The macrohaplogroup M (2.20% of Z1, 4.40% of M8, 5.49% of M, 4.40% of G, 2.20% of D5, 10.99 of D4, 1.10% of D1, 2.20% of C7, 3.30% of C5, and 9.89% of C4), macrohaplogroup R (5.49% of F1, 1.10% of B5, 5.49% of B4, and 3.30% R) and macrohaplogroup N (1.10% of N9, 6.59% of A) commonly observed in East Eurasia populations were detected in the Kirgiz group, respectively. Additionally, the Western Eurasian haplogroups including H (7.69%), HV (1.10%), J1 (2.20%), N1 (2.20%), N1 (2.20%), T (6.59%), U (4.40%), W6 (3.30%), X2 (2.20%), and R0 (1.10%) were also observed in the Kirgiz group.

DNA polymorphisms of the studied Kirgiz group were evaluated, and the results of indexes were presented as follow, number of polymorphic site (S: 107), number of haplotype (h: 81), haplotype diversity (Hd: 0.997), nucleotide diversity (Pi: 0.00842) and the average number of nucleotide difference (k: 9.017), these parameters indicated the relatively high genetic diversities of mtDNA control region sequences was in Kirgiz group. The Tajima's D test (-1.950, p-value < 0.05) and Fu's FS test (-25.247, p-value < 0.05) of mtDNA variants for the 91 Kirgiz individuals were both calculated to be relatively large negative values, and the results might indicate that the variants were significantly deviated from neutral mutations.

Population Expansion of the Kirgiz Group

The mismatch distribution graph shown in **Supplementary Figure S2** revealed a single peak, which indicated the Kirgiz group might occur the population expansion event according to the previous research method (Slatkin and Hudson, 1991). The SSD and Raggedness indexes of two models in the Kirgiz group were 0.00182 (SSD, p-value = 0.60000), 0.00689

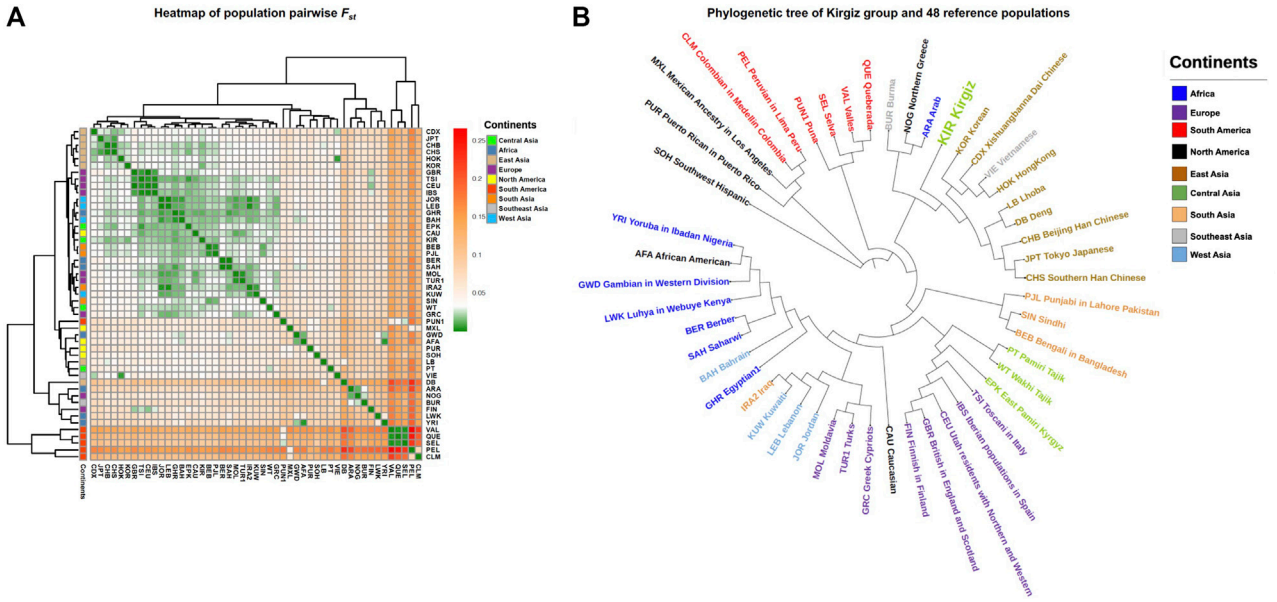

FIGURE 5 | The F_{ST} genetic distances and genetic relationships between Kirgiz group and other 48 reference populations. All the populations of different continents were labelled by different colors. **(A)** represented the heatmap of pairwise F_{ST} values between Kirgiz group and the 48 reference populations. F_{ST} colors were from green, white, orange to red in turn, the corresponding values changed from small to large. **(B)** showed the phylogenetic tree of Kirgiz group and the reference populations based on pairwise F_{ST} values.

(Raggedness index, p-value = 0.73000) for the sudden expansion model, 0.00216 (SSD, p-value = 0.57000) and 0.00689 (Raggedness index, p-value = 0.86000) for the spatial expansion model, respectively. The median network calculation was performed based on the mtDNA control region sequences of the 91 Kirgiz individuals, and the network plot of 91 Kirgiz individuals was shown in **Figure 3B**. Small branches belonged to the same haplogroup clustered together and separated from the other different haplogroup branches. Furthermore, as shown in **Figure 4**, Bayesian Skyline Plot (BSP) analysis was conducted based on the ancestral theory to quantify the evolutionary background of population size and history. The expansion time with the largest slope of the BSP abscissa was directly read. And the BSP abscissa showed that the Kirgiz group had grown at about *53.41 kya* (**Figure 4A**, the abscissa corresponding to the dash line). The annotated phylogenetic tree (**Figure 4B**) showed the estimated divergence time between *Homo sapiens* and *Homo* neanderthalensis was 0.5413 MYG, which was very close to the expected time of 0.55 MYG.

Genetic Relationship Analyses Among Kirgiz Group and 48 Reference Populations

As for an important genetic marker for population evolution analysis and ancestral information inference, mtDNA control region sequences originated from 91 Kirgiz individuals were combined with other 5,795 mtDNA control region sequences from those individuals of 48 reference populations, which were used to analyze genetic relationships between Kirgiz group and these reference populations. The detailed geographical coordinates and the macrohaplogroup compositions of the Kirgiz group and 48 reference populations were shown in **Figure 1** and **Supplementary Table S4**. As shown in **Figure 1**, with the advancement towards Africa, Europe, Asia, and America, the proportion of African-specific macrohaplogroup L gradually decreased and disappeared, and the proportions of macrohaplogroups M, N and R gradually increased. In macrohaplogroups N and R in this figure, the Western Eurasian specific haplogroups had been excluded. The 92.77% haplogroups of Caucasian population from North America (CAU) belonged to the Western Eurasian haplogroup because of the European origin.

Genetic Distance Analyses Based on the Haplogroup Frequencies of mtDNA Control Region Sequences

Pairwise F_{ST} values were calculated based on the haplogroup frequencies of mtDNA control region sequences to obtain insight into the genetic affinities of the Kirgiz group and 48 reference populations. A population clustering heatmap of the pairwise F_{ST} values was plotted and shown in **Figure 5A**. In addition, the NJ tree of these 49 populations was displayed in **Figure 5B**. In the NJ tree, the Kirgiz group gathered together with other East Asia populations and formed a larger cluster. As a whole, the NJ tree indicated that most geographically adjacent populations gathered together. However, some populations were outliers which might be due to the population histories or the deviations caused by the

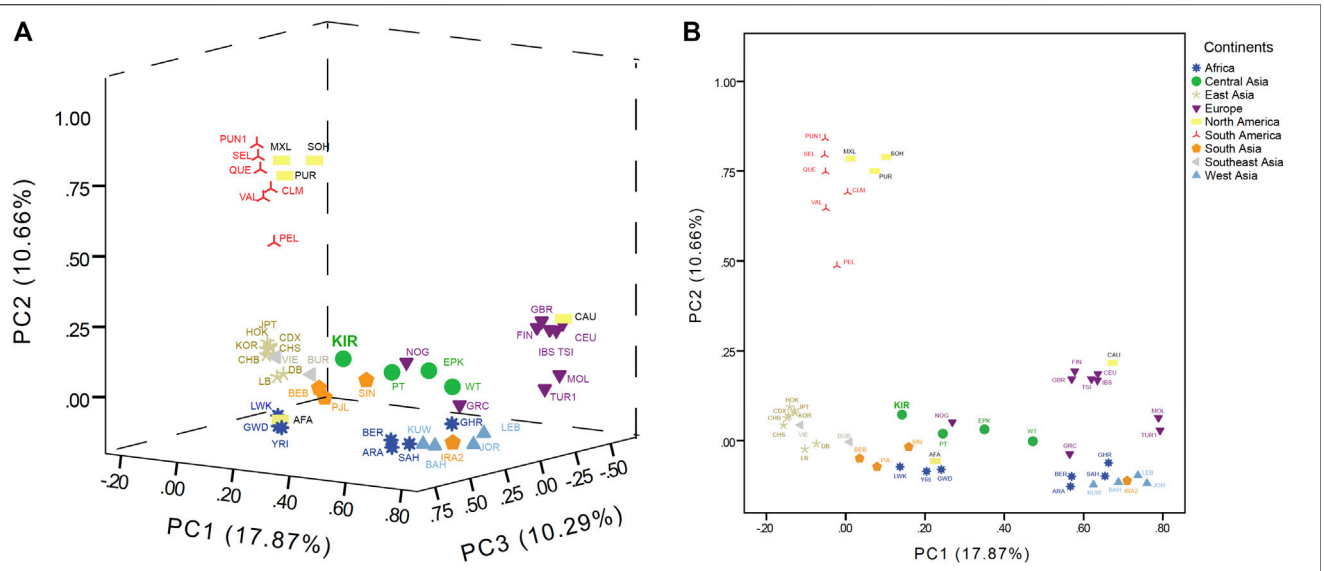

FIGURE 6 | The 3D (PC1, PC2, and PC3) in **(A)** and 2D (PC1 and PC2) in **(B)** of PCA plots among the Kirgiz and the 48 reference populations.

different methodologies. The pairwise F_{ST} values were shown in **Supplementary Table S6**. The pairwise F_{ST} values between the Kirgiz group and other reference populations ranged from 0.01142 to 0.1499, and the median was 0.02836. The largest three F_{ST} values with the Kirgiz group were identified in three populations including Peruvian in *Lima* (PEL), Valles (VAL) and Colombian in Medellin (CLM). Among Chinese nine populations, the smallest three F_{ST} values were observed between Kirgiz and Beijing Han Chinese (CHB), East Pamiri Kyrgyz (EPK) and Southern Han Chinese (CHS).

Principal Component Analyses Based on Haplogroup Frequencies

PCA was conducted to evaluate the genetic affinities among Kirgiz group and reference populations based on the haplogroup frequencies in terms of maternal inheritance, and the haplogroup distributions of Kirgiz group and 48 reference populations were shown in **Supplementary Table S5**. And the scree PCA plot was shown in **Supplementary Figure S3**, which represented the principal components and percentages of explained variances. The first three principal components could explain 38.82% of the total variation among these 49 populations, while the first principal component explained 17.87% variation, the second explained 10.66%, and the third explained 10.29%. Based on the first three principal components, a 3D-PCA (PC1, PC2 and PC3) and a 2D-PCA (PC1 and PC2) were constructed and presented in **Figure 6**. The populations from nine different continental regions were marked with nine different colors. The PCA plots in **Supplementary Figure S4** showed the 2D-PCA plots (PC1 and PC3, PC2 and PC3). As displayed in the PCA plots, the East Asia populations could be separated in PC1, the American populations could be separated in PC2, and the African populations could be separated in PC3.

Haplogroup Median Network Analyses of the Observed Haplogroups in Kirgiz Group

To evaluate the network relationships between Kirgiz and other reference populations, all the 91 Kirgiz mtDNA control region sequences were compared with the individual sequences from 48 reference populations, and six median network graphs were finally generated. In total, 460 different individual sequences of macrohaplogroup M were distributed into seven different haplogroups, as 62 of D4, 36 of M1, 216 of M, 44 of G, 63 of M8, eight of D1, and 31 of D5, respectively. The median network diagrams of macrohaplogroup M with the largest proportion of Kirgiz group were labeled by different colors for different geographical locations, and different haplogroups were shown in **Figure 7** and **Supplementary Figure S5**, respectively. In **Figure 7**, the red arrows indicated the 42 Kirgiz individual sequences were allocated into macrohaplogroup M. There were 58 individuals from eight populations from East Asia, one population from South Asia, two populations from Southeast Asia, one population from Europe, three populations from South America and two populations from North America which had homogenous haplogroup M with three Kirgiz individuals. Two East Asia populations were observed the direct link with Kirgiz group. Haplotype sequences apart from one median vector were observed between the Kirgiz individuals and one Southeast Asian, three East Asian individuals.

The commonly observed Western Eurasian haplogroup including the H, T, U, W, X2, J1, N1, HV and R0 was the second-largest haplogroup in the Kirgiz group. These 27 Kirgiz individuals were allocated into this category. The Western Eurasian haplogroup of 27 Kirgiz individuals was allocated into nine above-mentioned haplogroups and nine different network plots labeled with colors for different continents (**Figure 8**) and different haplogroups (**Supplementary Figure S6**), respectively. There were 582 mtDNA

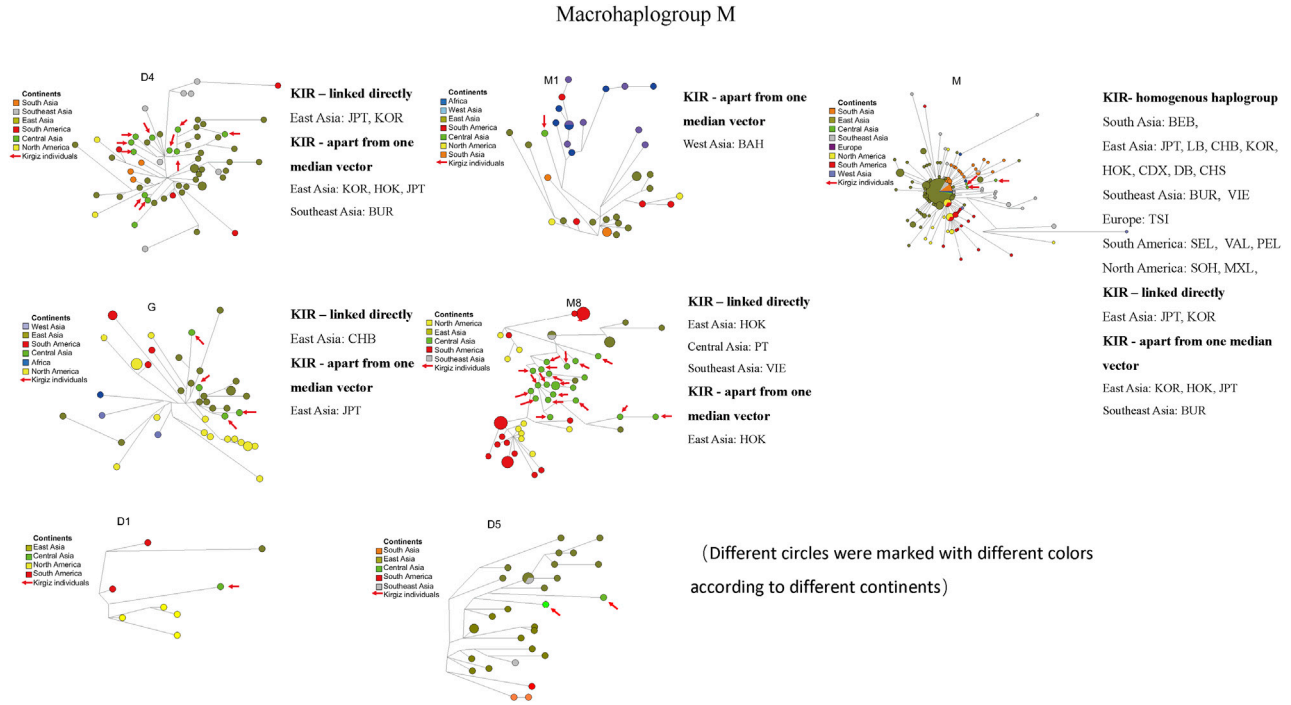

FIGURE 7 | Median network of macrohaplogroup M. Different continents were labelled by different colors. There were 460 individual sequences of macrohaplogroup M which were allocated into seven haplogroups (D4, M1, M, G, M8, D1 and D5). The red arrows represented the haplotype sequences of the 42 Kirgiz individuals. KIR was the abbreviation of Kirgiz group.

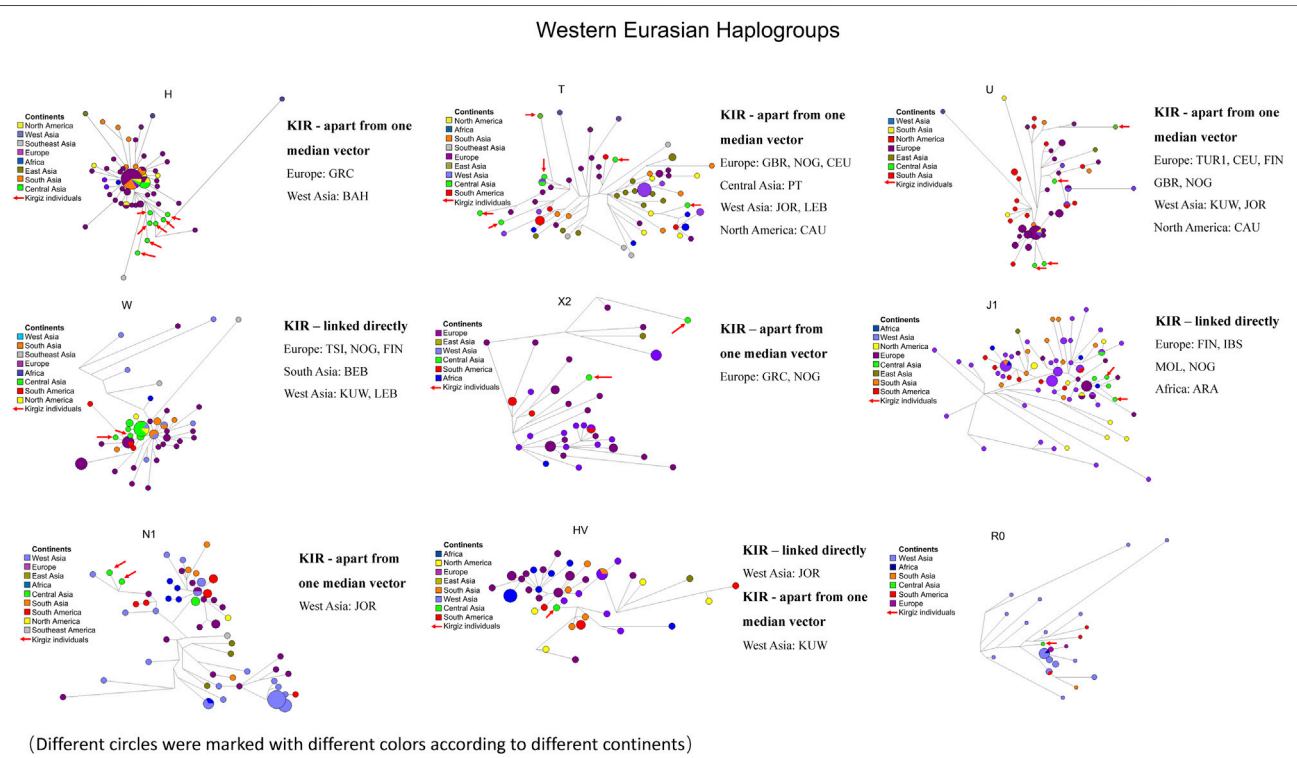

FIGURE 8 | Median network of Western Eurasian haplogroup. Different continents were labelled by different colors. There were 582 individual sequences of Western Eurasian haplogroups which were allocated into nine haplogroups (H, T, U, W, X2, J1, N1, HV and R0). The red arrows represented the haplotype sequences of 28 Kirgiz individuals. KIR was the abbreviation of Kirgiz group.

control region sequences including in these nine median network relationships: 80 of H, 77 of T, 65 of U, 64 of W, 42 of X2, 97 of J1, 74 of N1, 47 of HV, and 36 of R0. Compared with the Kirgiz group, the directly linked individuals were observed in five European populations including Toscani in Italy (TSI), Northern Greece (NOG), Finnish in Finland (FIN), Iberian populations in Spain (IBS) and Moldavia (MOL), two West Asia populations i.e. Kuwaiti (KUW) and Jordan (JOR), one South Asia group Bengali in Bangladesh (BEB), one South America population Lebanon (LEB) and one African population Arab (ARA). Individuals apart from one median vector with Kirgizs were mostly observed in European populations, followed by the Central Asia and West Asia populations.

The 14 individuals of Kirgiz group were allocated into the macrohaplogroup R. Combining with 263 individuals from reference populations, 277 mtDNA control region sequences were distributed into four different haplogroups. And these four different network plots were shown in **Supplementary Figure S7**. These four different networks were marked by different colors for different continents in **Supplementary Figure S7A**, whereas, the different haplogroups were labeled by different colors in **Supplementary Figure S7B**, respectively. The median networks of four haplogroups B4, B5, R, and F1 were reconstructed based on the 91, 82, 70, and 20 mtDNA control region sequences among 48 populations, respectively. One individual of Utah resident with Northern and Western European ancestry (CEU) shared the homogenous haplogroup with a Kirgiz individual. Some individuals from eight populations including seven populations of Asia and one of Europe linked directly with 14 Kirgiz individuals, respectively. Seven individuals from three South Asian populations and one Southeast Asian, one North American, one South American, and one African population were found to be apart from one median vector with 14 Kirgiz individuals, respectively.

The smallest macrohaplogroup proportion of 91 unrelated Kirgiz individuals was N macrohaplogroup. The 25 sequences from the reference populations and seven sequences from the Kirgiz group were used to draw the median network diagram of the N macrohaplogroup. And the diagram of each haplogroup was labeled by different continents and haplogroups with different colors (**Supplementary Figure S8**). Within the macrohaplogroup N, seven Kirgiz individuals clustered together and separated from the individuals of other reference populations.

DISCUSSION

In this research, mtDNA control region sequences were detected from 91 unrelated Kirgiz individuals to analyzed the genetic background of Kirgiz group located in the Chinese Northwest region, and laid the foundation for the mtDNA control region sequences in forensic practice applications and enriched the genetic information database of the Chinese populations. As a result, the Kirgiz group displayed relatively high genetic polymorphisms with the parameters including the number of haplotypes (h: 81), haplotype and nucleotide diversities (Hd:0.997 and Pi: 0.00842), and the average number of nucleotide differences (k:9.017) as Zimmermann said (Zimmermann et al., 2019). The 91 unrelated individuals can be separated from each other based on the mtDNA control region sequences, which might indicate the effectiveness of the 1122bp control region sequences for forensic maternal lineage analyses according to the recommendation of previous study (Cardoso et al., 2013). The neutrality test, mismatch distribution, median network analyses, and BSP inference are the most commonly used methods to evaluate the population historical dynamics and the population expansion time in the biogeographic analysis. The results of Tajima's D test (-1.950, p-value < 0.05) and Fu's FS test (-25.247, p-value < 0.05) indicated that the Kirgiz group experienced population expansion during its history as the previous study recommended (Ramos-Onsins and Rozas, 2002). The small and insignificant r values (raggedness index, 0.00689 of sudden expansion model and 0.00689 of spatial expansion model) and SSD indexes (0.00182 of sudden expansion model and 0.00216 of spatial expansion model) of the appropriateness test of the mismatch distribution also supported the hypothesis of population expansion of the Kirgiz group (Slatkin and Hudson, 1991). The expansion of the Kirgiz group occurred at about *53.41 kya* based on the curve of the BSP when the ancestries of early modern Kirgiz group began to expand. At that time, the individual fossils representing modern Asians and modern Europeans were discovered (Liu et al., 2021). Additionally, it was difficult to achieve smaller scale estimation of the expansion time to track smaller population expansion event because of the limited genetic information.

The population pairwise F_{ST} values were calculated based on haplogroup frequencies. And the F_{ST} heatmap showed the genetic differentiations between Kirgiz group and the reference populations. In the NJ phylogenetic tree, most of populations from the same continent clustered together except the immigrant populations like African America and Caucasians from North America. In addition, the Kirgiz group had the three smallest F_{ST} values with CHB, EPK and CHS populations from China, which indicated relatively close genetic affinities between Kirgiz and these three populations according to the criteria of genetic differentiations of Wright (Stoneking and Delfin, 2010). The largest three F_{ST} values were observed between Kirgiz group and PEL, VAL and CLM populations, which might indicate high differentiations and relatively remote genetic distances between Kirgiz group and these populations. What's more, the previous study indicated that the Kirgiz was a mixed Eurasian group (Fahu Chen et al., 2019), and the present results of the coefficient analysis of genetic differentiations also showed that the Kirgiz was close to the East Asia populations. Subsequent median network analyses based on the mtDNA control region sequences also showed the similar results mentioned above. No matter considering the homologous haplogroup, directly connection or apart from one median vector, the East Aisa individuals had the most

individuals who had close genetic affinities with the Kirgiz individuals. Several previously reported studies based on autosomal InDel markers and Y-STRs and SNPs (Guo et al., 2020; Song et al., 2021; Zhang et al., 2021) also supported that the Kirgiz had the close genetic relationships with the East Asia populations.

In summary, we provided 91 high quality mtDNA control region sequences of the Kirgiz ethnic group located in Northwest China. Fifty-two local mutations and three unexpected global mutations (199M, 451G, 502del) were identified among these mtDNA control region sequences of Kirgiz individuals. In addition, we evaluated the mtDNA genetic diversities in the Kirgiz group and gained insight into its genetic relationships with 48 reference populations all over the world. This study on the genetic diversities of mtDNA control region sequences in Kirgiz ethnic group might promote the understanding the genetic background of Kirgiz ethnic group and lay the foundation for the researches of maternal pedigrees, and promote the case investigations involving matrilineal families.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Xi'an Jiaotong University Health Science Center. The patients/participants provided their written informed consents to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

BZ designed the study, obtained the funds and revised the manuscript. HW and YF conducted the MPS experiments including sample DNA extractions, library and template preparations, and final MPS genotyping. HW and MC accomplished the data analyses and prepared the manuscript with equal contributions. CC, WC, and FL helped to perform the BSP inference and median network analyses. All the authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant numbers: 81930055 and 81772031).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.729514/full#supplementary-material

REFERENCES

Abramzon, S. M., and Tabyshaliev, S. T. (1990). The Kirgiz and Their Ethnogenetical Historical and Cultural Connections[J]. *Moscow* 1971, 30–81.

Amorim, A., Fernandes, T., and Taveira, N. (2019). Mitochondrial DNA in Human Identification: a Review. *PeerJ* 7, e7314. doi:10.7717/peerj.7314

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., et al. (1981). Sequence and Organization of the Human Mitochondrial Genome. *Nature* 290 (5806), 457–465. doi:10.1038/290457a0

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., and Howell, N. (1999). Reanalysis and Revision of the Cambridge Reference Sequence for Human Mitochondrial DNA. *Nat. Genet.* 23 (2), 147. doi:10.1038/13779

Bandelt, H.-J., and Dür, A. (2007). Translating DNA Data Tables into Quasi-Median Networks for Parsimony Analysis and Error Detection. *Mol. Phylogenet. Evol.* 42 (1), 256–271. doi:10.1016/j.ympev.2006.07.013

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., et al. (2019). BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis. *Plos Comput. Biol.* 15 (4), e1006650. doi:10.1371/journal.pcbi.1006650

Cardoso, S., Palencia-Madrid, L., Valverde, L., Alfonso-Sánchez, M. A., Gómez-Pérez, L., Alfaro, E., et al. (2013). Mitochondrial DNA Control Region Data Reveal High Prevalence of Native American Lineages in Jujuy Province, NW Argentina. *Forensic Sci. Int. Genet.* 7 (3), e52–e55. doi:10.1016/j.fsigen.2013.01.007

Excoffier, L., and Lischer, H. E. L. (2010). Arlequin Suite Ver 3.5: a New Series of Programs to Perform Population Genetics Analyses under Linux and Windows. *Mol. Ecol. Resour.* 10 (3), 564–567. doi:10.1111/j.1755-0998.2010.02847.x

Chen, F., Welker, F., Shen, C.-C., Bailey, S. E., Bergmann, I., Davis, S., et al. (2019). A Late Middle Pleistocene Denisovan Mandible from the Tibetan Plateau. *Nature* 569 (7756), 409–412. doi:10.1038/s41586-019-1139-x

Gallimore, J. M., McElhoe, J. A., and Holland, M. M. (2018). Assessing Heteroplasmic Variant Drift in the mtDNA Control Region of Human Hairs Using an MPS Approach. *Forensic Sci. Int. Genet.* 32, 7–17. doi:10.1016/j.fsigen.2017.09.013

Gordon, M. (2009). The Turks in World History (Review). *J. World Hist.* 20 (1), 151–153. doi:10.1353/jwh.0.0035

Guo, Y., Chen, C., Jin, X., Cui, W., Wei, Y., Wang, H., et al. (2018). Autosomal DIPs for Population Genetic Structure and Differentiation Analyses of Chinese Xinjiang Kyrgyz Ethnic Group. *Sci. Rep.* 8 (1), 11054. doi:10.1038/s41598-018-29010-8

Guo, Y., Xia, Z., Cui, W., Chen, C., Jin, X., and Zhu, B. (2020). Joint Genetic Analyses of Mitochondrial and Y-Chromosome Molecular Markers for a Population from Northwest China. *Genes* 11 (5), 564. doi:10.3390/genes11050564

Librado, P., and Rozas, J. (2009). DnaSP V5: a Software for Comprehensive Analysis of DNA Polymorphism Data. *Bioinformatics* 25 (11), 1451–1452. doi:10.1093/bioinformatics/btp187

Liu, Y., Mao, X., Krause, J., and Fu, Q. (2021). Insights into Human History from the First Decade of Ancient Human Genomics. *Science* 373 (6562), 1479–1484. doi:10.1126/science.abi8202

Parson, W., and Dür, A. (2007). EMPOP-A Forensic mtDNA Database. *Forensic Sci. Int. Genet.* 1 (2), 88–92. doi:10.1016/j.fsigen.2007.01.018

Parson, W., Gusmão, L., Hares, D. R., Irwin, J. A., Mayr, W. R., Morling, N., et al. (2014). DNA Commission of the International Society for Forensic Genetics: Revised and Extended Guidelines for Mitochondrial DNA Typing. *Forensic Sci. Int. Genet.* 13, 134–142. doi:10.1016/j.fsigen.2014.07.010

Chen, P., Zou, X., Wang, B., Wang, M., and He, G. (2019). Genetic Admixture History and Forensic Characteristics of Turkic-Speaking Kyrgyz Population via 23 Autosomal STRs. *Ann. Hum. Biol.* 46 (6), 498–501. doi:10.1080/03014460.2019.1676918

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* 67 (5), 901–904. doi:10.1093/sysbio/syy032

Ramos-Onsins, S. E., and Rozas, J. (2002). Statistical Properties of New Neutrality Tests against Population Growth. *Mol. Biol. Evol.* 19 (12), 2092–2100. doi:10.1093/oxfordjournals.molbev.a004034

Scally, A. (2016). The Mutation Rate in Human Evolution and Demographic Inference. *Curr. Opin. Genet. Dev.* 41, 36–43. doi:10.1016/j.gde.2016.07.008

Slatkin, M., and Hudson, R. R. (1991). Pairwise Comparisons of Mitochondrial DNA Sequences in Stable and Exponentially Growing Populations. *Genetics* 129 (2), 555–562. doi:10.1093/genetics/129.2.555

Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Röhl, A., et al. (2009). Correcting for Purifying Selection: an Improved Human Mitochondrial Molecular Clock. *Am. J. Hum. Genet.* 84 (6), 740–759. doi:10.1016/j.ajhg.2009.05.001

Song, F., Song, M., Luo, H., Xie, M., Wang, X., Dai, H., et al. (2021). Paternal Genetic Structure of Kyrgyz Ethnic Group in China Revealed by High-resolution Y-chromosome STRs and SNPs. *Electrophoresis* 42 (19), 1892–1899. doi:10.1002/elps.202100142

Stoneking, M., and Delfin, F. (2010). The Human Genetic History of East Asia: Weaving a Complex Tapestry. *Curr. Biol.* 20 (4), R188–R193. doi:10.1016/j.cub.2009.11.052

Strobl, C., Churchill Cihlar, J., Lagacé, R., Wootton, S., Roth, C., Huber, N., et al. (2019). Evaluation of Mitogenome Sequence Concordance, Heteroplasmy Detection, and Haplogrouping in a Worldwide Lineage Study Using the Precision ID mtDNA Whole Genome Panel. *Forensic Sci. Int. Genet.* 42, 244–251. doi:10.1016/j.fsigen.2019.07.013

Ta, M. T. A., Nguyen, N. N., Tran, D. M., Nguyen, T. H., Vu, T. A., Le, D. T., et al. (2021). Massively Parallel Sequencing of Human Skeletal Remains in Vietnam Using the Precision ID mtDNA Control Region Panel on the Ion S5 System. *Int. J. Leg. Med.* 135 (6), 2285–2294. doi:10.1007/s00414-021-02649-1

Wang, H.-D., Jin, X.-Y., Guo, Y.-X., Zhang, Q., Zhang, Y.-W., Wang, X., et al. (2019). KIR Gene Presence/absence Polymorphisms and Global Diversity in the Kirgiz Ethnic Minority and Populations Distributed Worldwide. *Mol. Biol. Rep.* 46 (1), 1043–1055. doi:10.1007/s11033-018-4563-3

World Medical Association (2013). World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *Jama* 310 (20), 2191–2194. doi:10.1001/jama.2013.281053

Xie, M., Li, Y., Wu, J., Song, F., Lu, Y., Liao, L., et al. (2020). Genetic Structure and Forensic Characteristics of the Kyrgyz Population from Kizilsu Kirghiz Autonomous Prefecture Based on Autosomal DIPs. *Int. J. Leg. Med.* doi:10.1007/s00414-020-02277-1

Zhang, W., Jin, X., Wang, Y., Chen, C., and Zhu, B. (2021). Genetic Structure Analyses and Ancestral Information Inference of the Chinese Kyrgyz Group via a Panel of 39 AIM-DIPs. *Genomics* 113 (4), 2056–2064. doi:10.1016/j.ygeno.2021.03.008

Zimmermann, B., Sturk-Andreaggi, K., Huber, N., Xavier, C., Saunier, J., Tahir, M., et al. (2019). Mitochondrial DNA Control Region Variation in Lebanon, Jordan, and Bahrain. *Forensic Sci. Int. Genet.* 42, 99–102. doi:10.1016/j.fsigen.2019.06.020

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership