GENETIC HISTORY OF HUMAN POPULATIONS ALONG THE ANCIENT SILK ROAD, 2nd Edition

EDITED BY: Shaoqing Wen, Horolma Pamjav and Maxat Zhabagin

PUBLISHED IN: Frontiers in Genetics and Frontiers in Ecology and Evolution

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-8325-4654-3
DOI 10.3389/978-2-8325-4654-3

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

GENETIC HISTORY OF HUMAN POPULATIONS ALONG THE ANCIENT SILK ROAD, 2nd Edition

Topic Editors:
Shaoqing Wen, Fudan University, China
Horolma Pamjav, Hungarian Institute for Forensic Sciences, Ministry of Interior, Hungary
Maxat Zhabagin, National Center for Biotechnology, Kazakhstan

Publisher's note: This is a 2nd edition due to an article retraction.

Citation: Wen, S., Pamjav, H., Zhabagin, M., eds. (2024). Genetic History of Human Populations Along the Ancient Silk Road, 2nd Edition. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4654-3

Table of Contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 24 January 2023
DOI 10.3389/fgene.2023.1130104

OPEN ACCESS

EDITED BY
Chuan-Chao Wang,
Xiamen University, China

REVIEWED BY
Lingxiang Wang,
Fudan University, China
Menghan Zhang,
Fudan University, China

*CORRESPONDENCE
Shaoqing Wen,
wenshaoqing1982@gmail.com

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

RECEIVED 22 December 2022
ACCEPTED 16 January 2023
PUBLISHED 24 January 2023

CITATION
Chang X, Pamjav H, Zhabagin M and Wen S (2023), Editorial: The genetic history of human populations along the ancient silk road.
Front. Genet. 14:1130104.
doi: 10.3389/fgene.2023.1130104

Editorial: The genetic history of human populations along the ancient silk road

Xin Chang[1], Horolma Pamjav[2], Maxat Zhabagin[3] and Shaoqing Wen[1,4,5]*

[1]Institute of Archaeological Science, Fudan University, Shanghai, China, [2]Hungarian Institute for Forensic Sciences, Institute of Forensic Genetics, Budapest, Hungary, [3]National Center for Biotechnology, Astana, Kazakhstan, [4]MOE Laboratory for National Development and Intelligent Governance, Fudan University, Shanghai, China, [5]Center for the Belt and Road Archaeology and Ancient Civilizations, Fudan University, Shanghai, China

KEYWORDS

Silk Road, STRs (short tandem repeats), SNPs (single nucleotide polymorphism), ancient DNA (aDNA), DNA sequencing, DNA genotyping, mtDNA, Y chromosomal DNA

Editorial on the Research Topic
The genetic history of human populations along the ancient silk road

Introduction

The Silk Road, a historical network of interlinking trade routes across the Afro-Eurasian landmass, was of great importance to the transport of peoples, goods, and ideas between the East and the West. Although its main use was for importing silk from China, traders moving in the opposite direction carried to Central China jewelry, glassware, and other exotic goods from the Mediterranean, jade from Khotan, and horses and furs from the nomads of the Steppe. In historical records, communication between China and Central Asia has been unbroken ever since the opening of the Silk Road in the Han Dynasty. However, relics unearthed from archaeological sites indicate that communication between people along the Silk Road began during the Bronze Age. The Silk Road brought together the achievements of the different peoples of Eurasia to advance the Old World as a whole.

Ethnic groups with different religions, cultures and customs inhabited the Ancient Silk Road and experienced complex histories. Patterns in genetic variation between individuals can tell us about the population history of these groups. In recent years, using relatively direct means of studying ancient samples through osseous material, alongside indirect means of analyzing the genomes of modern populations, demographic history—migrations, expansions and colonizing events - have been progressively revealed in numerous genetic studies of early human populations. However, until the present, the origins of the populations along the Ancient Silk Road and relationships have been examined in far less detail.

In this special editorial, we collected 15 genetic investigations involving the populations living along or related to the Silk Road from ancient times to the present day. These studies approach academic and public Research Topic (population origins, differentiation, and admixture) of the targeted populations through genome-wide

sequencing (Kairov et al.; Allen et al.; Guarino-Vignon et al.) or microarray technologies (Ma et al.; Wang et al.; Zhang et al.), or various kinds of markers, including mtDNA (Ning et al.; Xue et al.; Xiong et al.; Xiong et al.; Cardinali et al.), Y chromosome (Chen et al.; Khussainova et al.), forensic STRs (Chen et al.; Adnan et al.; Khussainova et al.), SNPs (Chen et al.; Ma et al.; Wang et al.; Zhang et al.; Xiong et al.), and InDels data (Fan et al.) For example, a novel 6-dye direct and multiplex PCR-CE-based typing system has now been validated and could be considered as a reliable tool for human identification and intercontinental population differentiation (Fan et al.)

Eastern end

The eastern end of the Silk Road lies on the middle reaches of the Yellow River. The Shimao site in Shaanxi Province is an important Neolithic archaeological culture in this area. To further reconstruct the genetic structure of Silk Road-related populations, it is necessary to understand the genetic compositions of such early local populations as these. The Shimao population showed a mostly local origin and showed a maternal affinity with Taosi site, a Longshan culture population in the middle Yellow River valley (Xue et al.).

The Hexi corridor connected the Central Plains with the Western Regions (present-day Xinjiang) and was an integral component of the eastern section of Silk Road. Heishuiguo site and Foyemiao site lie in the central and western portion of the Hexi corridor, respectively (Xiong et al.; Xiong et al.). The former site dates back to the Han Dynasty (118BCE-191CE) and the latter to the Wei - Jin Dynasties and Sui and Tang Dynasties (220CE- 907CE). For the two sites, from the paternal Y-chromosome perspective, all male individuals showed the Sino-Tibetan speaking origin of Yellow River-related populations, while Foyemiao samples demonstrated a higher proportion of an Altaic speaking and North Eurasian ancestral component, alongside a small proportion of southern East Asian ancestry. From the maternal perspective, female Heishuiguo individuals showed a northeast Asian origin and revealed a sex-biased migration from the middle and lower reaches of Yellow River to Hexi corridor. This was consistent with evidence in historical records, especially unearthed slips (Ge et al., 1997; Wen et al., 2004a). The female Foyemiao samples had a similar genetic structure with males. The genetic difference between Heishuiguo and Foyemiao populations reflected the distinct genetic diversity of the spatial-temporal Hexi Corridor. Furthermore, combined with multidisciplinary evidence, these two studies demonstrated the impact of sex-biased migration on the Hexi Corridor, providing a reference for other fields.

The Western regions

Population admixture also can be observed in the Western Regions (present Xinjiang), where individuals at the Xiabandi reveal genetic flow from Central Asian populations and thus the influence of Middle and Late Bronze Age steppe pastoral cultures (Ning et al.) Another site, Shichengzi, a Han Dynasty agricultural garrison, also located in ancient Xinjiang, has revealed to us genetic makeup of a Frontier population in early imperial China (Allen et al.). Archaeogenomics at Shichengzi has revealed two subgroups with East Asian origin and Northeast Asian origin, respectively, occupying a single burial space. Interestingly, stable isotope analysis showed that dietary patterns among site inhabitants could be split among agro-pastoral and agricultural groups. Considering ancient DNA and stable isotope evidence together, it has been argued that Northeast Asian origins of Altaic pastoralists and East Asian origins of Han agriculturalists lived together in the Shichengzi military outpost.

The northern steppe

The Altaic language family is divided into Mongolic, Tungusic and Turkic language groups. Altaic speaking nomads played an important role in shaping the northern steppe Silk Road. However, their history is often only recorded sporadically in ancient writings of surrounding civilizations. In this special edition, the genetic history of Uyghur (Adnan et al.), Kazakh (Kairov et al.; Adnan et al.; Khussainova et al.), Hui (Adnan et al.; Chen et al.; Ma et al.), Dongxiang (Ma et al.), Bonan (Ma et al.), Yugur (Ma et al.), Salar (Kairov et al.; Ma et al.), Mongol (Cardinali et al.), and Manchu (Zhang et al.) groups have been investigated using genetic markers and/or genome sequencing at the population level. Generally speaking, population from the same language group exhibit a closer genetic affinity. Furthermore, the Turkic-speaking population from China and Kazakhstan both present a closer genetic relationship to Central Asian populations, while Hui and Tungusic-speaking populations had a significant admixture with Han Chinese. Mongolian' mitochondrial genomes show a dominant East Asian related ancestry, an outcome of Bronze Age events and Mongol Empire expansion along the Silk Road.

Southern tea and horse ancient road

In addition to the northern steppe Silk Road, the Tea and Horse Ancient Road or South Silk Road were another part of the ancient Silk Road. Beginning at approximately 1200 years ago (Tang Dynasty), the Tea and Horse Ancient Road emerged as a famous caravan road system for tea, salt and horse trading in Southwest China. Guizhou Province, located near the Tea and Horse Ancient Road has been documented as a critical depository of substantial sociocultural, genetic, and linguistic diversity for studying southern Silk Road related populations. The southern Han constitute the majority ethnic group within Guizhou. Wang et al. found that the Guizhou Han were in turn a mixed population with shared excess ancestry with Longshan-culture-related middle Yellow River populations. Interestingly, Guizhou Han reveal significant genetic differentiation with geographically neighboring southern ethnic groups, such as the suspected descendants of Pengtoushan, Gaomiao, Daxi, Qujialing, Shijiahe, and other archaeological cultures (Wang et al.).

Western end

At the western end of the Silk Road, Guarino-Vignon et al. found that a bronze age individual from Oxus Civilization (or Bactrio-Margian Archaeological Complex, BMAC) at the Ulug-depe site in Turkmenistan, shared genetic affinity with a local BMAC population, and further revealed that modern Central

Asian Indo-Iranian–speaking populations primarily harbored ancient BMAC related ancestry (Guarino-Vignon et al.). The use of recent aboriginal paleogenomes, combined with the genomes of related modern humans, will help us to understand further details of the genetic history of regional populations.

Conclusion

In summary, this special edition focuses on the population history along the Silk Road, covering the eastern limits of the Silk Road, the Western Regions, the northern steppe, and southern Tea and Horse Ancient Road, and western end. The genetic diversity and population structure of modern populations in these regions are dissected using ancient forensic markers. Moreover, combined with modern human DNA data, ancient DNA researches discuss the formation of targeted populations in greater detail. In future work, fine temporal-spatial scales using enlarged sample sizes should be considered when outlining changes in population dynamics along the ancient Silk Road.

Author contributions

SW, HP, and MZ conducted the project and conceived the idea. XC and SW wrote the paper. All the authors revised the paper.

Funding

This work was supported by research grants from the National Natural Science Foundation of China (32070576), the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP09259560), the B&R Joint Laboratory of Eurasian Anthropology (18490750300), and European Research Council (ERC) grant to Dan Xu (ERC-2019-ADG-883700-TRAM).

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

References

Ge, J. X., Wu, S. D., and Chao, S. J. (1997). *Zhongguo yimin shi (the migration history of China)*. Fuzhou: Fujian People's Publishing House.

Kairov, U., Molkenov, A., Sharip, A., Rakhimova, S., Seidualy, S., Rhie, A., et al. (2022). Whole-genome sequencing and genomic variant analysis of Kazakh individuals. *Front. Genet* 13, 902804–302305. doi:10.3389/fgene.2022.902804

Wen, B., Li, H., Lu, D., Song, X., Zhang, F., He, Y., et al. (2004a). Genetic evidence supports demic diffusion of han culture. *Nature* 431, 302302–302305. doi:10.1038/nature02878

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 August 2021
doi: 10.3389/fgene.2021.727821

Fine-Scale Genetic Structure and Natural Selection Signatures of Southwestern Hans Inferred From Patterns of Genome-Wide Allele, Haplotype, and Haplogroup Lineages

Mengge Wang[1,2†], Didi Yuan[3†], Xing Zou[4], Zheng Wang[5], Hui-Yuan Yeh[6], Jing Liu[5], Lan-Hai Wei[7], Chuan-Chao Wang[7], Bofeng Zhu[8,9,10*], Chao Liu[1,2,8*] and Guanglin He[6,7*†]*

[1] Guangzhou Forensic Science Institute, Guangzhou, China, [2] Faculty of Forensic Medicine, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China, [3] Department of Forensic Medicine, College of Basic Medicine, Chongqing Medical University, Chongqing, China, [4] College of Basic Medicine, Chongqing University, Chongqing, China, [5] Institute of Forensic Medicine, West China School of Basic Science and Forensic Medicine, Sichuan University, Chengdu, China, [6] School of Humanities, Nanyang Technological University, Singapore, Singapore, [7] State Key Laboratory of Marine Environmental Science, State Key Laboratory of Cellular Stress Biology, Department of Anthropology and Ethnology, Institute of Anthropology, National Institute for Data Science in Health and Medicine, School of Life Sciences, Xiamen University, Xiamen, China, [8] Department of Forensic Genetics, School of Forensic Medicine, Southern Medical University, Guangzhou, China, [9] Key Laboratory of Shaanxi Province for Craniofacial Precision Medicine Research, College of Stomatology, Xi'an Jiaotong University, Xi'an, China, [10] Clinical Research Center of Shaanxi Province for Dental and Maxillofacial Diseases, College of Stomatology, Xi'an Jiaotong University, Xi'an, China

The evolutionary and admixture history of Han Chinese have been widely discussed *via* traditional autosomal and uniparental genetic markers [e.g., short tandem repeats, low-density single nucleotide polymorphisms). However, their fine-scale genetic landscapes (admixture scenarios and natural selection signatures) based on the high-density allele/haplotype sharing patterns have not been deeply characterized. Here, we collected and generated genome-wide data of 50 Han Chinese individuals from four populations in Guizhou Province, one of the most ethnolinguistically diverse regions, and merged it with over 3,000 publicly available modern and ancient Eurasians to describe the genetic origin and population admixture history of Guizhou Hans and their neighbors. PCA and ADMIXTURE results showed that the studied four populations were homogeneous and grouped closely to central East Asians. Genetic homogeneity within Guizhou populations was further confirmed *via* the observed strong genetic affinity with inland Hmong-Mien people through the observed genetic clade in Fst and outgroup f_3/f_4-statistics. qpGraph-based phylogenies and f_4-based demographic models illuminated that Guizhou Hans were well fitted *via* the admixture of ancient Yellow River Millet farmers related to Lajia people and southern Yangtze River farmers related to Hanben people. Further ChromoPainter-based chromosome painting profiles and GLOBETROTTER-based admixture signatures confirmed the two best source matches for southwestern Hans, respectively, from northern Shaanxi Hans and southern indigenes with variable mixture proportions in the historical period. Further three-way admixture models revealed larger genetic contributions from coastal southern East

OPEN ACCESS

Edited by:
Shaoqing Wen,
Fudan University, China

Reviewed by:
Wibhu Kutanan,
Khon Kaen University, Thailand
Habiba Alsafar,
Khalifa University,
United Arab Emirates
Lingxiang Wang,
Fudan University, China

***Correspondence:**
Chuan-Chao Wang
wang@xmu.edu.cn
Bofeng Zhu
zhubofeng@i.smu.edu.cn
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: 19 June 2021
Accepted: 29 July 2021
Published: 24 August 2021

Citation:
Wang M, Yuan D, Zou X, Wang Z, Yeh H-Y, Liu J, Wei L-H, Wang C-C, Zhu B, Liu C and He G (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. 12:727821. doi: 10.3389/fgene.2021.727821

Asians into Guizhou Hans compared with the proposed inland ancient source from mainland Southeast Asia. We also identified candidate loci (e.g., MTUS2, NOTCH4, EDAR, ADH1B, and ABCG2) with strong natural selection signatures in Guizhou Hans *via* iHS, nSL, and ihh, which were associated with the susceptibility of the multiple complex diseases, morphology formation, alcohol and lipid metabolism. Generally, we provided a case and ideal strategy to reconstruct the detailed demographic evolutionary history of Guizhou Hans, which provided new insights into the fine-scale genomic formation of one ethnolinguistically specific targeted population from the comprehensive perspectives of the shared unlinked alleles, linked haplotypes, and paternal and maternal lineages.

Keywords: allele-sharing, admixture history, Han Chinese, haplotype chunk, genetic origin, nature selection

INTRODUCTION

Southwestern East Asia is one of the most ethnolinguistically diverse regions around the world. Genetic origin, subsequent migration, isolation, plausible admixture, and local adaptation history of ethnolinguistic southern Chinese populations were widely discussed *via* different genetic markers, mainly including autosomal short tandem repeats (STRs), single nucleotide polymorphism (SNPs), and copy number variations (CNVs) (Chen et al., 2019; Zhang C. et al., 2019; He et al., 2021; Liu et al., 2021a). However, most of these studies focused on the genetic variations and forensic features of low-density genetic markers in the Han Chinese populations. Genome-wide data of Han people were relatively inefficient considering their largest population size and widely geographically distributed features. Previous chip-based population genetic analysis from the southernmost Han Chinese in Hainan Province revealed that these Han Chinese harbored more genetic materials from surrounding indigenous people (Austronesian, Austroasiatic, Tai-Kadai, and Hmong-Mien speakers) (He G. et al., 2020). Additionally, genetic admixture history from northern Hans and northwestern Hans also revealed that extensive admixture events, including the ancestral sources related to the southern East Asians, southern Siberians, and limited but important ancestral sources linked to the western Eurasians, participated in their genomic formation processes (He G.-L. et al., 2020; Yao et al., 2021). There are also important studies focused on the genetic relationships between central Hans and their neighbors (e.g., Han, Manchu, Mongolian, and Tujia). However, all of these studies mainly focused on the patterns from the shared alleles and sample frequency spectrum of independent SNPs (Chen et al., 2021; He et al., 2021) and lacked evidence from a fine-scale genetic structure based on the shared haplotype chunks (successive linked SNP fragments) and uniparental haplogroup lineages.

A previous genetic study based on the frequency spectrum of maternal and paternal founding lineages from the Neolithic to historical populations in North China has found that extensive population movement and admixture occurred here (Chen et al., 2019). Recently, admixture history and genetic structure patterns of modern and ancient East Asians were also discussed and characterized *via* genome-wide ancient DNA data. These earlier findings extracted from ancient genomes documented that the gene flow from the behaviorally and anatomically modern human flowed into Southeast Asians over 65 thousand years ago (kya), which had left major genetic traces in modern Austroasiatic people (McColl et al., 2018). Paleolithic genomes from Tianyuan Cave and Amur River Basin also revealed a complex genetic admixture landscape in northern East Asians since 40 kya (Mao et al., 2021). Yang et al. (2020) recently analyzed the genetic structure and population shift or admixture history of ancient northern and southern East Asians dating back to 9,500–300 years ago. They found that the gene flow among these populations had made contributions to the genetic patterns of all present-day East Asians since the Neolithic (Yang et al., 2020). Wang C. C. et al. (2021) also genotyped and analyzed the most comprehensive set of ancient genomic data from northern, central, and southern East Asia and reconstructed four Holocene population expansion events that shaped the modern genetic diversity of eastern Eurasians, including three eastern migration events spread Languages of Sino-Tibetan, southern Chinese multi-families and Altaic to the surrounding areas with the Yellow River Basin, Yangtze River Basin, and Mongolian Plateau as the centers and one western Eurasian eastward dispersal, which was consistent with the genetically attested population movements accompanied by subsistence shifts (Ning et al., 2020). However, the extent to which ethnically/geographically diverse modern populations obtained ancestry from these ancient ancestral sources remained to be further characterized.

Chronologically and historically, the Han Chinese could trace their origins back to the Huang Di's Tribe (Huaxia Tribe) in the central valley of the Yellow River about 5,000 years ago (China, 1995). Based on the millenarian antiquity of war and politics, the long-range evolution of agriculture technology, and the admixture movement of south–northern population migration, the Han Chinese gradually developed and formed *via* the indigenous populations with northward or southward incomers (China, 1995; Zhao et al., 2015). Subsequently, the northern Han Chinese embarked on a long-range period of continuous southward diffusion across various channels due to other political wars and natural famine over the past two millennia (Chen et al., 2009). They concurrently posed massive genetic admixtures with the native dwellers. Thus, the Han Chinese, together with surrounding indigenous residents and their ancestors, have played a predominant role in shaping the genetic

diversity of East Asians. Previous analyses of the North-to-South Han population structure were systematically explored by both uniparental markers (Y-chromosomal polymorphisms and mitochondrial DNA variations) (Wen et al., 2004; Wei, 2011; Chiang et al., 2018) and genome-wide autosomal SNPs with limited samples or direct co-analysis with the available ancient East Asians (Chen et al., 2009; Cao et al., 2020), which could only demonstrate a close correlation between geographical distribution and genetic structure categories or explore the extent of the impact of the demic or cultural diffusion on the formation of modern East Asians (Wen et al., 2004). Currently, the available ancient DNA studies found the genetic continuity among spatiotemporally diverse people in East Asians (Yang et al., 2020; Mao et al., 2021; Wang C. C. et al., 2021); thus, comprehensively representative modern publicly available datasets and available statistical methods [shared alleles in f-statistics (Patterson et al., 2012), shared haplotypes in fineSTRUCTURE v4 (Lawson et al., 2012) and GLOBETROTTER (Hellenthal et al., 2014), and shared haplogroup lineages] provided more information and the possibility to characterize a more complete picture of population origin, isolation, migration, and admixture processes of Han Chinese.

Guizhou, located in the southwestern region of China, has been documented to be an indispensable place with substantial sociocultural, genetic, and linguistic diversity, and forms the characteristics of a mountainous province with the densest population distribution. Guizhou is an important part of the Yungui Plateau, which is geographically close to ethnolinguistically diverse provinces of Yunnan, Guangxi, and Hunan in the southwest, south, and east, and to Chongqing and Sichuan in the northwest and north. The population in Guizhou is nowadays widely distributed among 18 local minorities including Miao, Bouyei, Dong, Tujia, Yi, Hui, Bai, Yao, Zhuang, Mongolian, Mulam, and Qiang. Han Chinese demographically accounted for the largest population in Guizhou ethnic groups and occupied more than 60% of the Guizhou population officially recognized by the local government (Chen et al., 2018). The overall language landscape of the Chinese population is dominated by more than 10 mainly language families (including Tai-Kadai, Hmong-Mien, Tungusic, Indo-European, Austroasiatic, Austronesian, Turkic, Mongolic, Koranic, and Sino-Tibetan) (Cavalli-Sforza, 1998). However, Guizhou populations are reported to be dominated by the Tai-Kadai, Hmong-Mien, and Sino-Tibetan-speaking families (Chen et al., 2018; Liu et al., 2020). A large number of population genetic researches have been conducted and primarily focused on the forensic polymorphism and genetic structure of the Guizhou Hans *via* STRs included in the AGCU X19 amplification system or InDels included the Investigator DIPplex kit (Chen et al., 2018; He et al., 2019; Liu et al., 2020). The previous genetic analyses focusing on Guizhou populations suggested that the genetic variations of Guizhou Hans were associated with geographical divisions and linguistic classifications. However, genome-wide SNP data have not been provided to investigate the fine-scale genetic structure of southwestern Hans in this ethnolinguistic region. Besides, we also found that ancient DNA in southwestern China is lacking due to the humid and acid–base environment which is not conducive to the preservation of ancient DNA (Yang et al., 2020; Mao et al., 2021; Wang C. C. et al., 2021). Thus, more genome-wide data of geographically denser modern populations and comprehensive population genetic analysis with the surrounding ancient genomes could provide some new insights into historical and prehistoric demographic processes of Guizhou populations. To this end, we generated and analyzed genome-wide SNP data of more than 700,0000 genome-wide SNPs from 50 Han samples across four regions in Guizhou Province to explore the population structure and genetic admixture of the Guizhou Hans and to identify candidate loci targeted for positive natural selection.

RESULTS

General Patterns of the Population Structure

We collected and generated genome-wide data of 50 Han Chinese individuals from Guizhou Province (**Figure 1A** and **Table 1**) and merged it with population data of over 3,000 modern Eurasians (mainly including Altaic, Sino-Tibetan, Austronesian, Austroasiatic, Hmong-Mien, and Tai-Kadai speakers within and around China) and ancient Eastern Eurasians from Nepal, China, Mongolia, Russia, Japan, and others[1] (Jeong et al., 2016; Ning et al., 2020; Yang et al., 2020; Chen et al., 2021; Liu et al., 2021b; Wang C. C. et al., 2021; Yao et al., 2021). We conducted a principal component analysis (PCA) based on the modern population dataset and projected ancient populations onto the basic framework constructed based on the modern genetic variations. We found that ancient populations from Mongolia and Russia formed a cline with modern Tungusic and Mongolic-speaking populations (modern/ancient northeastern Asian cline). Four studied Guizhou Hans were localized between the Tibeto-Burman-speaking population cluster and one meta-population cluster from southern China consisting of Austronesian, Austroasiatic, Tai-Kadai, and Hmong-Mien-speaking people (**Figure 1B**). Four Han Chinese populations were clustered tightly and deviated to southern East Asians compared with northern Han Chinese populations from Shaanxi, Shanxi, and Shandong provinces. Furthermore, Guizhou Hans had a close genetic relationship with other reference Han people and neighboring Tai-Kadai populations compared with the geographically close Guizhou Hmong-Mien-speaking Gejia, Dongjia, and Xijia people. Compared with the ancestry composition of geographically close Chuanqing people (Lu et al., 2020), four Han Chinese from Anshun, Qiannan, Qianxinan, and Qiandongnan cities harbored more shared ancestry related to northern Han Chinese populations. Patterns of the genetic relationship inferred from the first and third components showed the separation between Austronesian and Austroasiatic people (**Figure 1C**). Guizhou Hans still overlapped with southern Han Chinese populations and had a close genetic relationship with Hmong-Mien and Tai-Kadai people on this scale.

[1]https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data

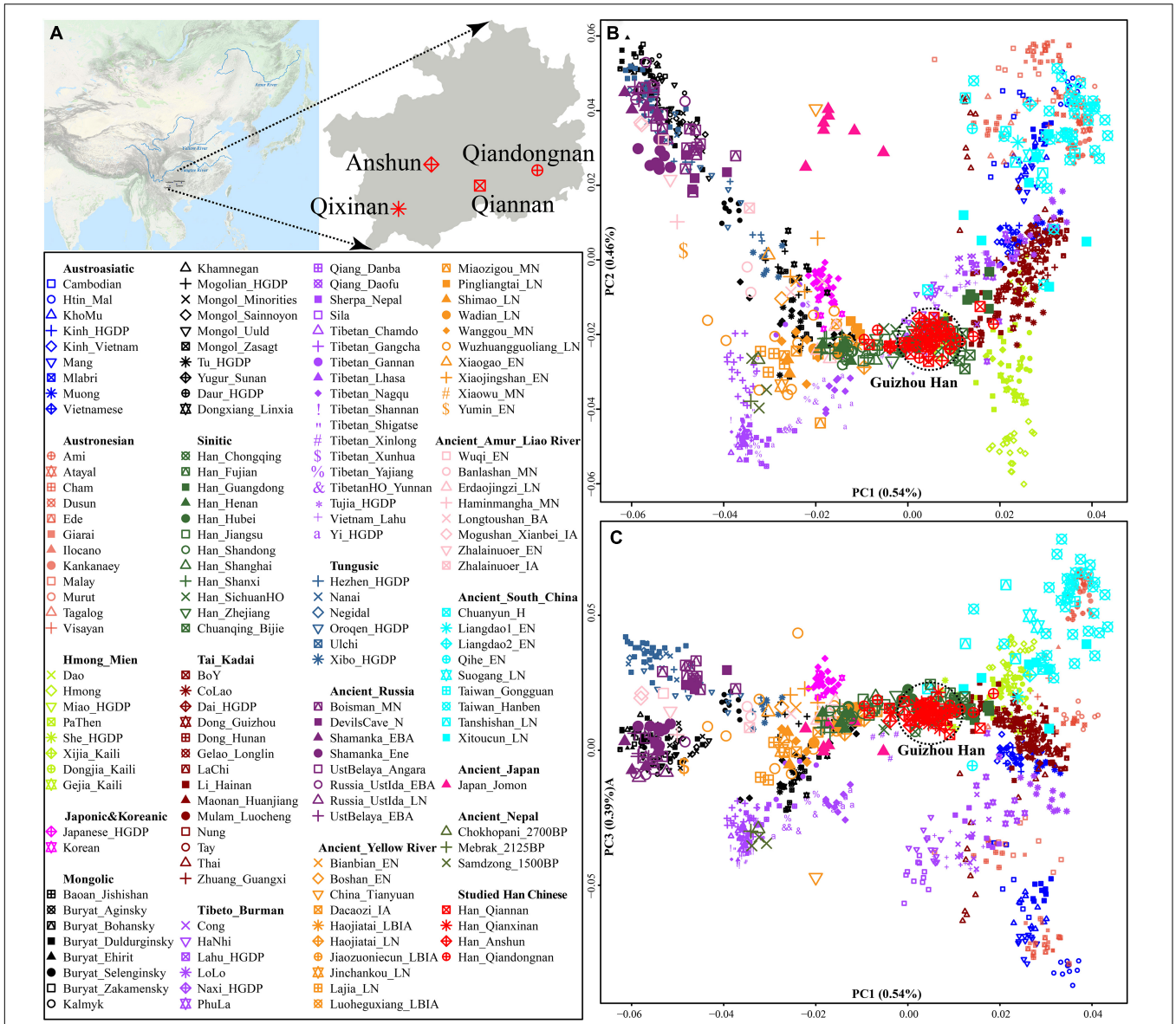


FIGURE 1 | General information of Guizhou Hans. **(A)** The geographic positions of four studied Guizhou Hans in southwestern China. **(B,C)** Principal component analysis based on the top three components among 1,220 modern and ancient East Asians from Altaic, Sino-Tibetan, Austronesian, Austroasiatic, Hmong-Mien, and Tai-Kadai people and other ancient eastern Eurasians (which were projected onto the basic background). Populations were shaped according to their unique ID from one language family. Different populations from one language families or groups were colored.

Furthermore, we further explored the ancestry composition based on the model-based ADMIXTURE analysis (**Figure 2**), which fitted the gene pool of the targeted populations using the unlinked SNP data with specific predefined ancestral sources (2–20). Here, we observed the relatively low cross-validation errors when the *K*-values were equal to 7∼9 (**Supplementary Figure 1**). We could observe four major ancestries in Guizhou Hans when seven ancestral sources were used: pink ancestry maximized in Taiwan Neolithic to Iron Age populations (Hanben and Gongguan) and modern Taiwan indigenous Austronesian Ami people; yellow ancestry dominant in Hmong-Mien-speaking Hmong and PaThen and some Austroasiatic populations; orange ancestry existed in modern Tibeto-Burman-speaking Tibetan and Qiang with high proportion; and limited green ancestry maximized in coastal Amur River Neolithic populations related to Boisman, DevilsGate, and modern Ulchi. Guizhou Hans harbored more Hmong-related ancestry compared with northern Hans (Henan, Shanxi, and Shandong Hans), and geographically close Neolithic to historical ancients in Henan possessed more Tibetan/Qiang-related ancestry. However, ancient people from Haojiatai, Xiaowu, and Pingliangtai had less Hmong-related ancestry compared with the geographically close modern northern Hans, which suggested that population movements and admixture shaped the spatiotemporal landscape in this

TABLE 1 | The demographical information and paternal and maternal haplogroups of our included 50 Han Chinese individuals.

Ind	Group	Status	MtDNA Haplogroup	Sex	Y-chromosome Haplogroup	Key Y-mutations
N0726	Han_Qiannan	Unrelated Healthy	Z3	Female	NA	NA
N0738	Han_Anshun	Unrelated Healthy	B5a1c1	Female	NA	NA
N0741	Han_Anshun	Unrelated Healthy	B5a1c1	Female	NA	NA
N0743	Han_Anshun	Unrelated Healthy	D5b1c	Male	O2a1c1a1a1a1e1a	['18404653 Y16154']
N0745	Han_Anshun	Unrelated Healthy	N9a4	Male	O2a1c1a1a1a1a1a1b	①
N0748	Han_Anshun	Unrelated Healthy	B5a1c1	Male	O2a2a1a2a1a	['16880955 F2309,' '19566267 F3085']
N0749	Han_Anshun	Unrelated Healthy	F1a'c'f	Male	O1b1a1a1b1a	⑤
N0750	Han_Anshun	Unrelated Healthy	B4	Female	NA	NA
N0751	Han_Anshun	Unrelated Healthy	B5a1c1	Female	NA	NA
N0753	Han_Anshun	Unrelated Healthy	D4a5	Female	NA	NA
N0754	Han_Anshun	Unrelated Healthy	C4a1'5	Female	NA	NA
N0764	Han_Anshun	Unrelated Healthy	F1a1	Female	NA	NA
N0765	Han_Anshun	Unrelated Healthy	M7b1a1	Female	NA	NA
N0775	Han_Anshun	Unrelated Healthy	B5a1c1	Female	NA	NA
N0733	Han_Anshun	Unrelated Healthy	D5a3	Female	NA	NA
N0735	Han_Anshun	Unrelated Healthy	D4j11	Male	O2a1c1a1a1a1e2a	['15953241 FGC54507']
N0736	Han_Qiandongnan	Unrelated Healthy	B4a2b	Female	NA	NA
N0740	Han_Qiandongnan	Unrelated Healthy	F1a1d	Female	NA	NA
N0742	Han_Qiandongnan	Unrelated Healthy	M7a1a	Male	O1b1a1a1a1a1a1b	['2738084 Z24091,' '23959373 Z24093']
N0746	Han_Qiandongnan	Unrelated Healthy	D4b2b	Female	NA	NA
N0747	Han_Qiandongnan	Unrelated Healthy	F1a1	Female	NA	NA
N0752	Han_Qiandongnan	Unrelated Healthy	F4a2	Female	NA	NA
N0761	Han_Qiandongnan	Unrelated Healthy	B5a1c1	Female	NA	NA
N0770	Han_Qiandongnan	Unrelated Healthy	B4h1	Male	O2a1c1a1a1a1a1a1a1	['22548606 F1495,' '23976986 F1418']
N0771	Han_Qiandongnan	Unrelated Healthy	D4a3b2	Female	NA	NA
N0774	Han_Qiandongnan	Unrelated Healthy	D4	Female	NA	NA
N0727	Han_Qiannan	Unrelated Healthy	B4a2b	Male	O2a2b1a2a1a2	①
N0737	Han_Qiannan	Unrelated Healthy	D4e1a2	Male	O1b1a1b1	③
N0739	Han_Qiannan	Unrelated Healthy	D5a2a1	Female	NA	NA
N0744	Han_Qiannan	Unrelated Healthy	M7b1a1	Female	NA	NA
N0728	Han_Qiannan	Unrelated Healthy	M7c3	Male	O2a1c1a1a1a1a1a1b	②
N0755	Han_Qiannan	Unrelated Healthy	M9b	Male	O2b1a	④
N0729	Han_Qiannan	Unrelated Healthy	M9b	Female	NA	NA
N0756	Han_Qiannan	Unrelated Healthy	F1a1	Female	NA	NA
N0757	Han_Qiannan	Unrelated Healthy	C7	Male	O1a1a1b2a1	['8598326 Z39268,' '22908919 SK1571']
N0758	Han_Qiannan	Unrelated Healthy	B5b2c	Male	O2a2b1a2a1a2	①
N0772	Han_Qiannan	Unrelated Healthy	B5a1c1	Female	NA	NA
N0773	Han_Qiannan	Unrelated Healthy	F1a1	Male	O2a1a1a	['14928001 F1867']
N0731	Han_Qiannan	Unrelated Healthy	B4c1a	Male	O2a2b1a2a1a2	①
N0734	Han_Qiannan	Unrelated Healthy	G2a	Male	O2a2b1a1a	['2800495 F8,' '6840710 F42']
N0759	Han_Qianxinan	Unrelated Healthy	B5a	Male	O2a1c1a1a1a1a1a1b	②
N0760	Han_Qianxinan	Unrelated Healthy	B4a	Female	NA	NA
N0762	Han_Qianxinan	Unrelated Healthy	F1c1a1	Female	NA	NA
N0763	Han_Qianxinan	Unrelated Healthy	B5a1c1	Female	NA	NA
N0730	Han_Qianxinan	Unrelated Healthy	B5a1	Female	NA	NA
N0766	Han_Qianxinan	Unrelated Healthy	F1d1	Female	NA	NA
N0767	Han_Qianxinan	Unrelated Healthy	F1a2a	Female	NA	NA
N0768	Han_Qianxinan	Unrelated Healthy	F1a2a	Female	NA	NA
N0769	Han_Qianxinan	Unrelated Healthy	B4b1a2	Male	D1a1a1a1a2a~	['16411247 Z44637']
N0732	Han_Qianxinan	Unrelated Healthy	M10a1b	Male	O1a1a1a1a1b1	['23159740 CTS11553']

① *['14173991 F242,' '14946079 F273,' '19371700 CTS10286,' '19436515 CTS10401,' '22821767 CTS10888,' '23024318 F634'].*
② *['6655747 F793,' '7316384 Z43869,' '8568806 F1316C,' '14746939 Z43872,' '15912372 F2035,' '16253175 F2108,' '18241027 Z43875,' '18705724 Z43876,' '21283525 Z43877'].*
③ *['8562287 F1309,' '14065037 F1685,' '14263051 F1742,' '16521659 F2189,' '17326573 F2445,' '22742290 F3353'].*
④ *['14336908 F1770,' '16718236 F2244,' '16733921 F2247,' '22676898 F3338'].*
⑤ *['15097700 B426,' '15320317 Z23672,' '16794293 Y9320,' '18816557 FGC29898'].*

region. We also found more Tibetan/Qiang ancestry in Guizhou Hans compared with their neighboring indigenes (e.g., Dong). When we used the increased predefined ancestral sources in the ancestry composition modeling, southern inland East Asian ancestries associated with Hmong-Mien and Tai-Kadai people were separated, and both contributed to the ancestry

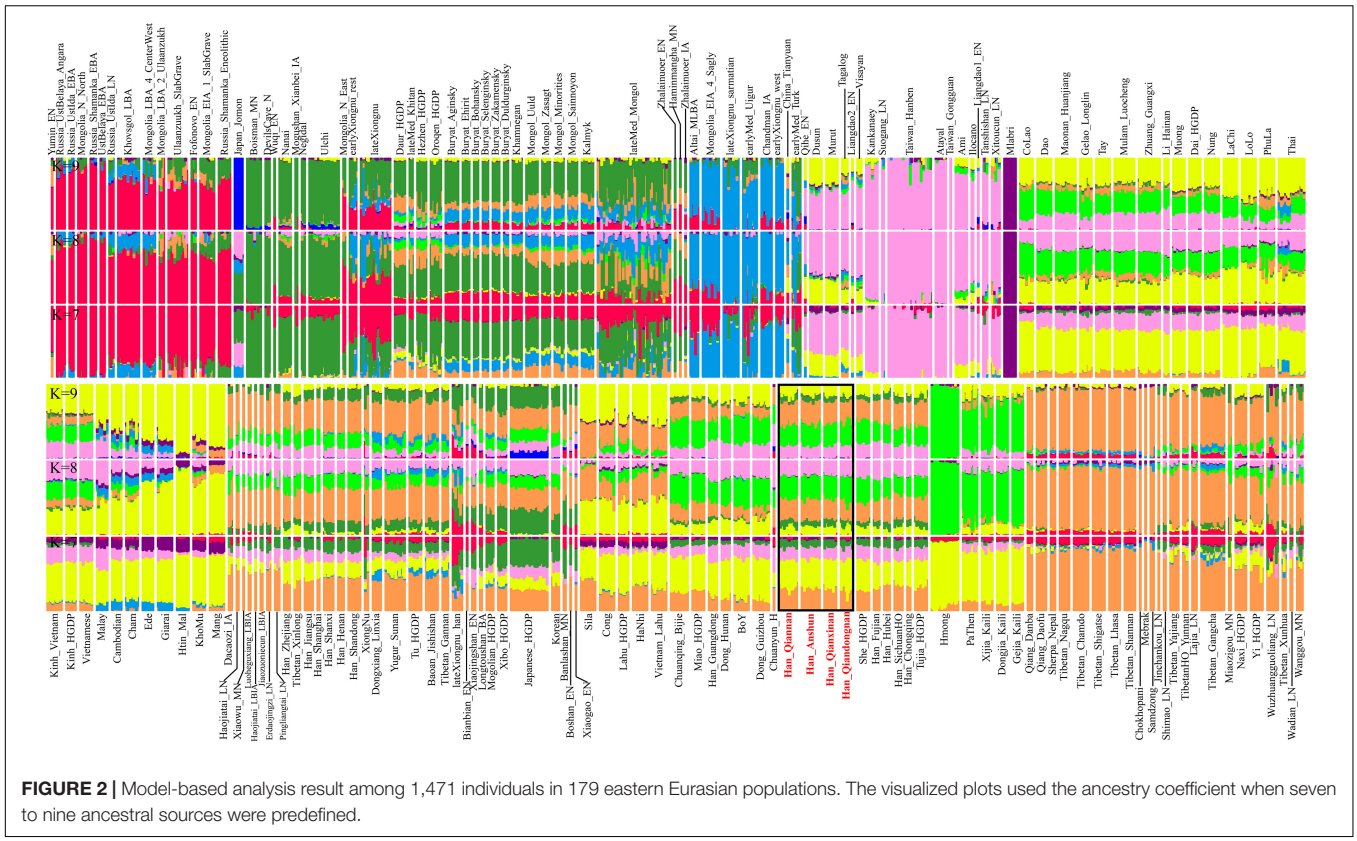


FIGURE 2 | Model-based analysis result among 1,471 individuals in 179 eastern Eurasian populations. The visualized plots used the ancestry coefficient when seven to nine ancestral sources were predefined.

compositions of four Guizhou Hans, which suggested that our four studied Han Chinese populations were mixture populations and obtained gene flow from multiple southern indigenous ancestral populations.

Genetic Origin and Admixture History Inferred From the Shared Ancestry in Allele-Based Analysis

To explore the intra/interpopulation genetic relationships between four genotyped Hans and other references (both modern and ancient populations), we calculated pairwise Fst genetic distances among 87 East Asian groups (**Supplementary Table 1**). We did not identify a genetic difference within the four studied populations, as we observed the most negative or smaller index values within four Hans. Compared with other adjoin populations, Qiannan Hans had a close genetic relationship with central and southern Hans (Hubei Hans: 0.0002, Sichuan Hans: 0.0003, Fujian and Zhejiang Hans: 0.0013). The other three Han Chinese populations possessed similar patterns of genetic correlation with our included reference populations, as the observed high correlation coefficient (larger than 0.99). We also assessed genetic affinity *via* the shared genetic drift measured from the outgroup f_3-statistics in the form f_3(reference modern and ancient populations, targeted populations; Mbuti). We first confirmed the close genetic relationship between Guizhou Hans and southern Han Chinese as the identified largest f_3-values among them (**Supplementary Table 2**), and then we observed affinity between Guizhou Hans and geographically close Hmong-Mien-speaking populations (0.3036 between Anshun Hans and Kali Gejias, 0.3047 between Kaili Gejias and Qiandongnan Hans, 0.3029 between Kaili Gejias and Qiannan Hans, and 0.3030 between Kaili Gejias and Qianxinan Hans). Among ancient reference populations, we also identified the largest values between studied Hans and Taiwan Gongguan and Yellow River Luoheguxiang ancient populations.

We further performed admixture f_3-statistics in the form f_3(source1, source2; targeted populations) to identify and describe plausible admixture events. We identified 4,451 pairs that showed statistically negative f_3-values in Anshun Hans which denoted that the allele frequency of the targeted Anshun Hans intermediated between that in source1 and source2. The observed significant f_3-values suggested that Anshun Han was a genetic admixture population. The most negative source pairs were produced from northern Hans and southern East Asian Tai-Kadai people (**Supplementary Table 3**). The plausible proxies of southern ancestral sources were related to Austronesian and Hmong-Mien-speaking populations, as negative values in f_3(northern sources, Austronesian/Hmong-Mien; Guizhou Hans). Northern Hans-linked ancestral sources were also related to Tibeto-Burman people and some of the Altaic-speaking populations. Indeed, these population pairs of one source from Tibeto-Burman or Altaic-speaking populations and the other from Austronesian and Tai-Kadai groups also produced significant admixture signals. Ancient people from Mongolia and the Yellow River Basin could also be used as

the candidates of northern sources, and ancient populations from Fujian and Taiwan in southeastern coastal regions and Southeast Asia also could be used as effective southern ancestral sources for four Guizhou Hans. We conducted 30,288 pairs of symmetric f_4-statistics in the form f_4(Eurasian1, Eurasian2; targeted Guizhou Hans, Mbuti) and found more shared ancestry between Guizhou Hans and Tai-Kadai-speaking populations (**Supplementary Table 4**), such as the most negative tests of f_4(Giarai, Zhuang_Guangxi; Han_Anshun, Mbuti) = –25.375*SE. Compared with modern coastal southern East Asians related to Amis, Guizhou Hans shared more alleles with inland modern southern East Asians related to Hmong, f_4(Amis, Hmong; Han_Qiandongnan, Mbuti) = –2.818, which was consistent with the observed patterns compared with Taiwan Hanben (Z-scores = –5.382). Affinity f_4-statistics in the form f_4(Eurasain1, Targeted Guizhou Hans; Eurasain2, Mbuti) were conducted to explore if some additional ancestries contributed to Guizhou Hans compared with other Eurasian comparative subjects. Compared with geographically close Guizhou Dong, Guizhou Hans harbored more ancestry related to Tibeto-Burman-speaking Tibetan (Chamdo, –5.274), as well as related to the Yellow River Basin ancient populations of middle Neolithic Wanggou people (–4.364), suggesting that more northern East Asian ancestry existed compared with southern indigenous people. Compared with Yellow River farmers (Haojiatai_LBIA), Guizhou Hans also possessed more ancestry related to southern East Asian indigenous populations related to Mlabri (–4.016) and others, which suggested that Guizhou Hans were formed with the gene pool from northern and southern sources.

Additionally, we reconstructed deep population admixture models for the formation of Guizhou Hans *via* qpGraph-based phylogeny framework with population split and admixture events (**Figure 3A**). Here, we used the late Neolithic Qijia culture-related ancient population as the northern ancestral source and used Neolithic to Iron Age Hanben people from southeastern China as the southern ancestral source. All four studied populations could be successfully fitted in this model with fluctuated proportions, which showed that the southwestern Hans from Guizhou Province harbored both northern and southern ancient East Asian ancestries. Totally, admixture processes kept a similar pattern with Chuanqing people from Guizhou Bijie City (modeled as the admixture of 0.31 of their ancestry related to Lajia and the remaining ancestry from Hanben). Compared with northern Hans from Shaanxi Province and recent southward Manchus and Mongolians in Guizhou Province, ancestral proportion related to the northern Lajia decreased in Guizhou Hans, but increased compared with Guizhou indigenous Hmong-Mien speakers (Gejia, Dongjia, and Xijia, **Figure 3B**). To further explore whether a differentiated genetic contribution from coastal and inland southern East Asians, we followingly conducted three-way admixture models with two sources from southern East Asia (inland and coastal) and one from northern East Asia focused on the four studied Guizhou Hans and four published Guizhou minorities (**Figure 4**). The putative northern sources included middle and upper Yellow River farmers of the Jiaozuoniecun_LBIA, Lajia_LN, Luoheguxiang_LBIA, Miaozigou_MN, Pingliangtai_LN, and Wadian_LN; the coastal southern sources comprised Neolithic to Iron Age populations (Taiwan_Hanben, Taiwan_Gongguan, and Xitoucun_LN); and inland sources were made up of Neolithic to Bronze Age populations of TamHang_BA, GuaCha_LN, and MaiDaDieu_LN. We obtained 603 fitted admixture models with major ancestry from northern sources and the second from the southern coastal sources and the last from the inland southern sources (**Supplementary Table 5**). It should be also noted that the putative inland southern East Asian ancient sources possessed some extent ancestry related to the indigenous Hòabìnhian hunter-gatherers (McColl et al., 2018), which may be biased against the true estimated proportion of inland ancestral sources. Ancient DNA from Daxi, Shijiahe, and other southwestern ancient people in the future will provide better fitted and more explicable admixture models for Guizhou Hans. Here, we also constructed one neighboring-joining phylogenetic tree among 86 modern and ancient eastern Eurasians. The unrooted tree was divided into two branches: the northern one comprised Mongolic and Tungusic speakers and Neolithic to Iron Age populations from Yellow River Basin, Mongolia Plateau, and southern Siberia and the southern branch was made up of Austronesian, Austroasiatic, and Hmong-Mien populations from southern China and Southeast Asia (**Figure 5**). Guizhou Hans and other Han Chinese groups were clustered between these northern and southern branches and had a close phylogenetic relationship with each other and then clustered with geographically close Sichuan, Hubei, and Fujian Hans.

Finer-Scale Population Substructure and Natural Selection Signatures Revealed *via* the Sharing Haplotype Based on the Linked Successive SNPs

Genetic analysis based on the unlinked SNPs can only capture the major information of population history encoded in the genomes. Thus, to further identify, date, and describe the fine-scale admixture events and decode more detailed information of population demographic history of Guizhou Hans, we merged population data with previous published East Asians genotyped using the same array (700K), including 11 Han populations from Shaanxi Province (He G.-L. et al., 2020), Lanzhou Hans from Gansu Province (Yao et al., 2021), Boshu Huis and Nanchong Hans (Liu et al., 2021b) from Sichuan Province, officially unrecognized populations [Chuanqings, Gejias, Dongjias, and Xijias (Lu et al., 2020)], and Manchus and Mongolians from Guizhou Province (Chen et al., 2021). We used SHAPEIT to phase the genome-wide data and obtain the phased SNPs. We first used the chromosome painting strategy instrumented in the ChromoPainter to paint the chromosome of the targeted population conditional on all potential DNA donors and to choose the best ancestry source based on the co-ancestry matrix. Both individual-level shared length and number of haplotype chunks were obtained. Moreover, it showed that the Guizhou Hans shared large and long ancestry fragments (also referred to as identity by the decedent, IBD) with geographically close populations, suggesting that Guizhou Hans had the most common ancestor among them. Furthermore, we used ChromoCombine to combine the shared number of ancestry fragments and converted the differentiated contributed

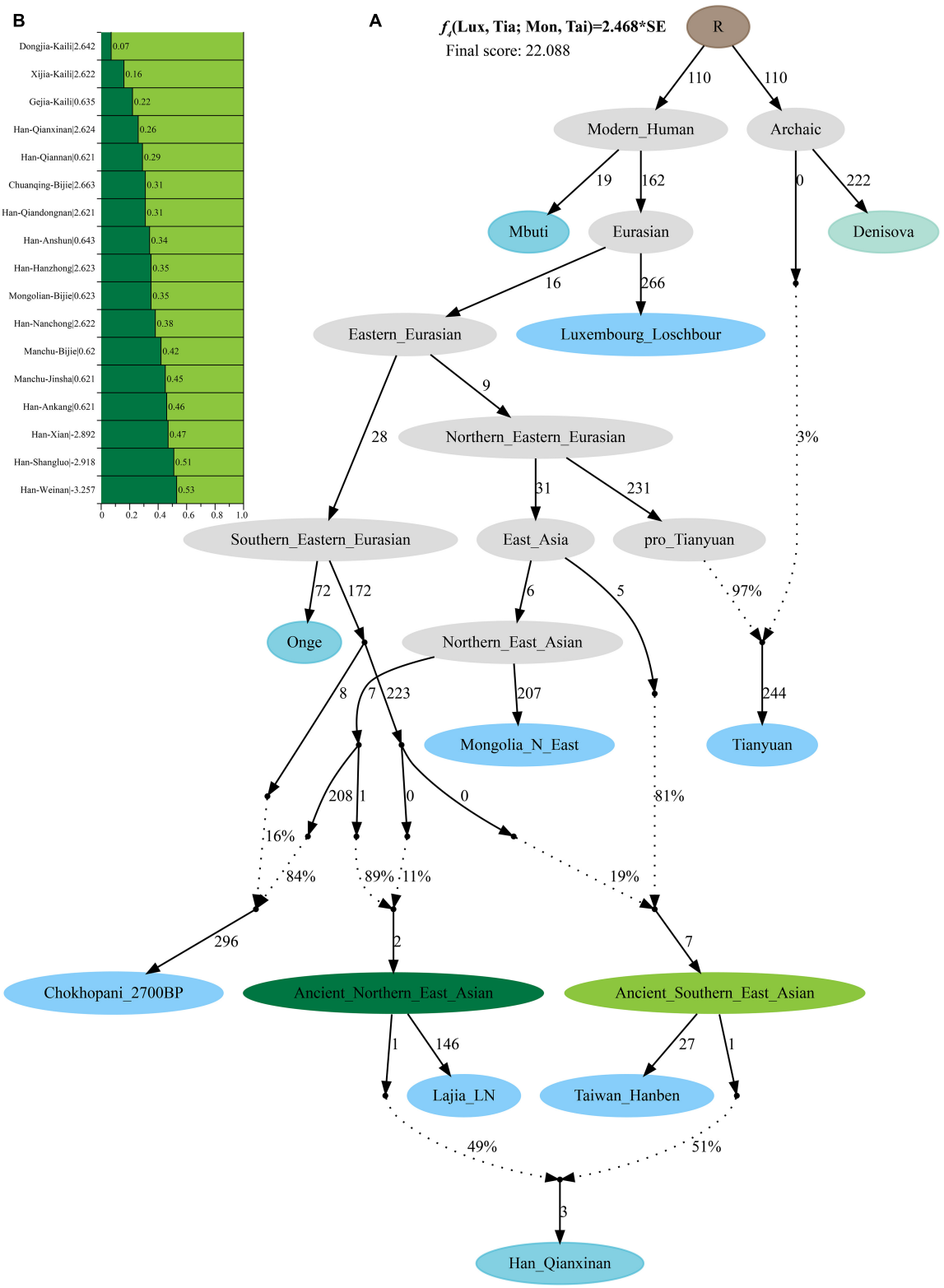


FIGURE 3 | Deep population admixture history reconstruction with the population splits and admixture events. **(A)** The qpGraph-based phylogenetic framework focused on the formation of Qianxinan Hans. **(B)** Admixture proportion and corresponding lowest *Z*-scores of the qpGraph models for other Hans and Guizhou aborigines. Different features of populations (ghost, archaic, ancient, and modern populations) were color-coded. Admixture events were denoted with dotted lines and marked the corresponding admixture proportion on it. Branch length was measured using f_2-values (1,000×).

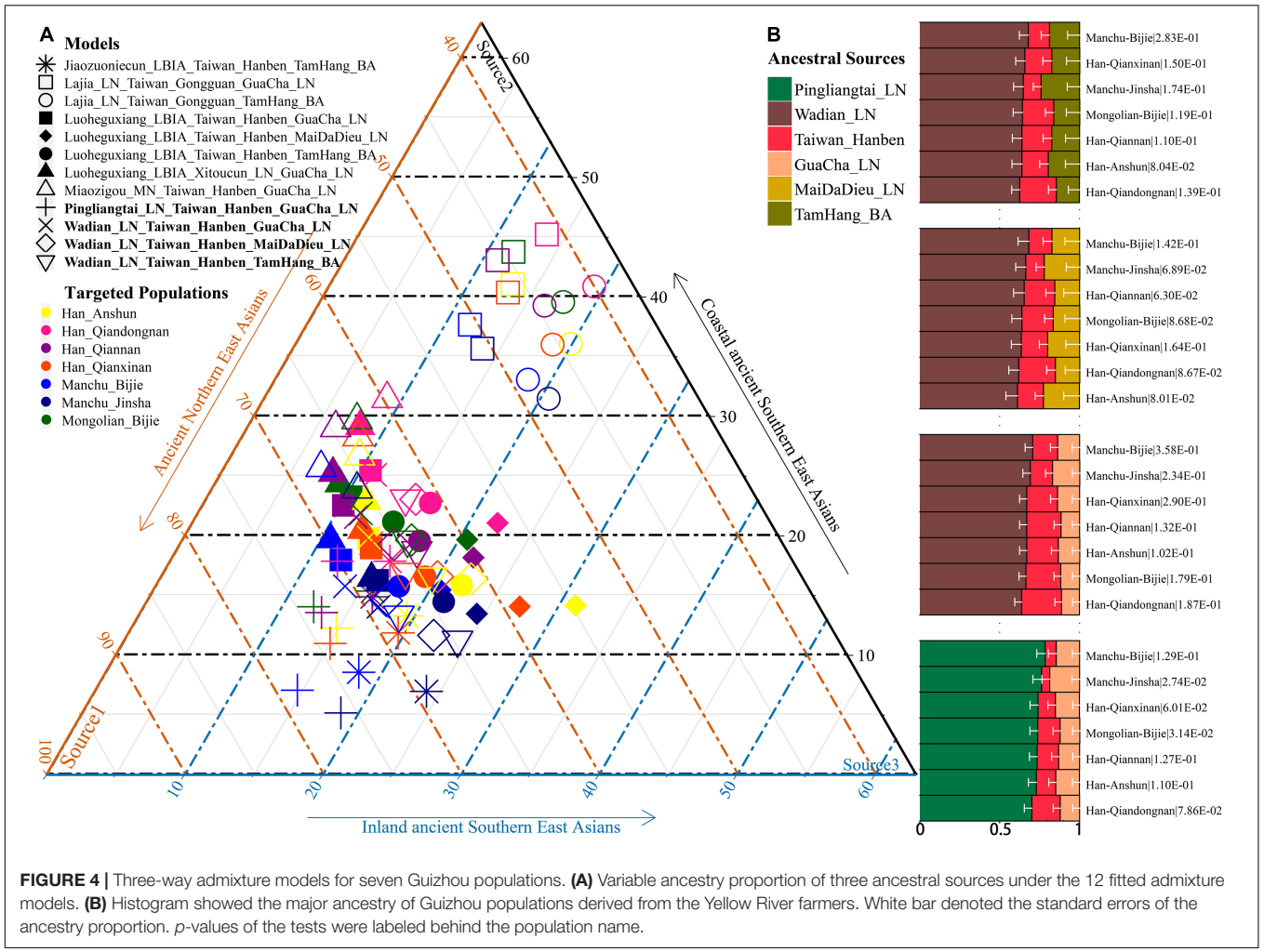


FIGURE 4 | Three-way admixture models for seven Guizhou populations. **(A)** Variable ancestry proportion of three ancestral sources under the 12 fitted admixture models. **(B)** Histogram showed the major ancestry of Guizhou populations derived from the Yellow River farmers. White bar denoted the standard errors of the ancestry proportion. *p*-values of the tests were labeled behind the population name.

haplotype states from all possible donors to the focused targeted populations as the co-ancestry matrix. Fine-scale population structure was characterized from individuals within populations using fineSTRUCTURE and grouped them into new genetically defined groups based on the genetic similarity. Here, the heatmaps of pairwise coincidence, co-ancestry, and average co-ancestry were visualized and used to explore the genetic background. Dendrograms and PCA clustering patterns based on the co-ancestry matrix were also conducted. We observed more subtle population substructures among northern Hans, western Hans, and Central Hans, as well as the clear population stratification between Guizhou Hans and Guizhou indigenous minorities (Gejia, Chuanqing, Dongjia, and Xijia), Huis, Mongolians, and Manchus (**Figure 6**). Most of the inferred genetically homogeneous populations were consistent with the geography-based defined population labels. All included individuals were classified into northern Shaanxi Hans, western Sichuan Hans, and southwestern Guizhou Hans and other minority clusters (Sichuan Huis, Guizhou Mongolians/Manchus, and Guizhou Hmong-Miens). Four Guizhou Hans formed into one branch and clustered with some Nanchong Hans, suggesting the homogenization of Guizhou Hans and the close genetic connection between southwestern Hans and western Hans (may be associated with the historically documented HuGuang Ruchuan migration).

Autosomal haplotype data also could provide new insights into the population admixture history. Previous genetic analyses focused on the shared haplotype have reconstructed the admixture sources and processes of Bantu expansion and Arab slave trade in Africa, the Mongol Empire and the first millennium CE migrations in Eurasia (Hellenthal et al., 2014). Here, we used the reconstructed shared haplotype chunk length from ChromoCombine to further explore the possible admixture events. We used genetically similar groups as the proxy of the true admixture source and employed northern Shaanxi Hans as the possible ancestral northern donors and Guizhou minorities as the southern source donors. We used GLOBETROTTER to analyze 573 individuals from 23 Chinese populations. Strong evidence of admixture was observed in four targeted populations ($p < 0.05$). Anshun Hans were inferred as the one-date admixture results in the best-guess inference, which was mixed from 0.17 haplotypes from local Kaili Xijia people and other 0.83 from northern Hanzhong Hans occurred around 13 generations ago. A similar pattern of one-date admixture model was also obtained

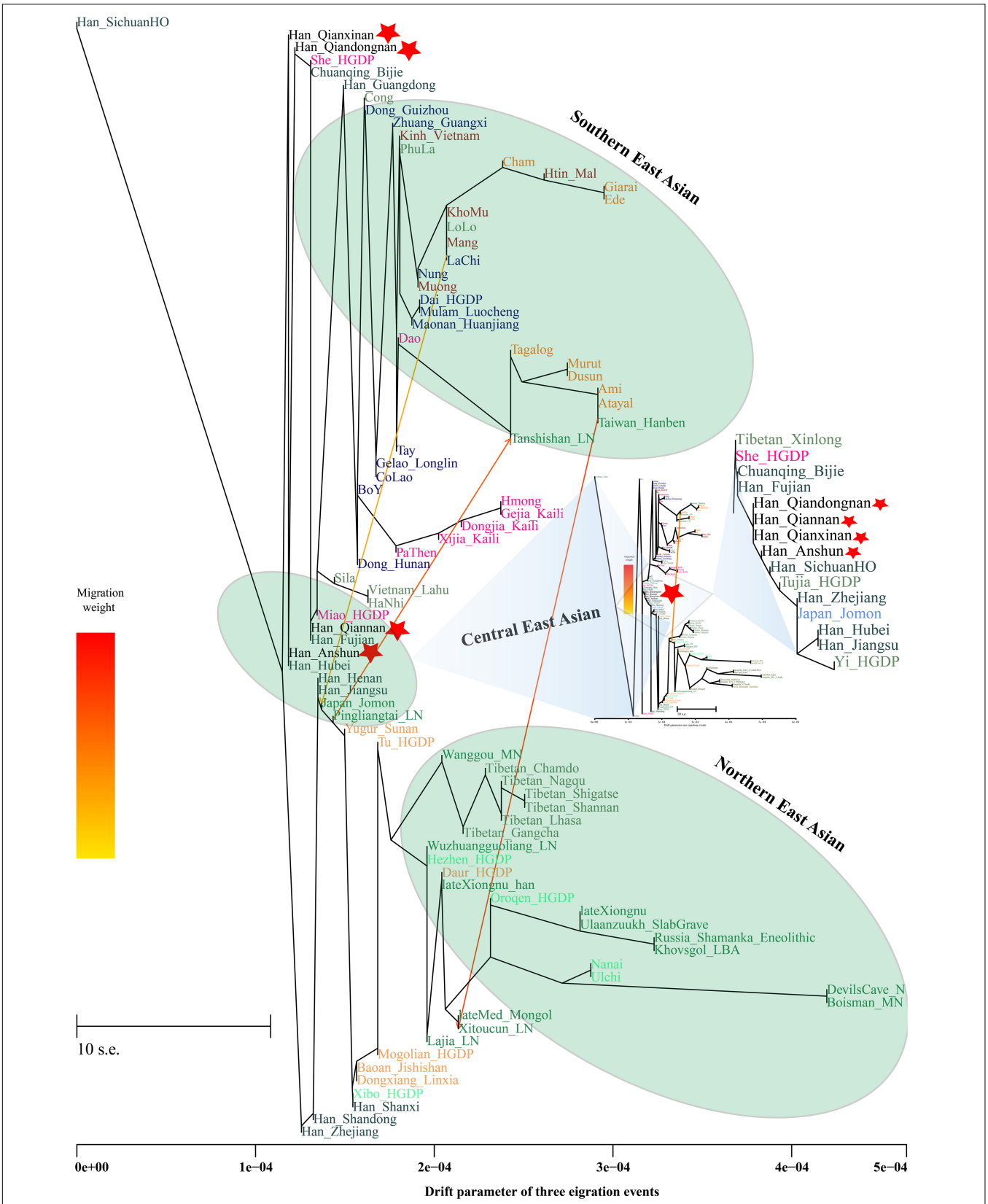


FIGURE 5 | TreeMix-based phylogenetic relationship between four Guizhou Hans and other modern and ancient East Asians with three migration events. Different populations from one language families/groups were color-coded.

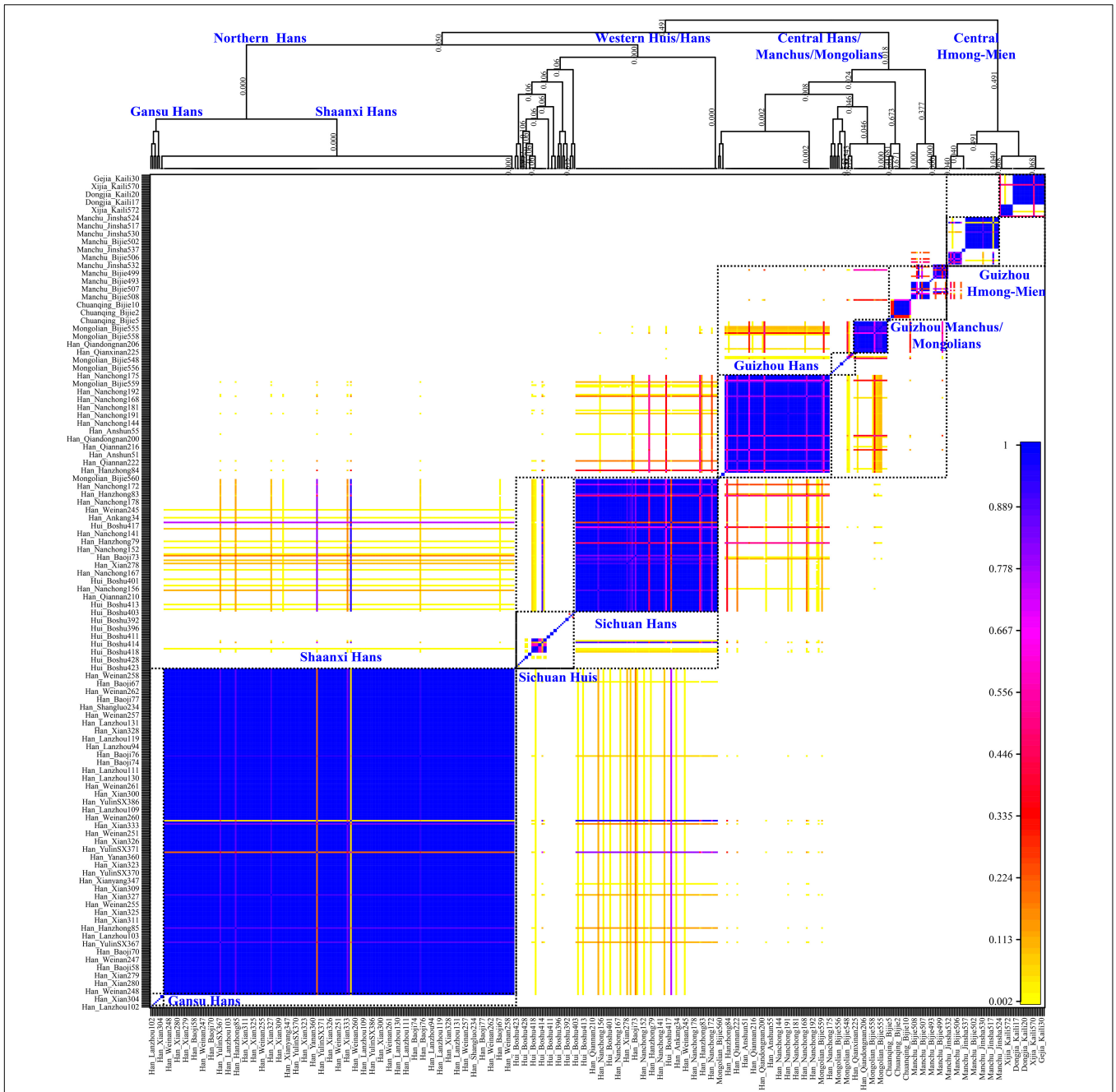


FIGURE 6 | Genetic structure of Chinese populations from haplotype analysis using fineSTRUCTURE. The top parts showed the dendrogram of the genetically homogeneous populations and the bottom heatmap showed the pairwise coincidence. No meaning of the branch lengths but the hierarchy denoted the branching pattern clustered based on the haplotype sharing.

in others with similar best-guess ancestral sources, admixture dates, and corresponding proportions, such as Qiannan Hans that were mixed with minor ancestry (0.2) from Xijia and major ancestry (0.8) from Hanzhong Hans at 15 generations ago.

We also identified and characterized the plausible exiting natural selection signals in Guizhou Hans based on the phased haplotypes using the integrated haplotype score (iHS), nSL, and integrated haplotype homozygosity pooled (iHH12). We used 329,863 phased loci from 100 haplotypes in Guizhou Hans (**Figure 7**). The most significant inferred selection signals were observed from chromosomes 6 and 13. These genes were associated with the susceptibility of complex diseases, including neurogenic locus notch homolog protein4 (NOTCH4) located in 6p21.3, MICB (Human MHC class I chain-related B gene), Microtubule-associated tumor suppressor candidate 2 (MTUS2), and others. Similar patterns of the natural selection

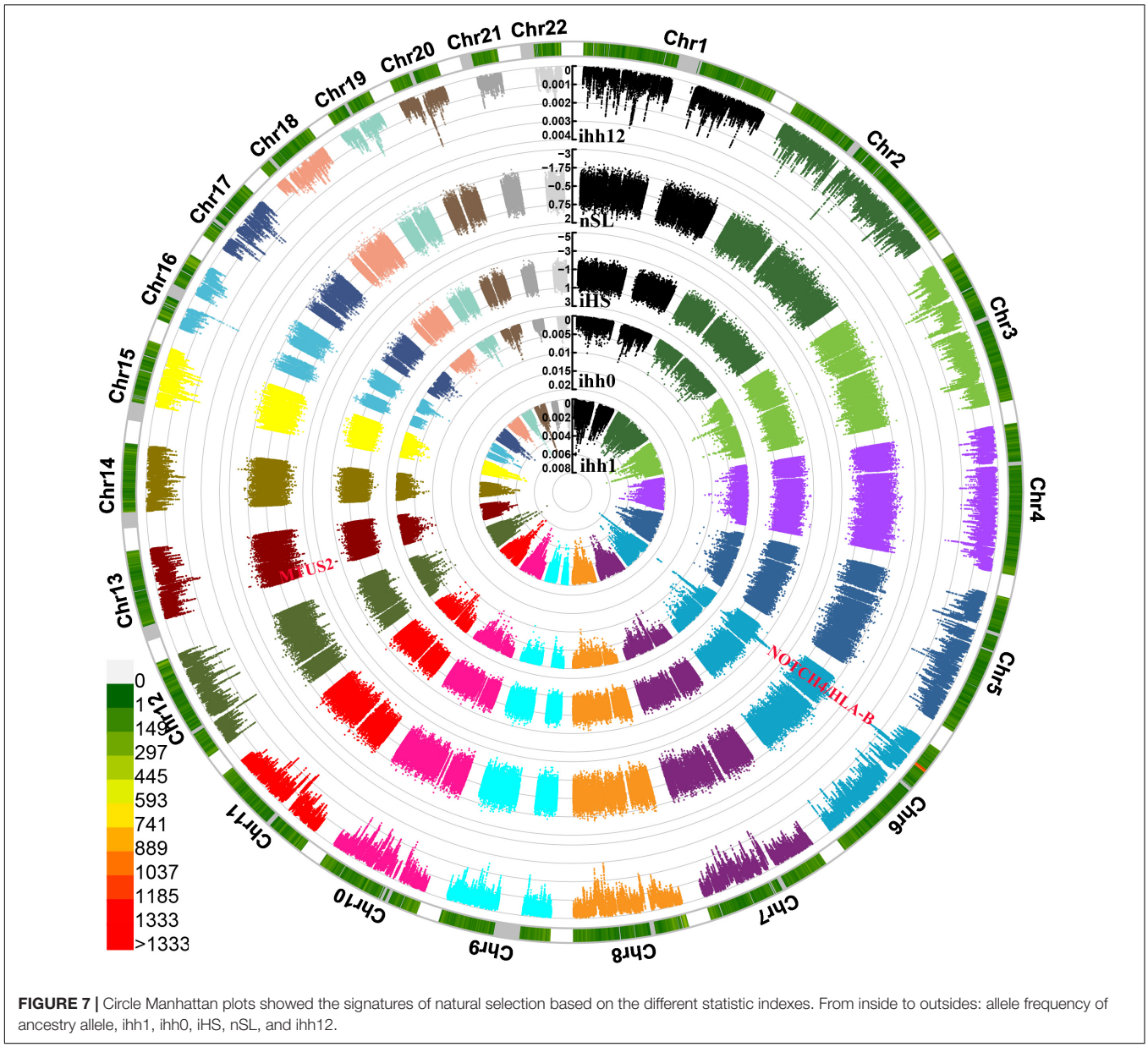


FIGURE 7 | Circle Manhattan plots showed the signatures of natural selection based on the different statistic indexes. From inside to outsides: allele frequency of ancestry allele, ihh1, ihh0, iHS, nSL, and ihh12.

signatures were further confirmed *via* the observed signatures from other natural selection indexes (nSL, iHH12, and iHH). Three SNPs from MTUS2 genes located in chromosome 13 (243198, 243199, and 243196) also possessed negative nSL values (−2.5738, −2.5716, and −2.4624).

Finally, we focused on the genetic diversity and structure within 230 individuals from 14 Guizhou or Chongqing populations. PCA results based on the allele frequency distribution of single SNPs showed three genetic clines (Manchus, Mongolians, and Hmong-Mien speakers) and Guizhou Hans clustered in the intermediated position among three clines (**Figure 8A**), which was further confirmed with the patterns of genetic relationships in the haplotype-based PCA inferred from the co-ancestry matrix (**Figure 8B**). Pairwise IBD and Fst genetic distance visualized in the heatmap (**Figures 8C,D**) revealed the close genetic relationship between Guizhou Hans and geographically close Miao and Tujia and a distant relationship with Tungusic-speaking Manchus and Hmong-Mien-speaking Gejias, Dongjias, and Xijias. Finer-scale genetic structure inferred from fineSTRUCTURE based on the co-ancestry matrix among 14 populations showed that Manchus formed one genetically separated group and Guizhou Hans clustered closed with Mongolians, Tujias, and Hans than with Hmong-Mien-speakers in both individual- and population-level shared chunks (**Figures 8E,F**). These substructures among Chinese southwestern populations were also observed as the identified similar clustered patterns in the TreeMix-based phylogeny framework (**Figure 8G**) and model-based ADMIXTURE models (**Figure 8H**). Focused on this meta-population from southwestern China, natural selection signatures inferred from

iHS showed different top signals, but most signals were observed in chromosome 6 (**Figure 8I**). The inferred loci here are also associated with the susceptibility of the disease.

Admixture Signatures Inferred From the Shared Paternal and Maternal Lineages

We genotyped 3,746 maternal lineage informative SNPs (LISNPs) in 50 female Hans and 24,047 LISNPs in 19 male Hans (**Table 1**). We identified 36 different terminal maternal lineages with the frequency ranging from 0.02 to 0.16. B5a1c1 was the dominant maternal lineage in Guizhou Hans (8/50), followed by F1a1 (4/50), and M7b1a1, B4a2b, M9b, and F1a2a. These observed maternal lineages generally retained relatively high frequencies in southern East Asia and Mainland Southeast Asia. A total of 15 terminal paternal lineages were identified with the frequency ranging from 0.0526 to 0.1579. O2a1b1a1a1a1a1a1b1-F793/Z43869/F1316C/Z43872/F2035/F2108/Z43875/Z43876/Z43877 and O2a2b1a2a1a1a2-F242/F273/CTS10286/CTS10401/CTS10888/F634 were the dominant lineages in the studied Hans.

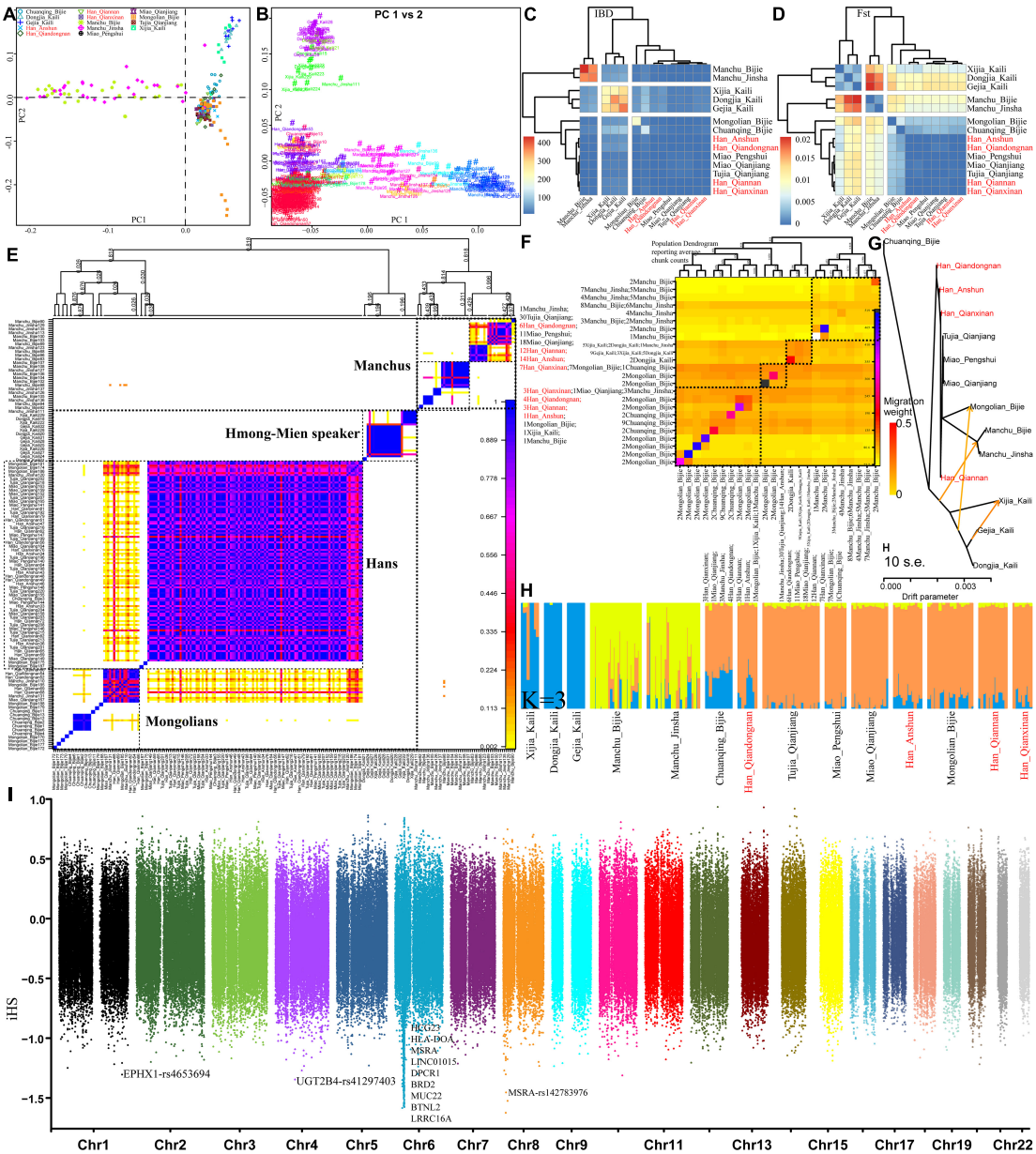


FIGURE 8 | Finer-scale genetic characteristics within 28 southwestern populations. PCA results based on the allele frequency **(A)** and the shared haplotypes **(B)**. Heatmap was visualized based on the pairwise Fst genetic distance **(C)** and the shared IBD segments **(D)**; individual **(E)** or population **(F)** clustering patterns inferred from the co-ancestry matrix in the fineSTRUCTURE analysis; descriptive analysis results from the TreeMix-based phylogeny **(G)** and model-based admixture results **(H)**; natural selection signals inferred from the iHS **(I)**.

The paternal lineage of O2a2a1a2a1a1-F2309/F3085 was also dominant in Guizhou Hans. We also identified one D lineage (D1a1a1a1a2a) and five O1b1a1. The Guizhou Hans with paternal lineages O2a1b1a1a1a1a1a1b1 and O2a2b1a2a1a1a2 might be the descendants of two Neolithic super-grandfathers: Oγ (O2a1b1a1a1a-F11) and Oβ (O2a2b1a2a1a-F46) (Yan et al., 2014). O2a2a1a2a1a1 is common in Hmong-Mien and Austroasiatic speakers (Xia et al., 2019). The infrequent D1a1a1a1a2a∼ is predominant in Tibetan-related populations (Shi et al., 2008; Qi et al., 2013), while O1b1a1 is mainly distributed in southern East Asian, Japanese, and Southeast Asian populations (Yan et al., 2011; Park et al., 2012; Kutanan et al., 2019; Lang et al., 2019).

DISCUSSION

Genetic Origins, Migration, and Admixture History of Guizhou Hans

East Asia is considered to be a region enriched with tremendous cultural and genetic diversity. Researches focused on the peopling of East Asia from ancient genome perspectives showed different ancient genetic ancestries in southern and northern East Asia (Mao et al., 2021; Wang T. et al., 2021). Genetic findings based on the genome-wide SNP data of modern people also demonstrated that the north–south/east–west genetic substructure profiles and demic diffusion of Han Chinese populations shared the overall genetic diversity and demographical history of modern East Asians (Chen et al., 2009; Chiang et al., 2018). In detail, genome-wide association study results based on the Han Chinese populations from 26 administrative regions found substantially genetic differentiation among them, in which its intricate substructures corresponded roughly to the northern Hans, central Hans, and southern Hans. These genetically attested genetic patterns were consistent with historical immigration, cultural exchanges, and geographical characteristics (Xu et al., 2009; Yang et al., 2021). Besides, Yang et al. (2020) recently reported a genome-wide ancient genome study from northern and southern East Asia in the early Neolithic period and found that the population movement and genetic admixture involving northern East Asian ancestry spread southward into Southeast Asia during the Neolithic period, which transformed the genetic ancestry of southern China. Wang T. et al. (2021) sequenced the ancient genomes from Fujian and Guangxi and also identified three ancient ancestry components (Qihe, Longlin, and Hoabinhian) in South China and Southeast Asia, which contributed to the formation of local Early Neolithic people *via* migration and admixture but a limited contribution to modern Guangxi and surrounding populations. Although advanced in deep population history reconstruction of ancient East Asians, the direct genetic contribution from ancient populations to Han Chinese populations needed to be further characterized. A large number of studies have been conducted to explore the genetic structure of Han Chinese groups across China, which suggested that genetic homogenization existed among the Han populations and the genetic differentiation was identified among populations from different language families (Guan et al., 2020; Li et al., 2020; Luo et al., 2020; Wu et al., 2020). Although these signs of progress have been achieved, the genetic history of Guizhou Hans, genetic admixture, or population relationship with neighboring ethnicities was vastly underrepresented due to the lack of sampling of present-day people and comparison with prehistoric East Asian populations. Thus, we generated new genome-wide data over 700K SNPs in Guizhou Hans and merged it with publicly available genomic data from the early Neolithic to the modern populations, and we conducted one comprehensive study focused on the genetic structure and genetic history of the Guizhou Han populations.

Genetic origins and admixture history of four geographically different Guizhou Hans and the finer-scale substructure among Han Chinese based on the genome-wide data were the main studied focus. We also included minorities (Geijia, Dongjia, Xijia, Manchu, and Mongolian) residing in Guizhou Province and other Hans from the surrounding provinces as our references (Lu et al., 2020; Chen et al., 2021; Liu et al., 2021b). PCA and ADMIXTURE results clustered Guizhou Hans in the intermediated position between northern East Asians and southern East Asians, consistent with the clustering patterns in the TreeMix-based phylogeny and the geographical location. Guizhou Province is one of the ethnolinguistically diverse provinces; thus, non-Hans played an important role in the gene pool of the Guizhou populations (Chen et al., 2021). Thus, the effect of the genetic material of Guizhou indigene on the genetic composition of Guizhou Hans is interesting. Indeed, comparative results in the descriptive analysis and qualitative measures in the *f*-statistics, as well as the fineSTRUCTURE-based finer-scale admixture evidence and corresponding admixture proportion from the putative ancestral sources, consistently showed that Guizhou Hans were mixed populations and shared excess ancestry with northern Hans and Henan late Neolithic to Iron Age ancient populations associated with Longshan culture and their dependents and also showed significant genetic differentiation with geographically close southern indigenes. Our findings were consistent with archeologically attested population history of ancient southwestern China, which has suggested that the main components of the southern indigenous people in Guizhou Province may be the direct descendants of prehistoric people associated with southwestern ancients linked with Pengtoushan, Gaomiao, Daxi, Qujialing, Shijiahe, and other cultures (Yu and Li, 2021). Thus, we could identify more southern East Asian ancestry related to Hanben, Hmong, or Tai-Kadai people in Guizhou minorities compared with Guizhou Hans.

Another point of this work was the focus on the patterns of the fine-scale genetic structure of Han Chinese from northern (Shaanxi) and southwestern Hans (Guizhou) and minority ethnic groups (Guizhou) and natural selection signatures based on the reconstructed ancestral chunks. We identified genetic differences among Shaanxi, Sichuan, and Guizhou Hans, as well as the genetic distinction between Han Chinese and minorities *via* the fineSTRUCTURE-based population clustering patterns. We also identified the recent admixture time based on the ALDER and GLOBETROTTER. The inferred that differentiated admixture signatures may be the plausible explanation for the observed genetic differentiation between Guizhou Hans and

their non-Han neighbors. In the qpGraph- and qpAdm-based admixture models, we identified all northern and southern Neolithic farmers who participated in the formation of Guizhou Hans, which was related to the descendants of Yangtze Valley farmers and Yellow River Basin farmers (Wang et al., 2020), suggesting population interactions between northern and southern China were more ancient than the simplified admixture models proposed by ALDER and GLOBETROTTER. Multiple and continuous waves of multiple sources may participate in the formation of ethnolinguistically diverse East Asians. Recent ancient DNA findings of differentiated allele sharing in affinity f_4-statistics between modern East Asians and Neolithic farmers from the northern Yellow River Basin and southern Yangtze River Basin can provide some clues for these hypothesized complex admixture models (Yang et al., 2020), which should be further explored *via* more modern and ancient deep sequencing data and more complex and reasonable biostatistic methods. Additionally, the primary ancestry of Han Chinese in Guizhou Province from northern China was further evidenced *via* the f_4-statistics and GLOBETROTTER-based admixture characterization, supporting much-shared gene ancestry between the southwestern Han Chinese and the present-day northern Sino-Tibetan populations (e.g., Sino-Tibetan and northern Sinitic-speakers), which provided more autosomal genetic supporting evidence for the common origins of Sino-Tibetan language and people from the Yellow River Basin in northern China (Zhang M. et al., 2019).

Positive Natural Selection Based on Successive Linked SNPs

We had explored the genome-wide candidate loci targeted by natural selections based on the reconstructed haplotype data. Among the identified natural selection loci, the top hundreds of SNPs were located in NOTCH4 and HLA in chromosome 6, which was associated with the susceptibility of complex diseases. SNP rs9262558 linked to the HLA gene family located in chromosome 6 was recently evidenced by obvious positive natural selection in Taiwan Hans ($|\text{iHS}| = 7.5$), and it also showed similar signatures in Guizhou Hans (−3.1243) (Yan et al., 2014).

Interestingly, we did not observe the most significant selection signals in EDAR (ectodysplasin A receptor, which was associated with facial and hair morphology in East Asians) and SLC24A5 (pigmentation gene in western Eurasians). Among 11 functional SNPs located in EDAR in chromosome 2 (rs260674, rs12466509, rs3827760, rs10865026, rs260687, rs260690, rs260714, rs6542787, rs6750964, rs922452, and rs72939934 with iHS values from −04017 to 1.2616), the first three SNPs were likely subjected to natural selection in Guizhou Hans with iHS values larger than 1. All of them except SNP rs72939934 harbored nSL values ranging from 1.1930 to 1.4646. SNP rs3827760, one variation being recently evidenced harboring greater genetic differentiation between Gansu Huis and Hans (Ma et al., 2021), also had relatively high natural selection signals in Guizhou Hans (iHS: 1.1139 and nSL: 1.2567). Among 3,133 SNPs located in the Solute Carrier Family (SLC) genes, we identified 70 loci that had large iHS values larger than 1 and 527 SNPs less than −1. The most natural selection marker was SNP rs11966200 located in SLC44A4 in chromosome 6, which was associated with susceptibility of postlingual non-syndromic mid-frequency hearing loss (Ma et al., 2017). SNP rs4148211 located in ABCG8 (iHS: −0.1227 and nSL: 0.0538 in Guizhou Hans) was also one identified SNP that possessed high genetic differentiation between Huis and Hans (Ma et al., 2021), but which is not significant in our studied Hans. Among 83 SNPs, we identified 15 SNPs with the natural selection signals ($\text{iHS} < -1$) located in ABCG1 or ABCG2 associated with the regulation lipid metabolism, especially for SNP rs3788008 (iHS: −2.1567). Endothelial Per-Arnt-Sim (PAS) domain protein 1 (EPAS1) and Hypoxia-inducible factor prolyl hydroxylase (EGLN) were evidenced as the key mutations of human adaptation to the high-altitude environment (Yi et al., 2010). Here, we analyzed 58 SNPs located in these two genes in lowland Hans and we only identified that two SNPs (rs3733829 and rs10151526) in EGLN and four SNPs (rs4953361, rs59901247, rs7577700, and rs187821419) in EPAS displayed natural selection signals. Seven out of 22 candidate loci located in the ADH in chromosome 4 also showed high $|\text{iHS}|$ scores, especially for four loci (rs1042026, rs2066701, rs2075633, and rs1229984) located in ADH1B, which was strongly associated with alcohol metabolism. Next, population genomic analysis with denser sapling with larger sample size based on the genotyping array, second-generation sequencing, and third-generation sequencing (nanopore sequencing) should be conducted to confirm our findings and explore more comprehensive admixture profiles of Guizhou Hans.

CONCLUSION

We conducted one comprehensive population genomic analysis based on three types of shared ancestries (sharing alleles, haplotype, and uniparental lineage) among Guizhou Hans and all publicly available genome-wide data from the early Neolithic to the modern Eurasian populations. We explored and reconstructed the genetic origin, migration, and admixture history of Guizhou Hans, as well as illuminated the candidate loci targeted for positive natural selection. Our survey illuminated that the present-day Guizhou Hans mainly derived the major ancestry from the Yellow River millet farmers and also obtained additional admixture ancestry from an indigenous southern source related to Yangtze River rice agriculturalists. Genetic clustering analysis based on the sharing of IBD patterns and paternal and maternal lineages also demonstrated that the Guizhou Hans were located in the middle position of the North–South genetic gradient consistent with their geographical origin. Additionally, we identified great genetic differentiation between Guizhou Hans and northern Shaanxi Hans, as well as between Guizhou Hans and Guizhou indigenous non-Han people based on the fineSTRUCTURE-based shared ancestry fragments, although we found a strong genetic homogeneity within four Guizhou Han populations. Finally, we searched for signatures of positive selection in the Guizhou Hans by scanning for SNPs that displayed unusually long haplotype lengths using iHS and

identified hundreds of loci located in chromosome 6 (including HLA and NOTCH4) associated with the susceptibility of the complex diseases and other loci located in EDAR, ADH1B, and ABCG2 associated with morphology formation, alcohol, and lipid metabolism.

MATERIALS AND METHODS

Sample Collection and SNP Genotyping

Following the recommendations of the Helsinki Declaration of 2013 (Wilson, 2013), this work was approved by the Ethics Committee of Xiamen University (XDYX201909). We collected 50 saliva samples from 50 unrelated individuals from four cities in Guizhou Province (Anshun, Qiannan, Qianxinan, and Qiandongnan, **Figure 1A**). All participants in this study were needed to be indigenous Han people residing in Guizhou Province for at least three generations. A high-density SNP genotyping array of Infinium® Global Screening Array (GSA) was used to genotype around 700K SNPs from both autosomal and uniparental chromosomes. Genotyping success rate and missing rate per loci or individual were further controlled *via* PLINK 1.9 following our recent similar work (Chang et al., 2015; Liu et al., 2021b). We merged our data with publicly available modern and ancient reference populations from the Human Origin dataset and the 1240K dataset[2] as the low-density dataset used for allele-based analysis. We also merged the newly generated data with other Chinese populations genotyped using the same genotyping chip as the high-density dataset for the haplotype-based analysis, including Hans, Huis, Mongolians, Manchus, Gejias, Xijias, and Dongjias (Chen et al., 2021; Liu et al., 2021b; Yao et al., 2021).

Allele-Based Population Genetic Analysis

Principal component analysis (PCA) among eastern Eurasian or other local-scale populations was conducted using the smartpca package (Patterson et al., 2012). Ancient populations from Mongolia, China, and other neighboring countries were projected onto the top two components. PLINK 1.9 (Chang et al., 2015) was used to calculate the pairwise Fst genetic distances. Model-based clustering analyses were conducted based on the PLINK-pruned unlinked data (–indep-pairwise 200 25 0.4) and conducted using ADMIXTURE 1.3.0 (Alexander et al., 2009) with the predefined ancestry sources ranging from 2 to 20. We run 100 times for each predefined admixture model. A shared genetic drift between Guizhou Hans and other reference populations was measured using the *qp3pop* packages (Patterson et al., 2012) with the tested form of f_3(Guizhou Hans, reference populations; Mbuti); here, the central African Mbuti population was used as the outgroup. Allele-based admixture signatures were explored using the admixture statistics in the form f_3(source1, source2; Guizhou Hans), in which the observed negative values with Z-scores less than –3 denoted a strong admixture evidence. Four population-based analyses (both affinity f_4- and asymmetric f_4-statistics) were conducted using the *qpDstat* package (Patterson et al., 2012) with the f_4 model used. The number of ancestral sources and corresponding admixture proportions was calculated using *qpWave/qpAdm* packages (Patterson et al., 2012), and qpGraph-based phylogeny with admixture events was constructed using *qpGraph* (Patterson et al., 2012). We also built the phylogenetic tree based on the allele frequency using TreeMix (Pickrell and Pritchard, 2012). ALDER (Loh et al., 2013) was used to estimate the admixture times with the predefined ancestral sources.

Haplotype-Based Population Genomic Analyses and Uniparental Haplogroup Assignment

We first phased all populations included in the high-density dataset using ShapeIT v2 (Browning and Browning, 2013). Four Guizhou Han Chinese populations were regarded as the targeted populations, and other reference populations were used as the surrogated populations. All targeted and surrogated populations were used as the donor populations to paint the used recipient populations using the chromosome painting strategies in the ChromoPainter v2 and ChromoCombine v2 (Lawson et al., 2012). FineSTRUCTURE v4 (Lawson et al., 2012) was used to reconstruct individual-based trees and group genetically homogeneous clusters based on the IBD number. We also used the GLOBETROTTER (Hellenthal et al., 2014) to explore, date, and characterize the admixture events based on the IBD length. We assigned and determined the terminal Y-chromosomal haplogroups using our in-house script based on the Y-chromosome tree resource version 2017 in the International Society of Genetic Genealogy (ISOGG)[3]. Maternal haplogroups were determined using HaploGrep 2 (Weissensteiner et al., 2016).

DATA AVAILABILITY STATEMENT

The original data contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directly obtained from the corresponding author/s.

ETHICS STATEMENT

This work was approved by the Ethics Committee of Xiamen University. The sample was collected with informed consent. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

C-CW, BZ, CL, and GH designed this study. MW, DY, and GH wrote the manuscript. MW, GH, XZ, ZW, H-YY, JL, and L-HW

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

[3]https://isogg.org/tree/index.html

conducted the experiment. MW, GH, XZ, ZW, H-YY, JL, and L-HW analyzed the results. C-CW, BZ, CL, and GH revised the manuscript. All the authors reviewed the manuscript.

FUNDING

This study was supported by China Postdoctoral Science Foundation (2021M691879), the National Natural Science Foundation of China (31801040), Nanqiang Outstanding Young Talents Program of Xiamen University (X2123302), and Fundamental Research Funds for the Central Universities (2021M691879).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.727821/full#supplementary-material

Supplementary Figure 1 | The distribution of cross-validation errors.

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

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., Xu, M., Feng, Z., Sun, X., and Lu, J. (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.U. S. A.* 95, 11501–11503.

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, 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, 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

Chen, P., He, G., Zou, X., Wang, M., Jia, F., Bai, H., et al. (2018). 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., 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

Chiang, C. W. K., Mangul, S., Robles, C., Sankararaman, S., and Mulligan, C. (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

China, N. (1995). War and Politics in Ancient China, 2700 BC to 722 BC: measurement and Comparative Analysis. *J. Conflict Resolut.* 39, 467–494. doi: 10.1177/0022002795039003004

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

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., 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

He, G.-L., Wang, M.-G., Li, Y.-X., Zou, X., Yeh, H.-Y., Tang, R.-K., et al. (2020). Fine-scale north-to-south genetic admixture profile in Shaanxi Han Chinese revealed by genome-wide demographic history reconstruction. *J. Syst. Evol.* 1–18. doi: 10.1111/jse.12715

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

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

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

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

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.

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., Zou, X., Zhang, G., Wang, H., Su, Y., Wang, M., et al. (2020). Population genetic analysis of Shaanxi male Han Chinese population reveals genetic differentiation and homogenization of East Asians. *Mol. Genet. Genomic Med.* 8:e1209.

Liu, Y., Wang, M., Chen, P., Wang, Z., Liu, J., Yao, L., et al. (2021a). Combined Low-/High-Density Modern and Ancient Genome-Wide Data Document Genomic Admixture History of High-Altitude East Asians. *Front. Genet.* 12:582357. doi: 10.3389/fgene.2021.582357

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., 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

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

Luo, L., Gao, H., Yao, L., Liu, H., Zhang, H., Wu, J., et al. (2020). Updated population genetic data of 15 autosomal STR loci in a Shandong Han population from East China and genetic relationships among 26 Chinese populations. *Ann. Hum. Biol.* 47, 472–477. doi: 10.1080/03014460.2020.1749928

Ma, X., Yang, W., Gao, Y., Pan, Y., Lu, Y., Chen, H., et al. (2021). Genetic origins and sex-biased admixture of the Huis. *Mol. Biol. Evol.* [Epub Online ahead of print]. doi: 10.1093/molbev/msab158.

Ma, Z., Xia, W., Liu, F., Ma, J., Sun, S., Zhang, J., et al. (2017). SLC44A4 mutation causes autosomal dominant hereditary postlingual non-syndromic mid-frequency hearing loss. *Hum. Mol. Genet.* 26, 383–394.

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.e13.

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.

Park, M. J., Lee, H. Y., Yang, W. I., and Shin, K.-J. (2012). Understanding the Y chromosome variation in Korea—relevance of combined haplogroup and haplotype analyses. *Int. J. Legal Med.* 126, 589–599. doi: 10.1007/s00414-012-0703-9

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

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

Qi, X., Cui, C., Peng, Y., Zhang, X., Yang, Z., and Zhong, H. (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

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

Wang, C.-C., Yeh, H.-Y., Popov, A. N., Zhang, H.-Q., Matsumura, H., and Sirak, K. (2020). The Genomic Formation of Human Populations in East Asia. *bioRxiv* [Preprint]. doi: 10.1101/2020.03.25.004606

Wang, C. C., Yeh, H. Y., Popov, A. N., Zhang, H. Q., Matsumura, H., and Sirak, K. (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., and Yang, Q. (2021). Human population history at the crossroads of East and Southeast Asia since 11,000 years ago. *Cell* 184, 3829–3841.e21.

Wei, L. (2011). Genetic evidences are against a common origin of the Altaic populations. *Commun. Contemp. Anthropol.* 5, 229–236/e38,

Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.-J., et al. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* 44, W58–W63.

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

Wilson, C. B. (2013). An updated Declaration of Helsinki will provide more protection. *Nat. Med.* 19:664. doi: 10.1038/nm0613-664

Wu, X., Zheng, J. L., Lou, Y., Wei, X. H., Wang, B. J., and Yao, J. (2020). Genetic polymorphisms of 20 autosomal STR loci in the Han population of Zhangzhou City, Southeastern China. *Leg. Med.* 46:101726. doi: 10.1016/j.legalmed.2020.101726

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. *bioRxiv* [Preprint]. doi: 10.1101/730903

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

Yan, S., Wang, C. C., Li, H., Li, S. L., Jin, L., and Genographic, C. (2011). An updated tree of Y-chromosome Haplogroup O and revised phylogenetic positions of mutations P164 and PK4. *Eur. J. Hum. Genet.* 19, 1013–1015. doi: 10.1038/ejhg.2011.64

Yan, S., Wang, C. C., Zheng, H. X., Wang, W., Qin, Z. D., and Wei, L. H. (2014). Y chromosomes of 40% Chinese descend from three Neolithic super-grandfathers. *PLoS One* 9:e105691. doi: 10.1371/journal.pone.0105691

Yang, M. A., Fan, X., Sun, B., Chen, C., Lang, J., and Ko, Y. C. (2020). Ancient DNA indicates human population shifts and admixture in northern and southern China. *Science* 369, 282–288. doi: 10.1126/science.aba0909

Yang, X., Wang, X. X., He, G., Guo, J., Zhao, J., Sun, J., et al. (2021). Genomic insight into the population history of Central Han Chinese. *Ann. Hum. Biol.* 48, 49–55. doi: 10.1080/03014460.2020.1851396

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., and Pool, J. E. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329, 75–78.

Yu, X., and Li, H. (2021). Origin of ethnic groups, linguistic families, and civilizations in China viewed from the Y chromosome. *Mol. Genet. Genomics* 96, 783–797. doi: 10.1007/s00438-021-01794-x

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

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

Zhao, Y. B., Zhang, Y., Zhang, Q. C., Li, H. J., Cui, Y. Q., Xu, Z., et al. (2015). Ancient DNA reveals that the genetic structure of the northern Han Chinese was shaped prior to 3,000 years ago. *PLoS One* 10:e0125676. doi: 10.1371/journal.pone.0125676

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 GH, H-YY, L-HW, C-CW, BZ, and 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

ORIGINAL RESEARCH
published: 22 September 2021
doi: 10.3389/fgene.2021.740167

Ancient Mitochondrial Genomes Reveal Extensive Genetic Influence of the Steppe Pastoralists in Western Xinjiang

Chao Ning[1,2†], Hong-Xiang Zheng[3†], Fan Zhang[1†], Sihao Wu[1], Chunxiang Li[1], Yongbin Zhao[4], Yang Xu[1], Dong Wei[5], Yong Wu[6], Shizhu Gao[7], Li Jin[3]* and Yinqiu Cui[1]**

[1] School of Life Sciences, Jilin University, Changchun, China, [2] Max Planck Institute for the Science of Human History, Jena, Germany, [3] State Key Laboratory of Genetic Engineering, School of Life Sciences, and Human Phenome Institute, Fudan University, Shanghai, China, [4] College of Life Science, Jilin Normal University, Siping, China, [5] School of Archaeology, Jilin University, Changchun, China, [6] Xinjiang Cultural Relics and Archaeology Institute, Urumchi, China, [7] School of Pharmaceutical Sciences, Jilin University, Changchun, China

OPEN ACCESS

Edited by:
Horolma Pamjav,
Hungarian Institute for Forensic Sciences, Hungary

Reviewed by:
Balazs Egyed,
Eötvös Loránd University, Hungary
Guanglin He,
Nanyang Technological University, Singapore

***Correspondence:**
Shizhu Gao
gaosz@jlu.edu.cn
Li Jin
lijin@fudan.edu.cn
Yinqiu Cui
cuiyq@jlu.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: 12 July 2021
Accepted: 20 August 2021
Published: 22 September 2021

Citation:
Ning C, Zheng H-X, Zhang F, Wu S, Li C, Zhao Y, Xu Y, Wei D, Wu Y, Gao S, Jin L and Cui Y (2021) Ancient Mitochondrial Genomes Reveal Extensive Genetic Influence of the Steppe Pastoralists in Western Xinjiang. Front. Genet. 12:740167. doi: 10.3389/fgene.2021.740167

The population prehistory of Xinjiang has been a hot topic among geneticists, linguists, and archaeologists. Current ancient DNA studies in Xinjiang exclusively suggest an admixture model for the populations in Xinjiang since the early Bronze Age. However, almost all of these studies focused on the northern and eastern parts of Xinjiang; the prehistoric demographic processes that occurred in western Xinjiang have been seldomly reported. By analyzing complete mitochondrial sequences from the Xiabandi (XBD) cemetery (3,500–3,300 BP), the up-to-date earliest cemetery excavated in western Xinjiang, we show that all the XBD mitochondrial sequences fall within two different West Eurasian mitochondrial DNA (mtDNA) pools, indicating that the migrants into western Xinjiang from west Eurasians were a consequence of the early expansion of the middle and late Bronze Age steppe pastoralists (Steppe_MLBA), admixed with the indigenous populations from Central Asia. Our study provides genetic links for an early existence of the Indo-Iranian language in southwestern Xinjiang and suggests that the existence of Andronovo culture in western Xinjiang involved not only the dispersal of ideas but also population movement.

Keywords: mitochondrial genome, ancient DNA, Eurasian Steppe, Silk Road, Andronovo

INTRODUCTION

Recent archaeogenetic studies showed that the expansion of western steppe herders (WSHs) had a marked impact on the demographic, cultural, social and linguistic development since the third millennium BCE on the Eurasian continent (Allentoft et al., 2015; Haak et al., 2015; Damgaard et al., 2018a; Jeong et al., 2018, 2020; Narasimhan et al., 2019; Wang C. C. et al., 2021). One of the earliest representatives, known as the Yamnaya culture (ca. 3,300–2,700 BCE) from the Pontic–Caspian steppe migrated into Europe and Asia, bringing with them metallurgy, animal herding skills, and possibly the Indo-European languages (Frachetti, 2009; Allentoft et al., 2015; Haak et al., 2015). By the middle and late Bronze Age, the Sintashta culture (ca. 2,200–1,800 BCE) arose near the Urals and succeeded a majority of ancestry from the preceding Yamnaya culture. It carried a similar

genetic profile with the Srubnaya and the Andronovo cultures that spread over a large part of the Eurasia landmass, extending westward into Europe, southward into Central Asia and the India subcontinent, and eastward into the Mongolian Plateau (Allentoft et al., 2015; Haak et al., 2015; Damgaard et al., 2018b; Narasimhan et al., 2019; Jeong et al., 2020; Wang C. C. et al., 2021). A number of studies provided the evidence that the steppe cultures from western Eurasia had also integrated into the early Bronze Age cultures of western China. A recent archaeobotanical study showed that both wheat and barley had already spread to the Altai Mountains as early as 5,200 years ago (Zhou et al., 2020). Additionally, domesticated sheep and cattle were also observed in the prehistoric cultures of northwestern China (e.g., Majiayao culture, 3,550–2,850 BC; Qijia, 2,450–1,650 BC) (Fu et al., 2009). The cultural influences from WSHs suggested that ancient mobile pastoralists had played an extremely significant role in the prehistoric *trans*-Eurasian exchanges and the formation of agropastoralism.

Located at the intersection of the ancient "Silk Road," Xinjiang has played an important role in bridging the exchanges of cultures, goods, languages, and population movements (Wood, 2002). A recent genome-wide study on 951 Uyghurs in Xinjiang revealed a complex demographic history of the present-day populations in this region. Four major ancestral components were identified, namely, European, South Asian, Siberian, and East Asian (Feng et al., 2017). Two waves of admixtures were further characterized, with the first wave dating back as early as 3,750 years ago (Feng et al., 2017). However, human populations always underwent frequent population migrations, admixtures, and replacements, which made it difficult to reflect the true ancestral components and population dynamics using extant population data alone. The high level of genetic diversity of present-day Xinjiang people was likely a result of recent admixture events. The opening of the "Silk Road" made the exchanges of different populations in Xinjiang more frequent than ever. In contrast, ancient DNA study has been proven to be a powerful tool to reconstruct human prehistory by providing direct tests on samples from a certain period. Previous genetic studies have delineated that modern and ancient Xinjiang populations had maternal genetic affinities with both the eastern and western Eurasians, displaying high genetic diversity and admixture (Yao et al., 2004; Li et al., 2010; Zhang et al., 2010; Zheng et al., 2017; Wang W. et al., 2021). A recent paleogenomic study on the Iron Age Shirenzigou individuals from the eastern Tianshan mountains further confirmed the previous observations and characterized that the West Eurasian ancestry was likely to be related to the Early Bronze Age steppe pastoralists such as Yamnaya and/or Afanasievo than the chronologically more recent Sintashta and Andronovo cultures (Ning et al., 2019). Wang W. et al. (2021) retrieved the whole mitochondrial genomes of ancient Xinjiang populations from the Bronze Age to Historic Era. Their results revealed that the Bronze Age Xinjiang populations had genetic affinities with Steppe-related and Northeastern Asian populations (Wang W. et al., 2021). All of the above studies had proven the very complex demographic landscape of the ancient Xinjiang populations.

However, all those ancient DNA studies of Xinjiang were confined to the northern and eastern parts of this region. Considering the large geographic range and diverse ecosystems of Xinjiang, such studies in western Xinjiang are in great need to gain a more comprehensive understanding of the prehistoric demography of Xinjiang populations. In recent decades, a number of cultural remains and archaeological sites in western Xinjiang, showing the traits that are characteristic of the middle and late Bronze Age Eurasian Steppe (Steppe_MLBA) cultures (e.g., Sintashta and Andronovo) (Shao and Zhang, 2019), were investigated. However, the stable isotope analysis of the Bronze Age Xiabandi (XBD) population provided direct evidence of wheat and millet consumption in the eastern part of the Pamir Plateau (Zhang et al., 2016), suggesting that the possible East–West cultural interactions and communications in westernmost Xinjiang can be dated to 1,500 BC. A craniometry study on individuals from the Liushui cemetery (∼2,950 BP) in western Xinjiang also showed that the population was already admixed between the East and West Eurasians but with the majority inherited from the former (Zhang et al., 2011). The above research presented a complex and confusing scenario of western Xinjiang. More genetic studies on ancient populations in this region will undoubtedly provide important clues to the issue.

In this study, we collected 15 ancient samples from the XBD cemetery, the earliest archaeological site excavated in western Xinjiang to the best of our knowledge. We then enriched and sequenced the complete mitochondrial genomes of the XBD individuals through designed target probes. By comparing the mitochondrial DNA (mtDNA) of the XBD individuals with that of ancient and extant Eurasians, we explored the early population movement in western Xinjiang.

MATERIALS AND METHODS

Archaeological Background, Sampling, and Sequencing

The XBD cemetery is located in the westernmost region of Xinjiang, adjacent to the eastern edge of the Pamir Plateau (**Figure 1A**). This region lies at the intersection of the southern and northern branches of the historical "Silk Road," making it an important melting place for populations from East and Central Asia, as well as those from the Eurasia steppe. The XBD cemetery was investigated by Xinjiang Cultural Relics and Archaeology Institute in 2003 and 2004. The whole cemetery can be divided into three phases, the earliest of which was dated to the Bronze Age (3,500–3,300 BP), and the remaining two phages were dated to Han-Tang (∼2,200–1,300 BP) and Ming-Qing dynasties (∼600–300 BP) (Wu, 2012). The excavations of the jars with contracting neck, the bowls, the trumpet-shaped earrings, as well as the wide band-shaped bracelets in the first phase suggest that the XBD cemetery belongs to the Andronovo culture (**Figure 1B**; Wu, 2012). The cemetery contained 92 burials from the Bronze Age, but only 27 human skeletons were excavated. We selected 15 well-preserved skulls and sampled the intact and sound teeth for genetic research (**Table 1**; **Supplementary Table 1A**). The permission for the use of the 15 Bronze Age samples of the

FIGURE 1 | (A) Geographic location and **(B)** representative archaeological excavations of the Xiabandi (XBD) cemetery.

TABLE 1 | Summary of XBD individuals included in this study.

Sample ID	Biological sex	Archaeological dating (BP)	Archaeological culture	Mean coverage	Contamination (%)	Haplogroup
XBD-M18	Female	3,500–3,300	Andronovo	1,578	1.3	R1b
XBD-40	Male	3,500–3,300	Andronovo	1,622	5.8	H6a1a
XBD-M24	Male	3,500–3,300	Andronovo	1,143	1.6	H5b
XBD-M29	Male	3,500–3,300	Andronovo	4,130	0.6	U4a1
XBD-M10	Male	3,500–3,300	Andronovo	187	2.8	U4c1
XBD-M9	Female	3,500–3,300	Andronovo	3,410	3.4	H6a1a
XBD-M36	Female	3,500–3,300	Andronovo	1,881	1.9	U1a1c1
XBD-36	Female	3,500–3,300	Andronovo	2,023	1.4	R1b1
XBD-M45	Male	3,500–3,300	Andronovo	897	2.6	H11b
XBD-M46	Female	3,500–3,300	Andronovo	383	2	HV14
XBD-37	Female	3,500–3,300	Andronovo	1,308	4.8	U2e2a1d
XBD-M48	Male	3,500–3,300	Andronovo	576	2.1	U2e3
XBD-M17	Male	3,500–3,300	Andronovo	812	2.5	I4a
XBD-M14	Male	3,500–3,300	Andronovo	1,084	4.1	U2e1
XBD-38	Male	3,500–3,300	Andronovo	581	1.1	T2a1b1

XBD, Xiabandi.

XBD cemetery was obtained from Xinjiang Cultural Relics and Archaeology Institute.

DNA was extracted from teeth powder (∼50 mg) with the method previously described (Ning et al., 2016). The libraries were prepared with the NEBNext Ultra DNA Library preparation kit (New England Biolabs, United Kingdom) following the manufacturer's protocol but with a 1:20 dilution of the adapter during the ligation step. The quality and concentration of the libraries were determined on an Agilent Bioanalyzer 2100 (Agilent Technologies, Palo Alto, CA, United States). Subsequently, targeted enrichment of the mtDNA was conducted with the MitoCap^{TM} kit (MyGenostics, Beijing, China). Sequencing was carried out on an Illumina HiSeq 2000 platform at Novogene Inc. (Beijing, China).

Sequence Mapping and Mitochondrial DNA Haplogroup Determination

Raw data was processed using EAGER v1.92.50 with default parameters, a pipeline specially designed to deal with ancient DNA data (Peltzer et al., 2016). Quality assessment was performed with FastQC software (Andrews, 2010). The adapters were trimmed with AdapterRemoval v2.2.0 with a minimum overlap of 1 bp and base quality larger than 20 (Schubert et al., 2016). Reads shorter than 30 bp were disregarded. BWA v0.7.12 was used to align the reads to the Revised Cambridge Reference Sequences (rCRS) with seed disabled (*-l 2000*). The duplicate reads were removed by the DeDup v0.12.1 (Peltzer et al., 2016). Ancient DNA deamination rates were calculated with MapDamage v2.0 (Jonsson et al., 2013). Single-nucleotide polymorphisms (SNPs) and insertions and deletions (INDELs) were called using SNVer-0.5.2 (Wei et al., 2011) and were checked by visual inspection. We used trimmed 10 bp at both 3′ and 5′ ends with TrimBam function in the BamUtils v1.0.13[1] to minimize the bias caused by ancient DNA deamination. The mitochondrial haplogroups were determined with Haplogrep2 (Weissensteiner et al., 2016) according to PhyloTree build 17 (Van Oven, 2015).

Analysis of Xiabandi Mitochondrial DNA Genomes

Haplogroup frequencies were estimated by simple counting. A principal component analysis (PCA) based on the frequencies of sub-haplogroups was performed with the R libraries "factoextra", "FactoMineR", and "ggplot2".

The coalescence time of each lineage was estimated using the ρ statistic-based method and the maximum likelihood (ML) method implemented in PAML software v4.9g (Yang, 2007) with the Soares rate for complete mitochondrial genomes (Soares et al., 2009). For the ρ-based method, the corresponding standard deviation (SD) was calculated following published methods (Saillard et al., 2000). With the knowledge of coalescence time of each haplogroup estimated by contemporary samples, a Bayesian method implemented in BEAST software v1.8.0 was used to infer the time of XBD samples (Drummond and Rambaut, 2007).

RESULTS

Mitochondrial DNA Authentication and Contamination Assessment

Strict contamination precautions for ancient DNA were taken, and wet lab works were carried out in a dedicated clean room facility specially designed for ancient DNA studies at Jilin University. All samples showed a short fragment length (55–90 base pairs) and postmortem chemical modifications at 3′ and 5′ ends that are expected for ancient DNA (Dabney et al., 2013). The contamination rates of those samples were further quantitatively evaluated with contamMix v1.0-10 (Fu et al., 2013). As a result, we obtained a low level of modern human DNA contaminations (<5%) with the exception of one individual to be 5.8% (**Table 1**), which restricted our analysis to the deaminated reads to rule out modern human contamination. By applying the above criteria, we confirmed the authentication of our ancient data.

Major Bronze Age Steppe Pastoralist Origin of the Xiabandi Mitochondrial Haplogroups

We obtained complete mitochondrial sequences to an average coverage of 187- to 4,130-fold across all 15 individuals sequenced in this study. A total of 14 different mitochondrial haplogroups belonging to five macro-haplogroups, such as U, H, T, R, and I, were observed (**Table 1**; **Supplementary Table 1A**). All these haplogroups commonly appear in present-day populations from Europe, Central Asia, and Central/Western steppe, and are uniformly considered to be of West Eurasian origin. Two out of the 14 haplogroups, namely, HV14 and U1a1c1, are prevalent in the extant Central and Western Asians (Palanichamy et al., 2015; Narasimhan et al., 2019; Shamoon-Pour et al., 2019). Haplogroup HV14 was present in two ancient individuals from Central Asia, one (3,000–2,200 BCE) from Turkmenistan and the other (2,100–1,800 BCE) from Uzbekistan (Narasimhan et al., 2019). Similarly, U1a1c1 was found in a historical individual (680–649 CE) from the Pontic steppe (Narasimhan et al., 2019) and the Bronze Age individuals from Iran (3,328–3,022 BCE) and Turkmenistan (2,500–1,700 BCE), who were associated with the bactria-margiana archaeological complex (BMAC). These results suggest that haplogroup HV14 and U1a1c1 are probably Central or Western Asian origin. The remaining 12 haplogroups (I4a, H6a1a, H5b, H11b, R1b, R1b1, T2a1b1, U2e1, U2e2a1d, U2e3, U4a1, U4c1), however, were detected in the Bronze Age steppe pastoralists, the ancient Xinjiang groups, and the prehistoric populations in Europe. For example, haplogroup U2e, the most abundant type in XBD (20%), was found in high frequency in the Sintashta (11.6%) and Andronovo (14.3%) populations. Haplogroup U4a1, which had a high frequency in the Andronovo population (19%), was also observed in one individual associated with the Afanasievo culture (Allentoft et al., 2015). Beyond their wide distributions in the Bronze Age steppe pastoralists, several haplogroups were detected in some pre-Bronze Age hunter-gatherers from the central steppe as well. For example, haplogroup R1b was identified in an Upper Paleolithic individual from the left bank of the Yenisei River dated to around 14,000 BP. In addition, haplogroups U4a1, R1b1, and U2e3 were observed in the Botai culture from northern Kazakhstan and in Eastern Europe hunter-gatherer (Mathieson et al., 2015; Fu et al., 2016; Mittnik et al., 2018). Notably, haplogroups I4a, R1b1, and U2e2a1d were found in individuals who were associated with the BMAC culture and dated to the beginning of this culture 451 in Central Asia. These earlier individuals shared the substratum with the BMAC group but harbored additional Bronze Age steppe pastoralist ancestry than the main BMAC group as evidenced by the autosomal data (Narasimhan et al., 2019). Genetic frequency-based principal component analysis (PCA) agrees with what we have observed in the mitochondrial haplogroup distributions that the XBD falls within the western Eurasian cluster (right) formed

[1]https://github.com/statgen/bamUtil

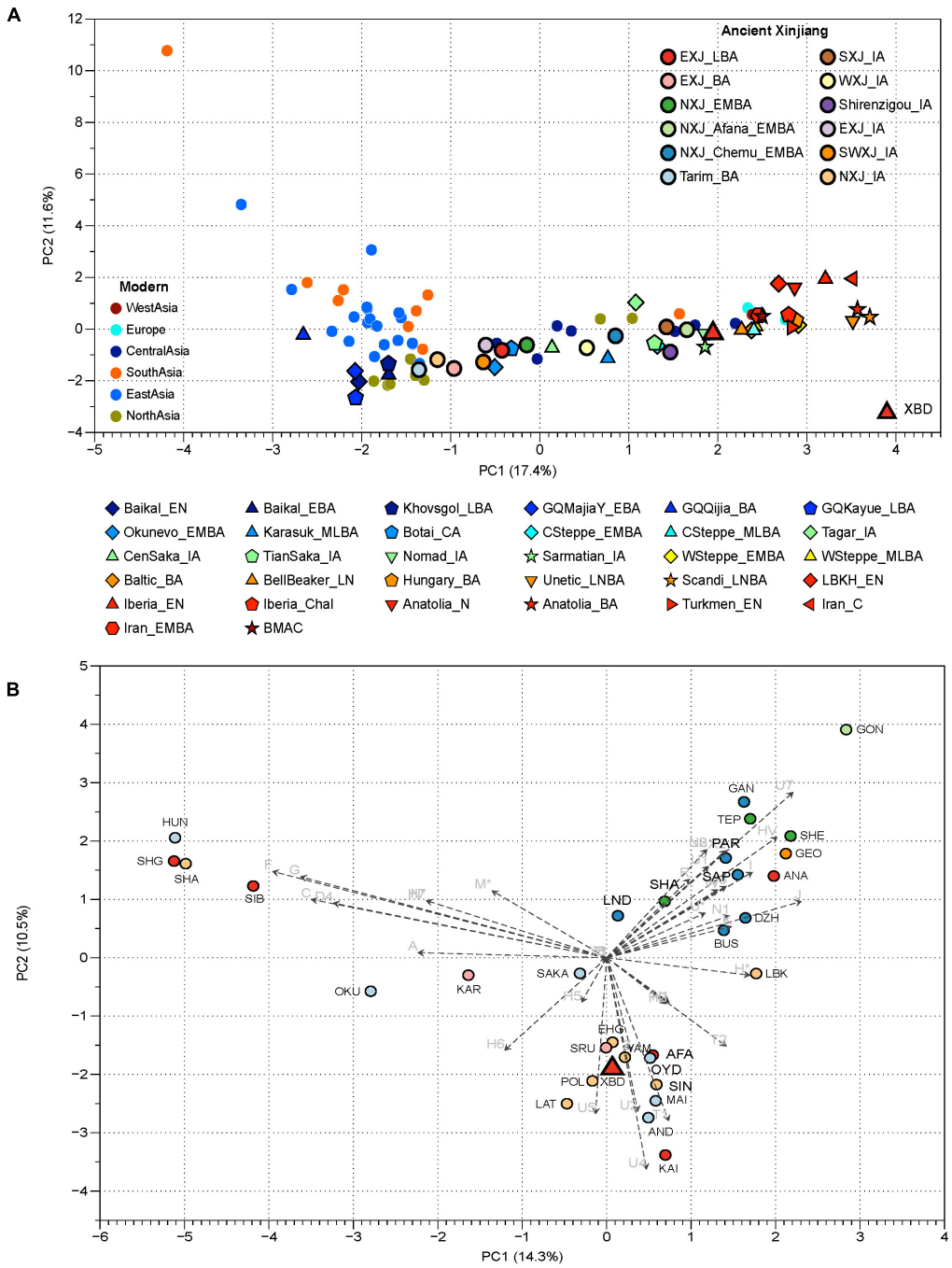


FIGURE 2 | Principal component analysis (PCA) based on mitochondrial DNA (mtDNA) haplogroup frequencies of ancient and present-day Eurasian populations. **(A)** The PCA constructed by populations from Xinjiang and neighboring regions. The different combinations of color and scheme on the plot represent different groups, present-day populations are marked with a circle; circles with dark black frames represent the published ancient Xinjiang samples, while shapes with light black frames are published ancient populations from the Eurasian continent. The single red triangle is the XBD populations in our study. NXJ, northern Xinjiang; EXJ, eastern Xinjiang; SXJ, southern Xinjiang; SWXJ, southwestern Xinjiang; WXJ, western Xinjiang. Detailed descriptions and references of comparative populations are provided in **Supplementary Tables 1B, C**. **(B)** The PCA constructed on ancient populations from the Eurasian continent. The two dimensions account for 25.8% of the total variance. Haplogroup contributions are represented by loading vectors marked as grep arrows. Population abbreviations are as follows: XBD, Xiabandi; SMK, Shamanka_EBA; ESHG, Scandinavian hunter-gatherers; ANA, Anatolia_N; LAT, Mesolithic/Neolithic Hunter gatherers from Lativ; LBK, LBK_EN; SIB, ancient Siberians; AFA, Afanasievo; KAI, Kairan_MLBA; AND, Andronovo; HUN, Tianshan_Hun; SAK, Tianshan_Saka; MAI, Maitan_MLBA; OKU, Okunevo; OYD, Oy_Dzhaylau_MLBA; POL, Poltavka; SIN, Sintashta; SRU, Srubnaya; YAM, Yamnaya; KAR, Karasuk; BUS, Bustan_BA; DZH, Dzharkutan_BA; SAP, Sappali_Tepe_BA; GEO, Geoksyur_EN; GAN, Ganj_Dareh; IND, Indus_Periphery_BA; PAR, Parkhai_EN; SHA, Shahr_I_Sokhta_BA; SEH, Seh_Gabi; TEP, Tepe_Hissar_CHL/EN; SHE, She_Gabi_ChL/EN. Detailed descriptions and references of comparative populations are provided in **Supplementary Table 1D**.

by the ancient nomads, WSteppe_EMBA and WSteppe_MLBA, represented by the Yamnaya and Andronovo, respectively (**Figure 2A**; **Supplementary Table 1F**). When compared to the other ancient populations from Xinjiang, we found that the XBD clustered with the NXJ_Afana_EMBA and Shirenzigou_IA, both of which were previously proven to share significant genetic affinity with the Bronze Age steppe pastoralists (Ning et al., 2019; Wang W. et al., 2021). In a finer scale PCA plot, the XBD also clustered with multiple WSH groups but shifted toward the South/Central Asian populations (top right) slightly (**Figure 2B**; **Supplementary Table 1E**), documenting that the majority of XBD mitochondrial haplogroups (12/14) can trace their origin from the Eurasia steppe pastoralist while the minority (2/14) from West or Central Asia.

Expansion of the Bronze Age Steppe Pastoralists as a Dynamic Process to Form the Genetic Landscape of Xiabandi Individuals

We used 540 present-day mitochondrial sequences obtained from PhyloTree database (van Oven and Kayser, 2009) who were genetically close to XBD individuals to construct the mtDNA phylogeny (**Supplementary Table 2**). Coalescence times of 14 mtDNA haplogroups related to XBD samples were estimated employing the ρ-based and the ML methods. The estimates obtained by both methods showed consistency (**Table 2**), suggesting the reliability of our estimates. Out of the 14 haplogroups, seven (U2e2a1d, I4a, U1a1c1, U4a1, U4c1, H6a1a,

TABLE 2 | TMRCA of XBD mtDNA haplogroups estimated from modern mtDNA genomes.

Haplogroup/sample	Coalescence time (kya)		
	ρ-based method[a]	ML method	Bayesian method[b]
I4a	6.89 ± 0.93	5.88 (4.05, 7.73)	–
S02914	–	–	5.25 (0.11, 7.71)
H6a1a	6.34 ± 0.66	5.94 (4.56, 7.33)	–
S02912	–	–	5.33 (0.54, 7.38)
S04760	–	–	4.63 (1.28, 7.33)
H5b	9.21 ± 1.71	9.20 (6.29, 12.16)	–
H5b + G263A + G7897A	7.22 ± 2.67	6.23 (2.13, 10.44)	–
R1769	–	–	4.38 (0.51, 8.62)
H11b	10.19 ± 2.81	9.26 (3.60, 15.11)	–
R1772	–	–	4.61 (0.00, 10.43)
HV14	13.70 ± 4.36	15.36 (5.43, 25.81)	–
R1773	–	–	14.66 (1.92, 22.25)
R1b + A9894G + C10160T + T16224C + G16390A	13.37 ± 4.13	14.18 (2.40, 26.74)	–
S02915	–	–	5.91 (0.00, 17.68)
R1b1	19.79 ± 3.43	20.95 (12.51, 29.72)	–
R1b1 + T16189C!	–	–	17.83 (7.81, 29.03)
S04757	–	–	12.27 (1.03, 26.14)
T2a1b1	12.22 ± 2.97	10.04 (7.12, 13.02)	–
S04759	–	–	3.77 (0.00, 7.65)
U1a1c1	13.91 ± 2.59	13.75 (8.18, 19.50)	–
R1767	–	–	11.29 (3.61, 17.59)
U2e1	15.21 ± 2.08	15.85 (12.36, 19.41)	–
U2e1 + C2526T + G12618A	7.89 ± 2.30	9.20 (3.09, 15.54)	–
S02913	–	–	6.08 (1.31, 11.40)
U2e2a1d	4.16 ± 1.08	4.47 (1.87, 7.12)	–
S04758	–	–	3.10 (0.00, 5.23)
U2e3	11.79 ± 3.65	11.37 (0.78, 22.63)	–
U2e3 + C16400T	11.99 ± 3.90	7.71 (1.54, 14.11)	–
U2e3 + C16400T + A3892G + T6050C + T10463C + T16017C + T16224C	–	–	3.23 (0.15, 9.59)
R1774	–	–	2.09 (0.00, 8.45)
U4a1	8.93 ± 1.47	9.44 (6.53, 12.41)	–
U4a1 + C7868T + A13773G	–	–	6.00 (2.70, 10.75)
R1770	–	–	4.07 (0.02, 9.85)
U4c1	6.09 ± 0.88	5.64 (3.81, 7.49)	–
U4c1 + A827G	–	–	5.56 (2.94, 7.68)
R0341	–	–	1.97 (0.00, 7.16)

[a] *We excluded Xiabandi and other ancient samples when using ρ-based method and ML method to estimate the coalescence time of each haplogroup.*
[b] *We used a Bayesian method to infer the time of Xiabandi samples with the knowledge of coalescence time of each haplogroup estimated by contemporary samples.*
ML, maximum likelihood; mtDNA, mitochondrial DNA; XBD, Xiabandi.

and U2e1) showed a star-like phylogeny of their ancestral node, indicating strong population expansions. Among the seven star-like lineages (**Figure 3**), four (I4a, H6a1a, U2e2a1d, and U4c1) were estimated of rather time to most recent common ancestor (TMRCA) of <6,000 BP with the most recent expansion lineage (U2e2a1d) estimated at approximately 4,470 BP (**Table 2**). This time is within the range of the presence of Early Bronze Age steppe pastoralists represented by the Yamnaya culture (3,300–2,700 BCE) in the Pontic steppe and the Afanasievo culture (3,300–2,500 BCE) in the Altai Mountains and fits well with the onset of the Sintashta culture (2,200–1,800 BCE). The Sintashta culture first emerged in the Urals at around 2,200 BC with multiple technological innovations, such as the earliest known chariots and training horses (Kristiansen and Larsson, 2005), and gave rise to the Andronovo culture (1,500–1,700 BC) (Kuznetsov, 2006; Hanks et al., 2007; Allentoft et al., 2015). Those innovations together with the populations quickly spread across much of the Eurasia Steppe (Narasimhan et al., 2019; Jeong et al., 2020). The genetic observations here, as well as the archaeological evidences, suggest that the XBD population originated in a large extent from the middle and late Bronze Age steppe pastoralists, who expanded to the western Xinjiang carrying their technologies along. However, two haplogroups (HV14 and U1a1c1) with Western or Central Asian origin were estimated of rather ancient TMRCA (14,660 and 11,290 BP, respectively) (**Table 2**), suggesting Western or Central Asian to be the source of these two haplogroups. A scenario to explain this phenomenon is that the Bronze Age steppe pastoralists expanded from the western and central steppe southward into Central Asia and admixed in a small scale with the indigenous populations there to form the ancestor of the XBD population, who then migrated eastward over the Pamir Plateau into western Xinjiang. This scenario is consistent with the recent ancient genomic study that the Bronze Age steppe pastoralists only marginally admixed with the indigenous population in Central Asia they met and moved farther southward into South Asia and admixed extensively with the local populations there (Narasimhan et al., 2019).

DISCUSSION

The prehistory of Xinjiang is of considerable interest given its special geographic location in connecting the East and the West Eurasians. Multiple genetic studies showed that since the Bronze Age, the populations in Xinjiang had exhibited high genetic diversity and extensive admixture with various populations (Yao et al., 2004; Zhang et al., 2010). The admixture dating analysis based on linkage disequilibrium for the present-day populations in Xinjiang suggested multiple waves of admixture events (Zhang et al., 2010; Shan et al., 2014a,b; Feng et al., 2017). However, tracing the population prehistory with present-day individuals is prone to be distorted by recent admixture events, which is especially the case for Xinjiang populations because the opening of the well-known "Silk Road" made the gene flow among different populations in this region even more frequent. Ancient DNA studies in this region had shown that populations in eastern Xinjiang were already admixed between the East and the West Eurasians as early as the Second Millennium BCE (Li et al., 2010; Wang W. et al., 2021). Population genetic history of western Xinjiang, however, is still largely unknown. By analyzing the XBD mitochondrial genomes, we show here that XBD was genetically admixed from the middle and late Bronze Age steppe pastoralists and the indigenous populations in Central Asia, who probably migrated into Xinjiang through the Pamir Plateau.

The discovery of the Tocharian manuscript from the northern rim of the Tarim Basin and the Indo-Iranian manuscripts from the southern edge provides direct evidence for the dispersal of Indo-European languages into the region (Di Cosmo, 2002). It is now a general consensus among the linguistics that the dispersal of both languages is related to the Bronze Age Steppe herders

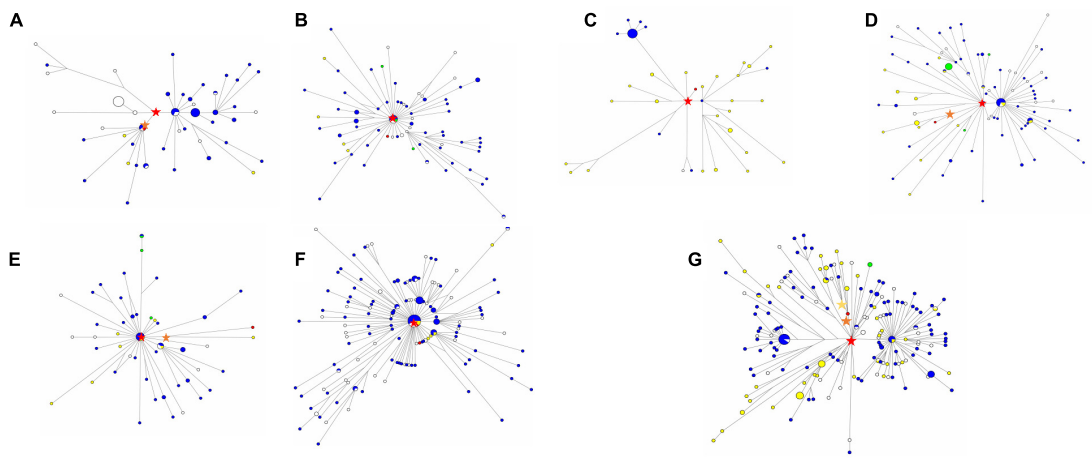


FIGURE 3 | Median-joining networks of star-like Xiabandi (XBD)-related mitochondrial DNA (mtDNA) haplogroups. The median-joining network of haplogroup U2e2a1d **(A)**, I4a **(B)**, U1a1c1 **(C)**, U4a1 **(D)**, U4c1 **(E)**, H6a1a **(F)**, and U2e1 **(G)** constructed from the complete mitochondrial sequences of both XBD and relevant present-day Eurasians. The red stars mark XBD individuals, the blue plots represent the Europeans, the green plots represent North Asians, the yellow plots represent West and East Asians, and the white plots represent the individuals that are of unknown location. The red stars are the original haplogroup, the orange star in **(A)** is haplogroup U2e2a1d, the orange star in **(E)** is haplogroup U4c1 + A827G, the orange star in **(D)** is haplogroup U4a1 + C7868T + A13773G, the orange star in **(G)** is haplogroup U2e1 + C2526T + G12618A, and the yellow star in **(G)** is haplogroup U2e1e.

(Walter, 1998). The Tocharians may have moved eastward earlier than the Indo-Iranians. The Tocharians are likely to be closely associated with the Afanasievo culture in the Altai Mountains who were a successor of the Yamnaya culture in the Pontic–Caspian Steppe. The middle and late Bronze Age steppe pastoralists, such as the Sintashta, Andronovo, and Srubnaya, are believed to be associated with the dispersal of Indo-Iranian languages (Lamberg-Karlovsky, 2002). The Iron Age individuals from northeastern Xinjiang were proved by autosomal DNA to harbor the Yamnaya/Afanasievo ancestry instead of the Steppe_MLBA, providing a strong genetic link of the "steppe hypothesis" over the "oasis hypothesis" and genetic support for the introduction of the Tocharian languages into Xinjiang (Ning et al., 2019). Our study here suggests a different genetic profile of totally west Eurasian origin, and provides a genetic link for the existence of Indo-Iranian languages in western Xinjiang at least 3,300 years ago.

CONCLUSION

Taken together, the systematic mtDNA analysis on ancient samples from the westernmost part of Xinjiang provides us a unique opportunity to investigate the population origin of Xinjiang with a broader geography. We find that the 15 XBD individuals fall within the range of the ancient western Eurasian variation, and the formation of the ancestry legacy of XBD is related to the expansion of the middle and late Bronze Age steppe herders who might speak Indo-Iranian languages and admixed with the indigenous populations in the West or Central Asia during their expansion. Additionally, integrating the archaeological and genetic evidences in this study, the existence of the Andronovo culture in western Xinjiang involved not only the dispersal of ideas but also population movement. We recognize that such study on samples from a broader region and time sequences is required to obtain a more comprehensive understanding of the population prehistory of Xinjiang.

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: The BIG Data Center Genome Sequence Archive (GSA) under accession number HRA001154 (http://bigd.big.ac.cn/gsa-human).

AUTHOR CONTRIBUTIONS

YC, LJ, and SG conceived and supervised the study. CN, YZ, YX, and CL performed the laboratory work. YW and DW provided archaeological materials and associated information. CN, H-XZ, FZ, and SW analyzed the data. CN, YC, FZ, SG, and H-XZ wrote the manuscript with the input from all co-authors.

FUNDING

This work was supported by the Major project of Humanities and Social Sciences Key Research Base of the Ministry of Education (16JJD780005), National Key R&D Program of China (2016YFE0203700), Scientific and Technological Developing Scheme of Jilin Province (20190701077GH) and National Natural Science Foundation of China (31401062 and 31271338).

ACKNOWLEDGMENTS

We thank Xinjiang Cultural Relics and Archaeology Institute, Urumchi, China, for their support in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.740167/full#supplementary-material

REFERENCES

Allentoft, M. E., Sikora, M., Sjogren, K. G., Rasmussen, S., Rasmussen, M., Stenderup, J., et al. (2015). Population genomics of Bronze Age Eurasia. *Nature* 522, 167–172. doi: 10.1038/nature14507

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., et al. (2013). Complete mitochondrial genome sequence of a middle pleistocene cave bear reconstructed from ultrashort DNA fragments. *PNAS* 110, 15758–15763. doi: 10.1073/pnas.1314445110

Damgaard, P. B., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., Korneliussen, T., et al. (2018a). 137 ancient human genomes from across the Eurasian steppes. *Nature* 557, 369–374. doi: 10.1038/s41586-018-0094-2

Damgaard, P. B., Martiniano, R., Kamm, J., Moreno-Mayar, J. V., Kroonen, G., Peyrot, M., et al. (2018b). The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* 360:eaar7711. doi: 10.1126/science.aar7711

Di Cosmo, N. (2002). *The Tarim Mummies: ancient China and the mystery of the earliest peoples from the West*. London: Thames & Hudson, 279–281.

Drummond, A. J., and Rambaut, A. (2007). BEAST: bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7:214. doi: 10.1186/1471-2148-7-214

Feng, Q., Lu, Y., Ni, X., Yuan, K., Yang, Y., Yang, X., et al. (2017). Genetic History of Xinjiang's Uyghurs Suggests Bronze Age Multiple-Way Contacts in Eurasia. *Mol. Biol. Evol.* 34, 2572–2582. doi: 10.1093/molbev/msx177

Frachetti, M. D. (2009). *Pastoralist Landscapes And Social Interaction In Bronze Age Eurasia*. Berkeley, California: University of California Press, doi: 10.1525/9780520942691

Fu, L., Yuan, J., and Li, S. (2009). The origin and character of domesticated animals during Neolithic Age in Gansu–Qinghai region. *Archaeology* 5, 80–86.

Fu, Q., Mittnik, A., Johnson, P. L. F., Bos, K., Lari, M., Bollongino, R., et al. (2013). A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* 23, 553–559. doi: 10.1016/j.cub.2013.02.044

Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., et al. (2016). The genetic history of Ice Age Europe. *Nature* 534, 200–205. doi: 10.1038/nature17993

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. doi: 10.1038/nature14317

Hanks, B. K., Epimakhov, A. V., and Renfrew, A. C. (2007). Towards a refined chronology for the Bronze Age of the southern Urals, Russia. *Antiquity* 81, 353–367. doi: 10.1017/S0003598X00095235

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. 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. U. S. A.* 115, E11248–E11255. doi: 10.1073/pnas.1813608115

Jonsson, H., Ginolhac, A., Schubert, M., Johnson, P. L., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29, 1682–1684. doi: 10.1093/bioinformatics/btt193

Kristiansen, K., and Larsson, T. B. (2005). *The Rise Of Bronze Age Society : travels, Transmissions And Transformations*. Cambridge; New York: Cambridge University Press.

Kuznetsov, P. F. (2006). The emergence of Bronze Age chariots in eastern Europe. *Antiquity* 80, 638–645. doi: 10.1017/S0003598X00094096

Lamberg-Karlovsky, C. C. (2002). Archaeology and language: the indo-iranians. *Current Anthropology* 43, 63–88. doi: 10.1086/324130

Li, C., Li, H., Cui, Y., Xie, C., Cai, D., Li, W., et al. (2010). Evidence that a West-East admixed population lived in the Tarim Basin as early as the early Bronze Age. *BMC Biol.* 8:15. doi: 10.1186/1741-7007-8-15

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

Mittnik, A., Wang, C. C., Pfrengle, S., Daubaras, M., Zarina, G., Hallgren, F., et al. (2018). The genetic prehistory of the Baltic Sea region. *Nat. Commun.* 9:442. doi: 10.1038/s41467-018-02825-9

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. doi: 10.1126/science.aat7487

Ning, C., Gao, S., Deng, B., Zheng, H., Wei, D., Lv, H., et al. (2016). Ancient mitochondrial genome reveals trace of prehistoric migration in the east Pamir by pastoralists. *J. Hum. Genet.* 61, 103–108. doi: 10.1038/jhg.2015.128

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

Palanichamy, M. G., Mitra, B., Zhang, C. L., Debnath, M., Li, G. M., Wang, H. W., et al. (2015). West Eurasian mtDNA lineages in India: an insight into the spread of the Dravidian language and the origins of the caste system. *Hum. Genet.* 134, 637–647. doi: 10.1007/s00439-015-1547-4

Peltzer, A., Jager, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., et al. (2016). EAGER: efficient ancient genome reconstruction. *Genome Biol.* 17:60. doi: 10.1186/s13059-016-0918-z

Saillard, J., Forster, P., Lynnerup, N., Bandelt, H. J., and Norby, S. (2000). mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am. J. Hum. Genet.* 67, 718–726. doi: 10.1086/303038

Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* 9:88. doi: 10.1186/s13104-016-1900-2

Shamoon-Pour, M., Li, M., and Merriwether, D. A. (2019). Rare human mitochondrial HV lineages spread from the Near East and Caucasus during post-LGM and Neolithic expansions. *Sci. Rep.* 9:14751. doi: 10.1038/s41598-019-48596-1

Shan, W., Ablimit, A., Zhou, W., Zhang, F., Ma, Z., and Zheng, X. (2014a). 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

Shan, W., Ren, Z., Wu, W., Hao, H., Abulimiti, A., Chen, K., et al. (2014b). Maternal and paternal diversity in Xinjiang Kazakh population from China. *Genetika* 50, 1374–1385. doi: 10.7868/s0016675814110149

Shao, H., and Zhang, W. (2019). Review of the Research of Andronovo Cultures in Xinjiang (in Chinese). *Western Reg. Stud.* 2, 113–121. doi: 10.16363/j.cnki.xyyj.2019.02.012

Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Rohl, A., et al. (2009). Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am. J. Hum. Genet.* 84, 740–759. doi: 10.1016/j.ajhg.2009.05.001

Van Oven, M. (2015). PhyloTree Build 17: growing the human mitochondrial DNA tree. *Forensic Sci. Int. Genet. Suppl. Ser.* 5, e392–e394. doi: 10.1016/j.fsigss.2015.09.155

van Oven, M., and Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* 30, E386–E394. doi: 10.1002/humu.20921

Walter, M. N. (1998). *Tokharian Buddhism In Kucha: buddhism Of Indo-European Centum Speakers In Chinese Turkesta Before The 10th Century Ce*. United States: University of Pennsylvania.

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, W., Ding, M., Gardner, J. D., Wang, Y., Miao, B., Guo, W., et al. (2021). Ancient Xinjiang mitogenomes reveal intense admixture with high genetic diversity. *Sci. Adv.* 7:eabd6690. doi: 10.1126/sciadv.abd6690

Wei, Z., Wang, W., Hu, P., Lyon, G. J., and Hakonarson, H. (2011). SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.* 39:e132. doi: 10.1093/nar/gkr599

Weissensteiner, H., Pacher, D., Kloss-Brandstatter, A., Forer, L., Specht, G., Bandelt, H. J., et al. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* 44, W58–W63. doi: 10.1093/nar/gkw233

Wood, F. (2002). *The Silk Road: two thousand years in the heart of Asia*. Berkeley, California: Univ of California Press.

Wu, Y. (2012). On the Bronze Culture of Xiabandi Cemetery in Kashi, Xinjiang (in Chinese). *Western Reg. Stud.* 4, 36–44. doi: 10.16363/j.cnki.xyyj.2012.04.013

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

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, 2265–2280. doi: 10.1093/molbev/msh238

Zhang, F., Xu, Z., Tan, J., Sun, Y., Xu, B., Li, S., et al. (2010). Prehistorical East-West admixture of maternal lineages in a 2,500-year-old population in Xinjiang. *Am. J. Phys. Anthropol.* 142, 314–320. doi: 10.1002/ajpa.21237

Zhang, J., Wu, X., Li, L., Jin, L., Li, H., and Tan, J. (2011). Cranial Non-metric Evidence for Population Admixture Between East and West Eurasia in Bronze Age,Southwestern Xinjiang (in Chinese). *Renleixue Xuebao* 30, 379–404. doi: 10.16359/j.cnki.cn11-1963/q.2011.04.006

Zhang, X., Wei, D., Wu, Y., Nie, Y., and Hu, Y. (2016). Carbon and nitrogen stable isotope ratio analysis of Bronze Age humans from the Xiabandi cemetery, Xinjiang, China: implications for cultural interactions between the East and West. *Chin. Sci. Bull.* 61, 3509–3519. doi: 10.1360/n972016-00514

Zheng, H. X., Li, L., Jiang, X. Y., Yan, S., Qin, Z., Wang, X., et al. (2017). MtDNA genomes reveal a relaxation of selective constraints in low-BMI individuals in a Uyghur population. *Hum. Genet.* 136, 1353–1362. doi: 10.1007/s00439-017-1829-0

Zhou, X., Yu, J., Spengler, R. N., Shen, H., Zhao, K., Ge, J., et al. (2020). 5,200-year-old cereal grains from the eastern Altai Mountains redate the trans-Eurasian crop exchange. *Nat. Plants* 6, 78–87. doi: 10.1038/s41477-019-0581-y

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: 30 September 2021
doi: 10.3389/fgene.2021.754492

Genomic Insight Into the Population Admixture History of Tungusic-Speaking Manchu People in Northeast China

Xianpeng Zhang[1†], Guanglin He[2,3,4,5†], Wenhui Li[1], Yunfeng Wang[6], Xin Li[1], Ying Chen[1], Quanying Qu[1], Ying Wang[1], Huanjiu Xi[1], Chuan-Chao Wang[2,3,4] and Youfeng Wen[1]**

[1] Institute of Biological Anthropology, Jinzhou Medical University, Jinzhou, China, [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] School of Humanities, Nanyang Technological University, Singapore, Singapore, [6] Xinbin Manchu Autonomous County People's Hospital, Fushun, China

OPEN ACCESS

Edited by:
Horolma Pamjav,
Hungarian Institute for Forensic Sciences, Hungary

Reviewed by:
Atif Adnan,
China Medical University, China
Ranajit Das,
Yenepoya University, India

***Correspondence:**
Youfeng Wen
wenyf@jzmu.edu.cn
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: 06 August 2021
Accepted: 30 August 2021
Published: 30 September 2021

Citation:
Zhang X, He G, Li W, Wang Y, Li X, Chen Y, Qu Q, Wang Y, Xi H, Wang C-C and Wen Y (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

Manchu is the third-largest ethnic minority in China and has the largest population size among the Tungusic-speaking groups. However, the genetic origin and admixture history of the Manchu people are far from clear due to the sparse sampling and a limited number of markers genotyped. Here, we provided the first batch of genome-wide data of genotyping approximate 700,000 single-nucleotide polymorphisms (SNPs) in 93 Manchu individuals collected from northeast China. We merged the newly generated data with data of publicly available modern and ancient East Asians to comprehensively characterize the genetic diversity and fine-scale population structure, as well as explore the genetic origin and admixture history of northern Chinese Manchus. We applied both descriptive methods of ADMIXTURE, fineSTRUCTURE, F_{ST}, TreeMix, identity by decedent (IBD), principal component analysis (PCA), and qualitative f-statistics (f_3, f_4, qpAdm, and qpWave). We found that Liaoning Manchus have a close genetic relationship and significant admixture signal with northern Han Chinese, which is in line with the cluster patterns in the haplotype-based results. Additionally, the qpAdm-based admixture models showed that modern Manchu people were formed as major ancestry related to Yellow River farmers and minor ancestry linked to ancient populations from Amur River Bain, or others. In summary, the northeastern Chinese Manchu people in Liaoning were an exception to the coherent genetic structure of Tungusic-speaking populations, probably due to the large-scale population migrations and genetic admixtures in the past few hundred years.

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

INTRODUCTION

The Manchu is the third-largest ethnic minority with a population size of over 10 million in China; and they mainly live in Liaoning, Hebei, Heilongjiang, and other northern provinces. Liaoning Province was the traditional homeland of the Manchu people, and the first capital of the Qing Dynasty was located there. There are more than 5 million Manchus in Liaoning, accounting for

more than 50% of the total population of Manchu. The term "Manchu" can be dated back to the 16th century, but the history of Manchu can be traced back further. According to early historic records, the Manchu ancestors were known as Donghu, a tribal confederation of nomadic people that was first recorded from the seventh century BCE. After then, the history of Manchu had been associated with many ancient tribes that once lived in this region during different periods, such as Sushen, Yilou, Wuji, and Mohe. In the late historic period, there are two dynasties associated with the Manchus: the first one was Jin Dynasty[1] (1115–1234 AD) founded by Jurchen, and the second was the Qing Dynasty[2] (1636–1912 AD) founded by ancient Manchu people. After the Jin Dynasty was annihilated by the Mongols in 1234 AD, the surviving Jurchen people gradually developed into Manchu. When Manchu regained control of Manchuria, they moved further south and gradually controlled all sections of China, and they were suggested to have left detectable genetic imprints on the modern north and northeast Chinese, especially in the Qing Dynasty (Xue et al., 2005; Zhao et al., 2011; He and Guo, 2013).

Previous genetic studies of the Manchu population were predominantly based on very limited numbers of forensic markers, such as short tandem repeats (STRs) and single-nucleotide polymorphisms (SNPs) on the autosome (Liu et al., 2013), Y-chromosome (He and Guo, 2013), X-chromosome (Xing et al., 2019), and mtDNA (Zhao et al., 2011). The autosomal STR study of Manchu suggested that there were only small genetic distances between the Liaoning Manchus and Qinghai and Liaoning Hans (Liu et al., 2013). From a paternal Y chromosomal perspective, there were no significant differences in the haplotype composition between Liaoning Manchus and Northern Hans or Chinese Mongolians, and Manchus displayed a very typical East Asian affinity. Besides, there were only minor differences between Manchus and East Asian populations such as the Southern Han population, Chinese Korean, Japanese, and South Korean (He and Guo, 2013); Manchus shared similar Y-haplotypes with Xibe, Outer Mongolians, Inner Mongolians, Ewenki, Oroqen, and Hezhen (Xue et al., 2005). The investigation of the Y-chromosomal profile of the Aisin Gioro clan who are the Qing Dynasty nobility found that Manchus might be descendants of ancient populations in the Transbaikal region (Yan et al., 2015; Wei et al., 2017; Wang et al., 2019). The X-chromosomal profile further showed that there was an affinity between the Liaoning Manchus and Liaoning Koreans and Hans from Henan and Shanghai (Xing et al., 2019). From a maternal mtDNA perspective, Liaoning Manchus displayed an admixture signal between northern and southern East Asians, and they had a close genetic affinity with neighboring populations, such as the Mongolians, Liaoning Hans, and Korean (Zhao et al., 2011). The genetic distances between the Manchus and the Altaic language-speaking populations such as Hezhen, Daur, and Oroqen were smaller than those of other language-speaking populations (Zhao et al., 2011; Liu et al., 2013; Xing et al., 2019). Most genetic investigations based on the high-density genome-wide genetic variations from non-Altaic people in East Asia have revealed a fine-scale genetic landscape of genetic diversity and population admixture among the populations from different-language families (He et al., 2021a,c; Liu et al., 2021a,b; Wang et al., 2021d; Yao et al., 2021). Besides, recent genome-wide studies among Altaic-speaking populations in Northeast Asia have also found differentiated genetic admixture profiles between northern and southern Altaic-speaking populations and eastern and western Mongolians (He et al., 2021b). The reconstructed demographic models using ancient genomes further showed the Tungusic people keep a strong genetic homogeneity within populations, and the type of Tungusic-dominant ancestry probably originated in a vast region from Mongolian Plateau to Amur River Basin about at least 16,000 years ago (He et al., 2021b; Mao et al., 2021). However, there are few genome-wide SNP data from Tungusic-speaking Manchu people reported so far.

Manchu language belongs to the Tungusic group of the Altaic language family, but previous genetic studies indicated that Manchus were genetically different from other Tungusic-speaking groups, which is probably due to the genetic influence from surrounding Han Chinese into Manchus (Xue et al., 2005; He and Guo, 2013). Chen et al. recently reported that Guizhou Manchus in southwest China had a strong genetic affinity with southern East Asians and found that Guizhou Manchus could be modeled as deriving a large proportion of southern ancestry related to Austronesian, Tai-Kadai, and Austroasiatic speakers, suggesting that Manchu gradually mixed with the southern natives along with their southward migration (Chen et al., 2021). Genetic diversity and genetic admixture scenarios of northern East Asian Altaic people were mainly collected from Mongolic and Tungusic people (Jeong et al., 2019; Wang et al., 2021b,c). The genetic structure, population origin, and admixture history of Manchus due to the paucity of genome-wide data from northeast China—the origin center of ancient Manchu people—are now far from clear. In this study, we reported for the first time the genome-wide SNP data of 93 Manchu individuals who have lived in the Xinbin Manchu Autonomous County (the location of Hetu Ala city), Liaoning Province in northeast China. Our aim was to comprehensively infer the genetic origin and explore the population admixture history of the Manchu people by coanalyzing both modern and ancient Eurasian genomes.

MATERIALS AND METHODS

Sampling and Genotyping

We collected a total of 93 peripheral blood samples from unrelated Manchu individuals aged over 58 in different villages in the Xinbin Manchu Autonomous County, Liaoning Province, northeast China (**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

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

[2]https://en.wikipedia.org/wiki/Qing_dynasty

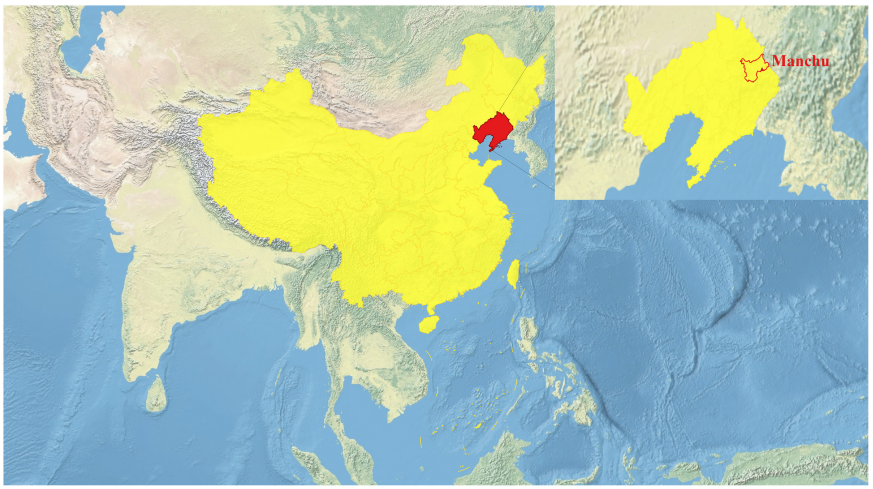


FIGURE 1 | Map of the sampling location for this study (https://www.qgis.org/).

history and corresponding family records. Our study and sample collection were reviewed and approved by the Ethics Committee of Jinzhou Medical University (JZMULL2021010) and followed the recommendations provided by the revised Declaration of Helsinki of 2000. Verbal and written informed consent was obtained from all participants. We used the Infinium® Global Screening Array (GSA) covering 659,509 SNPs to genotype targeted SNPs in Manchus. Genotype calling was carried out following the default parameters. Raw data were initially filtered using PLINK 1.9 (Purcell et al., 2007) based on our predefined threshold of the genotyping success rate, missing site rates, minor allele frequency, and Hardy–Weinberg equilibrium (−maf:0.01, −hwe:0.0001, −mind:0.01, and −geno:0.01). A final dataset with 293,307 SNPs was used to perform the following population genetic analysis.

Data Merging

We merged our newly genotyped data with previously published modern and ancient population data from the Affymetrix Human Origins (HO) Array dataset (Wall and Yoshihara Caldeira Brandt, 2016) and the 1240K dataset from the Allen Ancient DNA Resource (AADR).[3] The included modern genome-wide SNP reference data were collected from nine language families or groups (Tungusic, Mongolic, Turkic, Sinitic, Tibeto-Burman, Austronesian, Austroasiatic, Tai-Kadai, and Hmong-Mien) in China, South Siberia, and Southeast Asia (Patterson et al., 2012; Lipson et al., 2018; Liu et al., 2020; Kutanan et al., 2021; Wang et al., 2021a). Ancient reference populations were also collected from China and surrounding countries (Ning et al., 2020; Yang et al., 2020; Wang et al., 2021a). We also included recently published population data from the Guizhou Manchu people (Chen et al., 2021). Finally, two combined datasets covering 44,476 SNPs in the merged HO set and 127,435 SNPs in the merged 1240K set were used in the subsequent analysis.

[3]https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data

Principal Component Analysis and Admixture

Principal component analysis (PCA) was carried out using *smartpca*, part of the EIGENSOFT package (Patterson et al., 2006). The additional parameter of the lsqproject was set as YES (lsqproject: YES), and other parameters were used as the default setting. A total of 70 present-day and ancient worldwide populations were selected for PCA. Ancient people were projected onto the background of modern genetic variations. The PCA result was plotted by R software.[4] We applied ADMIXTURE (Alexander et al., 2009) to conduct the model-based clustering analysis based on the merged HO dataset, which consists of 1,385 individuals from 89 populations. Prior to the analysis, we pruned SNPs in strong linkage disequilibrium with each other using PLINK 1.9 (Purcell et al., 2007) with the parameters "-indep-pairwise 200 25 0.4." We run ADMIXTURE with the K values (number of assumed ancestral sources) ranging from 2 to 20. An optimal value of K was selected using 10-fold cross-validation errors, which is shown in **Supplementary Table 1**. The results of admixture analysis were visualized using AncestryPainter (Feng et al., 2018).

F_{ST} Analyses

We calculated pairwise F_{ST} genetic distance between Liaoning Manchu and other included modern and ancient reference populations using *smartpca* of EIGENSOFT package (Patterson et al., 2006) (fstonly: YES, fsthiprecision: YES).

f-Statistics

All *f*-statistics were calculated using ADMIXTOOLS (Patterson et al., 2012). We computed outgroup f_3-statistics in the form of f_3 (Manchu_Liaoning, X; Mbuti) to examine the shared genetic drift between Liaoning Manchus and non-African reference populations. Central African of Mbuti was used as the outgroup.

[4]https://www.r-project.org/

We also used the f_3-test in the form of f_3 (source1, source2; Manchu_Liaoning) to formally test whether there was an admixture signature in Liaoning Manchus with different source pairs from modern and ancient eastern Eurasians. Negative values with the absolute Z-score of less than −3 indicated the included two source-related populations may be the plausible ancestral sources. We applied the f_4-test in the forms of f_4 (X, Y; Manchu_Liaoning, Mbuti) and f_4 (X, Manchu_Liaoning; Y, Mbuti) to estimate the differentiated allele sharing between Liaoning Manchus and other representative sources compared with focused comparative subjects, where X and Y are the included ancient and present-day populations. The tested form f_4 (X, Y; Manchu_Liaoning, Mbuti) was designed to examine the differentiated genetic ancestry between Manchus and X or Y, in which significantly negative values indicate more shared alleles between Y and Manchus compared with X, significantly positive values indicated more shared alleles between X and Manchus relative to Y, and nonsignificant values (Z-scores ranged from −3 to 3) indicated that X and Y formed one clade compared with the Manchus (Patterson et al., 2012).

qpAdm and *qpWave*

We used formal tests of the *qpWave*/*qpAdm* programs in ADMIXTOOLS (Patterson et al., 2012) to determine the minimum number of streams of potential source populations contributing to the tested populations and also estimate the admixture proportions. We used the following eight outgroups including Atayal, Mbuti, Papuan, French, DevilsCave, Jomon, Malaysia_LN, and Tianyuan, which are unlikely to have been affected by recent gene flow with proposed ancestral sources and might be differentially related to the tested populations. We chose the included outgroups as the distant outgroups to dissect the northern and southern East Asian ancestries that participated in the formation of modern Manchus based on recent modern and ancient genetic studies (Chen et al., 2019; He et al., 2020; Wang et al., 2021a).

TreeMix and ALDER

To explore the genetic relationship between the Manchu population and other references East Asians, we used the TreeMix v1.13 (Pickrell and Pritchard, 2012) program to construct the maximum likelihood trees with variable predefined mixture events. The level and time of admixture events were estimated by using ALDER v.1.0.3 (Wu, 2020).

Identity by Decedent and Chromosome Painting

We used shapeit v2 to phase successive SNPs into haplotype data and following used refined identity by decedent

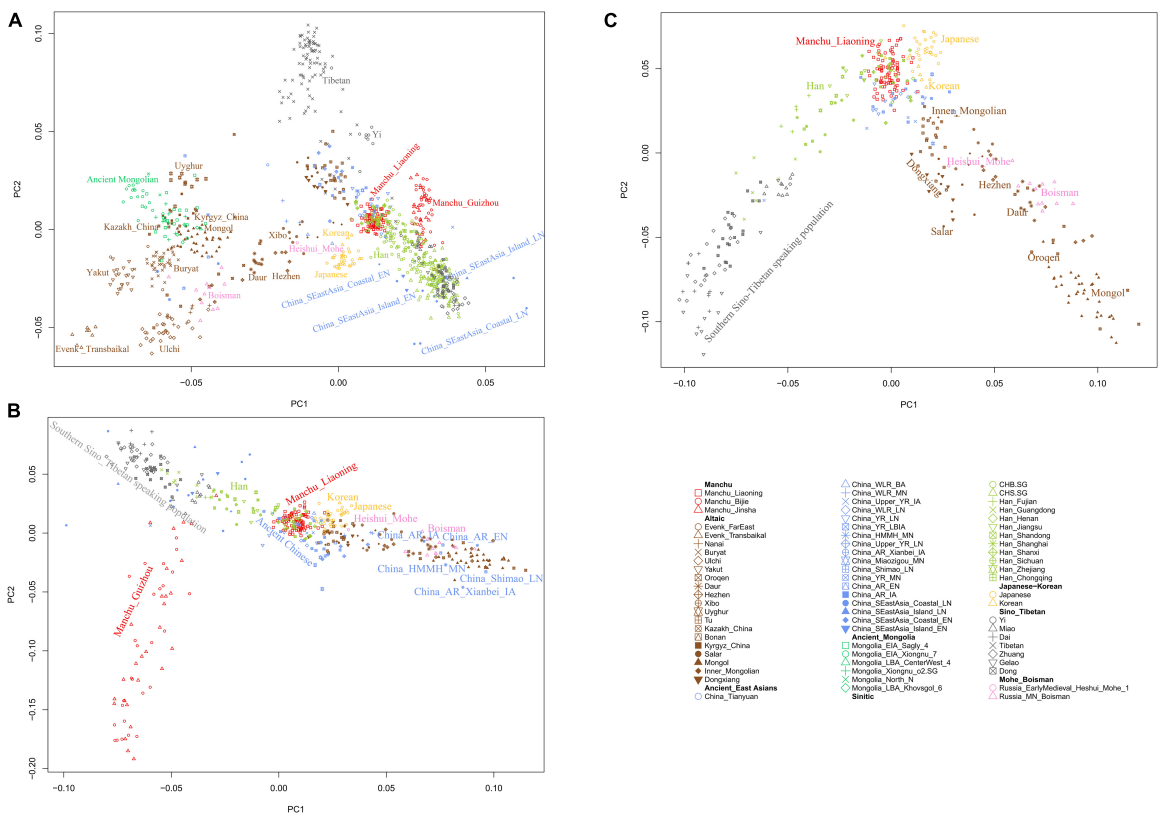


FIGURE 2 | Patterns of genetic relationship among East Asian populations and Liaoning Manchus inferred from the principal component analysis **(A–C)**. Modern populations were color-coded on the basis of their language family categories, and different shapes represented different populations. All ancient populations were projected onto it.

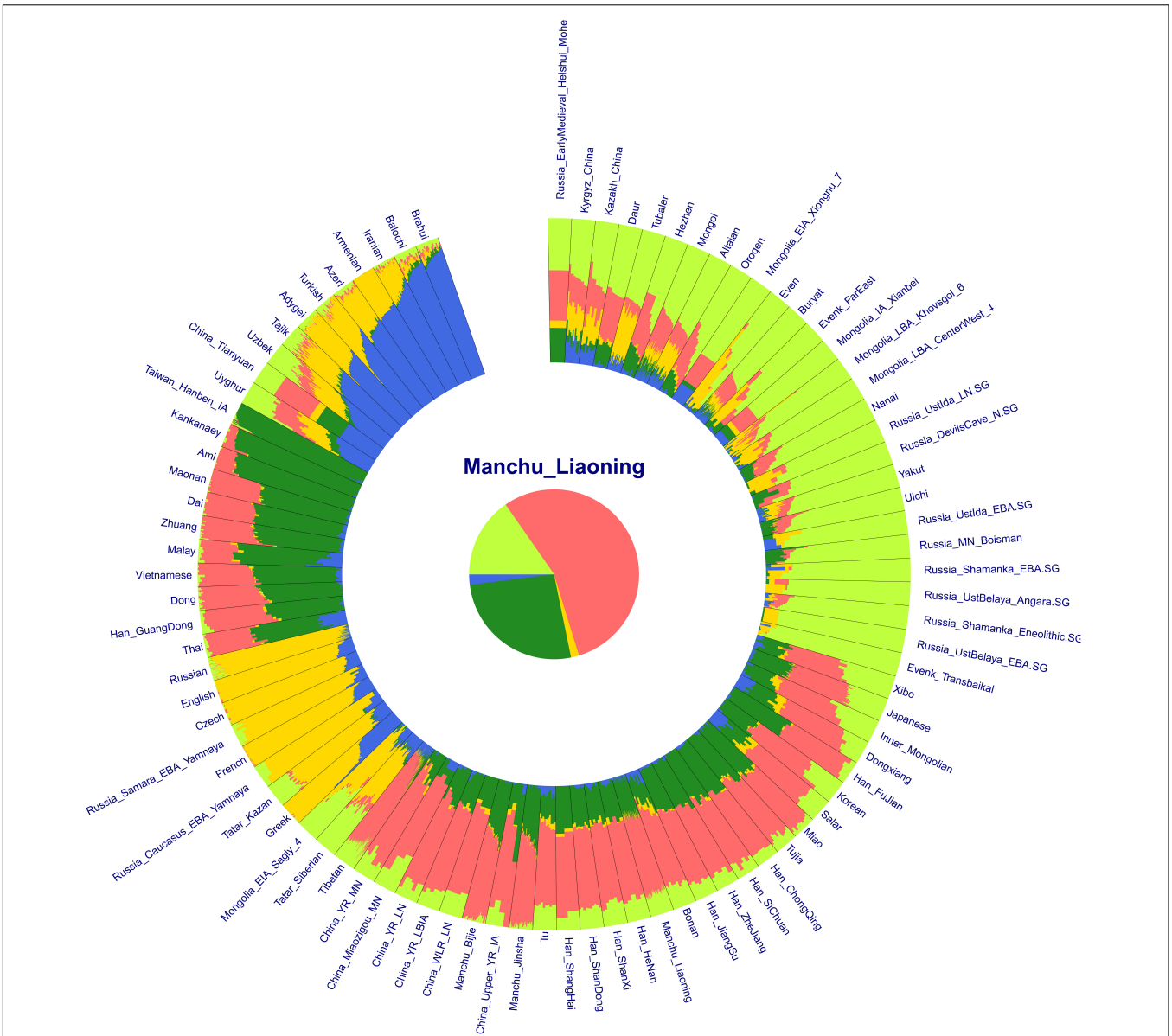


FIGURE 3 | Results of model-based ADMIXTURE clustering analysis. Clustering patterns were visualized with the predefined ancestral sources at $K = 5$. All of these ancestral components were revealed by different colors.

(IBD) to calculate the pairwise shared IBD (Browning and Browning, 2011). ChromoPainter (Lawson et al., 2012) and fineSTRUCTURE (Lawson et al., 2012) were further used to explore the fine-scale population structure based on the sharing haplotypes.

Y Chromosomal and mtDNA Lineages

The mtDNA haplogroups were classified using HaploGrep2 (Weissensteiner et al., 2016) based on PhyloTree17[5], and we used in-house scripts to assign the Y-chromosomal paternal lineage following the basic regulations reaccommodated via the International Society of Genetic Genealogy (2018)[6].

RESULTS

Principal Component Analysis and Admixture

To understand the general pattern of the genetic structure of Liaoning Manchus, we first performed PCA to explore the two-dimensional genetic relationship between Manchus

[5]https://www.phylotree.org/tree

[6]https://isogg.org/

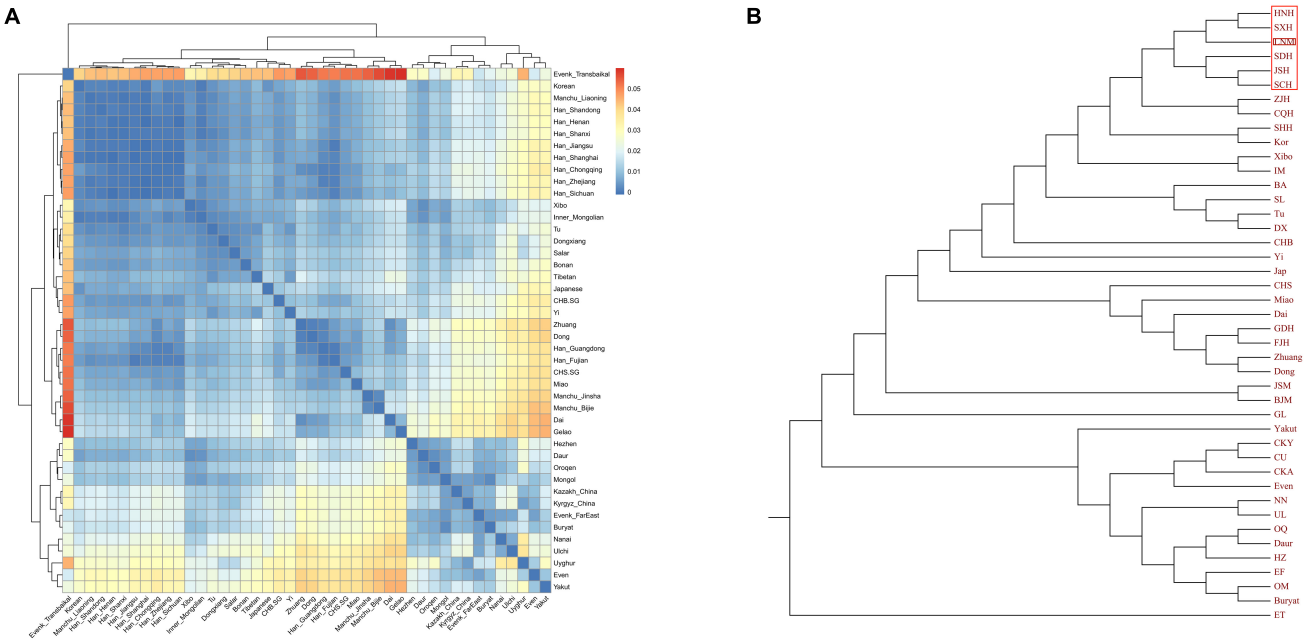


FIGURE 4 | Results of pairwise F_{ST} among East Asian populations. The dark blue color in panel **(A)** reveals a close genetic distance between two different populations, and panel **(B)** displays the phylogenetic tree based on F_{ST} values.

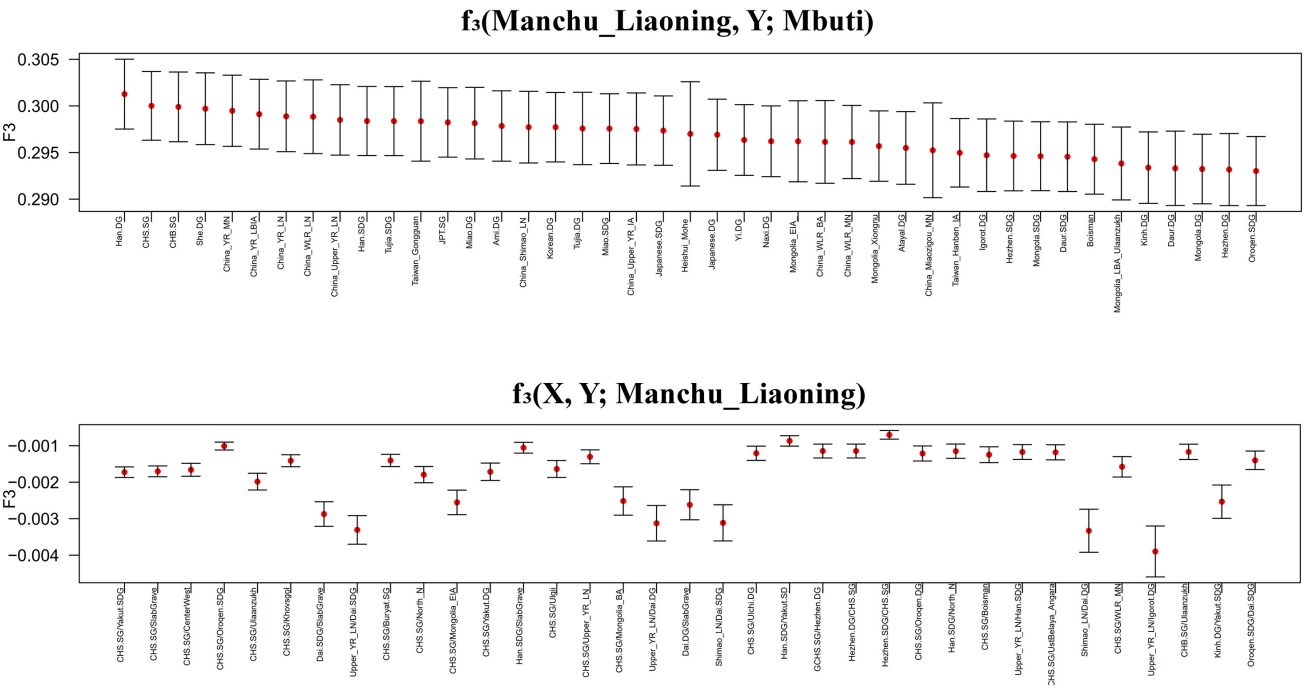


FIGURE 5 | Shared genetic drift estimated via f_3-statistics in different testing forms based on the merged 1240K dataset focused on Liaoning Manchus. Bar denotes three times of the error bar.

and other reference eastern Eurasians. We found that the observed genetic clusters were consistent with the geographic and linguistic categories within East Asia (**Figure 2**). We observed in the PCA that Liaoning Manchus are genetically different with Turkic-, Tungusic-, and Mongolic-speaking groups in northwest China, Mongolia, and Siberia. Liaoning Manchus clustered closely to northern Han Chinese from Shanxi, Shandong, and Henan Provinces and also to Neolithic

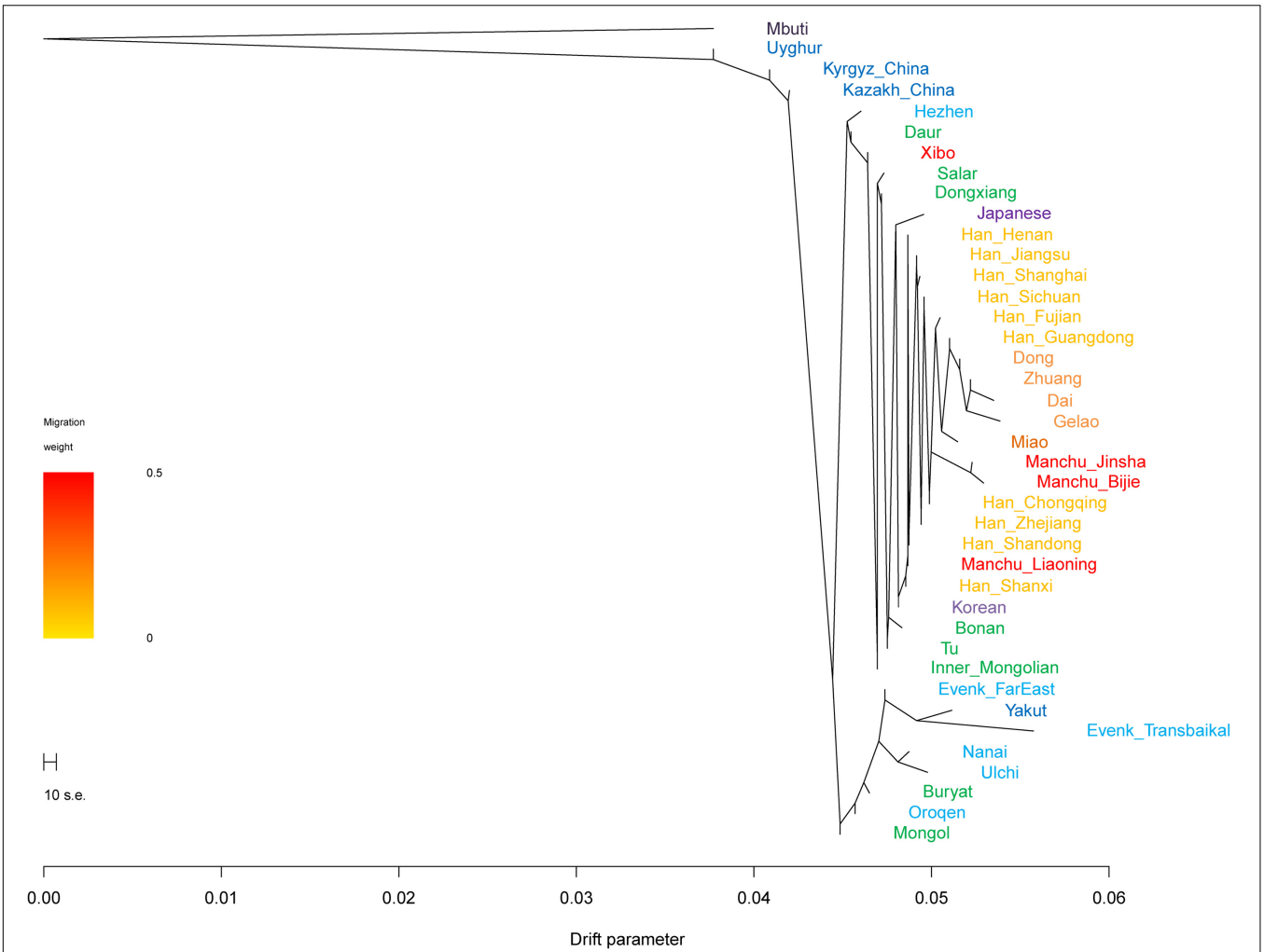


FIGURE 6 | Phylogenetic relationships among Turkic, Tungusic, Mongolic, Tai-Kadai, and Hmong-Mien speakers. Mbuti population from central Africa was set as the root. Different populations were marked according to their linguistic affinity.

northern East Asians from the Yellow River and Western Liao River Basins.

In the model-based ADMIXTURE clustering analysis, we used cross-validation to identify an "optimal" number of clusters. We found the lowest CV error at $K = 5$ (**Supplementary Table 1**). At $K = 5$ (**Figure 3**), we observed there were three components of light green, dark green, and pink color reaching high proportions in Liaoning Manchus. The light green ancestry was enriched in the Tungusic people and ancient populations from the Baikal Lake region. Dark green ancestry with maximum proportion was observed in the southern Chinese populations and Southeast Asians, especially in Taiwan Hanben people. Pink ancestry was maximized in the Yellow River millet farmers. Therefore, Manchus had ancestry related to northeast Asians, Yellow River farmers, and southern East Asians. Similar genetic profiles were observed in the northern Han Chinese, suggesting a close genetic relationship between Manchu and northern Han. In pairwise F_{ST} analysis, we observed a similar pattern that Liaoning Manchus had the smallest genetic distance with Han Chinese in northern China such as in Shandong, Henan, and Shanxi (**Figure 4** and **Supplementary Table 2**).

f_3- and f_4-Statistics

To further investigate the genetic origin and admixture of Liaoning Manchus, we performed outgroup-f_3 and admixture-f_3 statistics to measure allele sharing and detect admixture signals. In outgroup-f_3 (Manchu_Liaoning, Y; Mbuti) (**Figure 5** and **Supplementary Table 3A**), we found that Liaoning Manchus shared the most derived alleles with northern and southern Han Chinese, She, Tujia, Ami, Miao, Japanese, and Korean. When Y represented ancient individuals, Liaoning Manchus share more alleles with ancient East Asians from the Yellow River and Western Liao River Basins, consistent with the observed patterns in PCA and ADMIXTURE. We next used admixture-f_3 statistics in the form of f_3 (X, Y; Manchu_Liaoning) to detect possible admixture signatures, in which X and Y were modern or ancient populations that might be the candidate sources for modeling the admixture of Liaoning Manchus. We observed significant signals

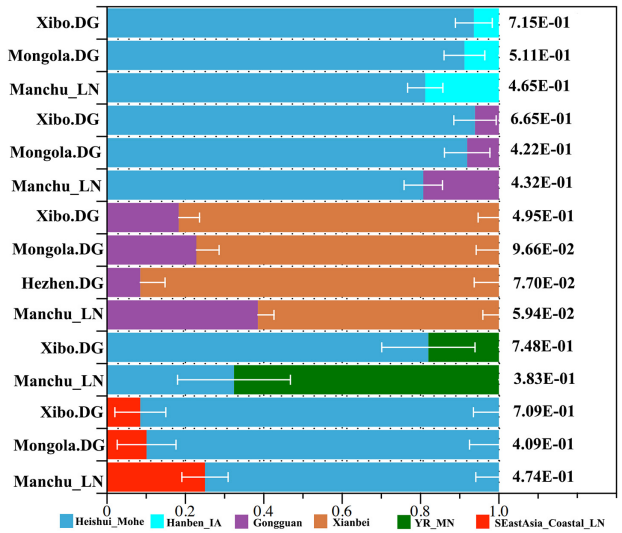


FIGURE 7 | Model of ancient population admixture by *qpWave* and *qpAdm* method. All of these ancestral components were revealed by different colors.

of admixture ($Z < -5$) in the Liaoning Manchus when using CHS or Dai as the southern source and present-day Oroqen, Yakut, Buryat, and Hezhen or ancient Mongolia populations as the northern ancestral source. We listed the $Z < -5$ in the **Supplementary Material** (**Supplementary Table 3B**).

We performed f_4-statistics to explore genetic substructure between studied groups and other modern/ancient populations in the forms of f_4 (X, Y; Manchu_Liaoning, Mbuti) and f_4 (X, Manchu_Liaoning; Y, Mbuti) (**Supplementary Table 4**). When compared with modern Yakut, Tu, Buryat, and Thai populations, Liaoning Manchus shared more derived alleles with Han Chinese and Japanese. Compared with northern Tungusic, Mongolic, and Turkic people in Siberia, Manchu shared more alleles with the Tungusic- and Mongolic-speaking populations in China. When compared with the ancient populations from the Eurasian steppe, such as Yamnaya, EIA_Sagly, and EBA_Chemurchek, Liaoning Manchus shared more derived alleles with ancient populations from Yellow River Basin and West Liao River Basin, Boisman, and some ancient Mongolian such as EIA_SlabGrave and LBA_CenterWest. To further explore the differentiated allele sharing status between Manchu and Han Chinese, we conducted f_4 (Han, Manchu_Liaoning; X, Mbuti); and we found significant negative values when X was the Tungusic and Mongolic people, suggesting that Liaoning Manchus harbored more Tungusic-related ancestry than did Han Chinese (**Supplementary Table 5A**). We further observed significant negative f_4 values in the form f_4 (Mongolic/Tungusic populations, Manchu_Liaoning; southern modern East Asians, Mbuti), suggesting that Manchu people shared more alleles with southern East Asians compared with Tungusic and Mongolic people in the Amur River Region. We also observed significant negative values in f_4 (Xianbei, Manchu_Liaoning; X, Mbuti) when X represented Han, Dai, Tujia, or other southern populations, suggesting that there was gene flow from the southern part of China into Manchu after the Xianbei period. When X was ancient Eurasians, Liaoning Manchus shared more derived alleles with populations from Yellow River Basin and West Liao River Basin (**Supplementary Tables 5B–H**).

TreeMix and qpAdm

In the TreeMix analysis (**Figure 6**), we found that Tungusic-, Turkic-, and Mongolic-speaking groups in northern China tended to cluster together, but Liaoning Manchus clustered with northern Han Chinese and Guizhou Manchus clustered with southern Han Chinese and southern Tai-Kadai-speaking groups. The result was consistent with the patterns observed in the aforementioned PCA, ADMIXTURE, F_{ST}, and f-statistics; Liaoning Manchus and Guizhou Manchus had experienced genetic influence from surrounding Han Chinese after they were separated from northern Tungusic- and Mongolic-speaking ancestors.

We applied *qpWave* and *qpAdm* methods to further infer the possible ancestral populations and estimate the admixture proportions. We used the related available ancient northern populations (Heishui_Mohe, Boisman, Xiongnu, Xianbei, Yamnaya, Afanasievo, SlabGrave, Mongolia_North_N, CenterWest, MongunTaiga, Munkhkhairkhan, and Sagly) as the northern sources, used all available ancient populations (Miaozigou_MN, Shimao_LN, Upper_YR_IA, YR_MN, YR_LBIA, YR_LN, WLR_LN, and Upper_YR_LN) as the source of Yellow River sources, and used Iron Age Hanben (Hanben_IA) and Gongguan samples from Taiwan and Neolithic southern populations (SEastAsia_Coastal_EN, SEastAsia_Coastal_LN, SEastAsia_Island_LN, SEastAsia_Island_EN) as the southern sources. We observed that Manchus can be modeled as deriving 32.4% ancestry from Mohe people and the remaining ancestry from the farming-related ancient populations in the Yellow River Basin (**Figure 7**). When using Late Neolithic to Iron Age ancient southern populations (Hanben_IA, Gongguan and SEastAsia_Coastal_LN) as the southern sources, we found the proportion of northern ancestry (Heishui_Mohe and Xianbei) spanned from 61.5 to 81.2% (**Figure 7**). Compared with other Tungusic or Mongolic populations Hezhen, Mongolia, and Xibo, we observed Liaoning Manchus had derived more ancestry from farming-related populations in the Yellow River Basin and southern China and less ancestry from northern ancient groups, such as Heishui_Mohe and Xianbei.

ALDER, Uniparental Haplogroups, and Chromosome Painting

We next used ALDER software to estimate when the admixture occurred. We tried different modern populations from the north and south of East Asia and Siberia as possible ancestral groups. We observed in most cases that the average time that the admixture occurred was around 500 AD (for example, 46.36 ± 20.93 generations) in the Southern and Northern Dynasties period.[7] That period witnessed large-scale population migrations and admixtures due to the turbulence and war between Xianbei and Han Chinese.

[7]https://en.wikipedia.org/wiki/Northern_and_Southern_dynasties

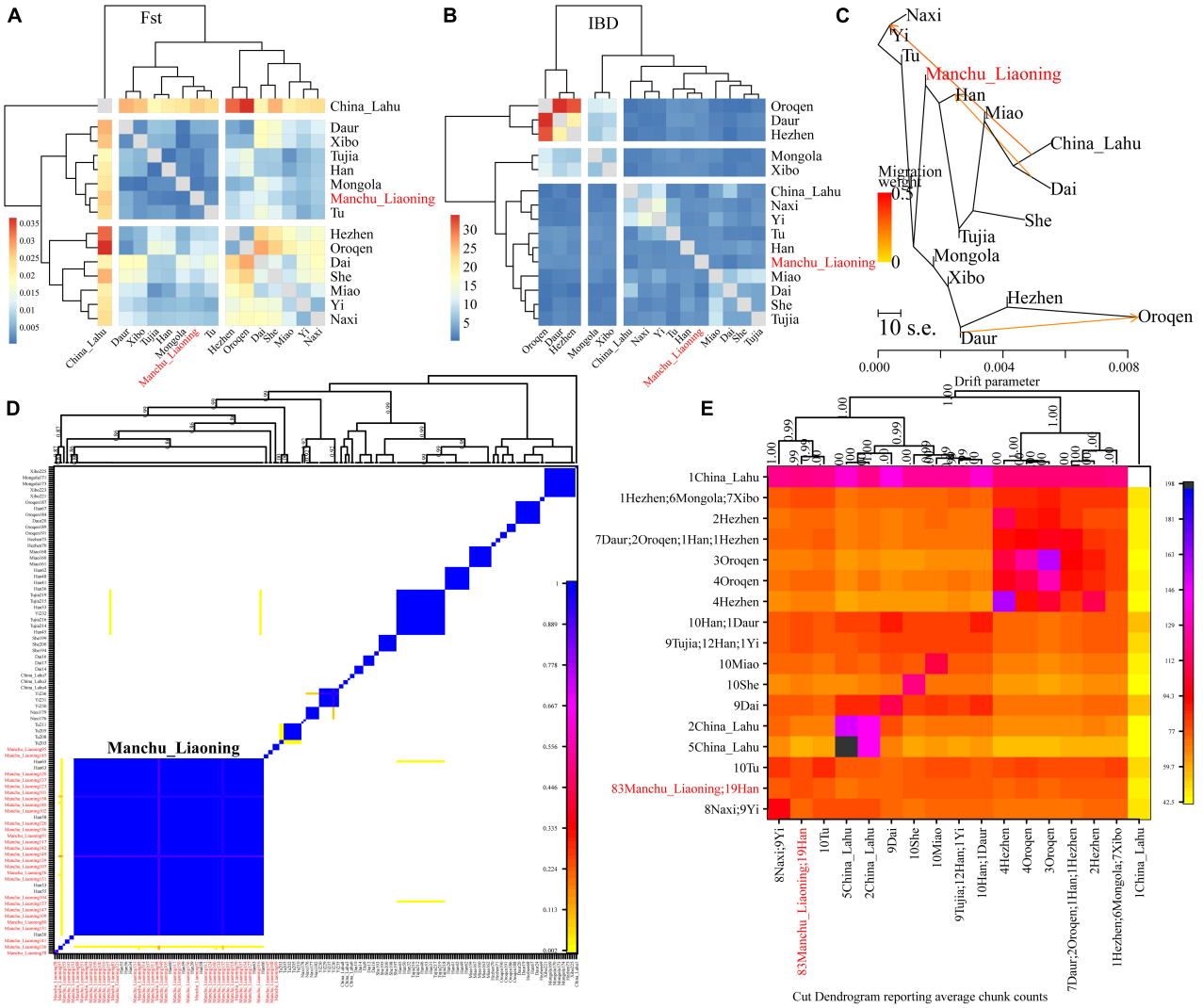


FIGURE 8 | The deep population history of Manchus inferred from both sharing alleles and haplotypes. **(A,B)** Heatmap shows the pairwise F_{ST} genetic distance and identity by decedent (IBD). **(C)** The TreeMix-based phylogenetic tree shows the genetic relationship between Manchu people and other modern East Asians. **(D,E)** Clustering patterns based on the co-ancestry matrix.

We successfully identified 34 uniparental Y-chromosome lineages and 93 mtDNA lineages in Liaoning Manchus as shown in the **Supplementary Material** (**Supplementary Table 6**). We found that D4, A, and M8 were the dominant maternal lineages, and O2a2b1a2 was the dominant paternal lineage. Those paternal and maternal haplogroups are also dominant in Han Chinese, suggesting the possible gene flow from Han Chinese into the gene pool of Manchus.

Finally, to explore the fine-scale population structure of Manchu and other East Asians based on a higher-density dataset, we merged our data with 14 East Asian populations that were whole-genome sequenced in the Human Genome Diversity Project (HGDP) project. We performed the IBD-based clustering and calculated the pairwise F_{ST} genetic distances. We found that Manchus shared the most IBD with Han Chinese but had the smallest F_{ST} genetic distance with northern East Asians (Mongolia and Tu, **Figures 8A,B**). We have not detected the gene flow into Manchu people from our used plausible sources in TreeMix-based phylogeny with three gene flow events, but we found that Manchu clustered in between northern Mongolic and Tungusic people and southern Hans and other Hmong-Mien populations. Manchu clustered the closest with Han Chinese and Tujia (**Figure 8C**), which was further confirmed via the clustering patterns based on the fineSTRUCTURE-based chromosome painting patterns (**Figures 8D,E**).

DISCUSSION

Previous genetic studies have found that northern Han Chinese from Liaoning and Jilin Provinces had genetic admixture with Liaoning Manchus from the analysis of

Y-chromosome and autosomal STRs (Yao and Wang, 2016; Xu et al., 2019). In this study, we also found that Liaoning Manchus had a significant admixture signal with Han Chinese, especially with the northern Han. For example, Manchus had more Sinitic-related ancestry component than other Altaic-speaking populations in ADMIXTURE. We proposed that Liaoning Manchus are an admixture of Han Chinese-related farming populations and local Tungusic-speaking populations in northeast Asia.

Manchus and their ancestors had lived in Manchuria for thousands of years before they moved south, invaded the Central Plains, and even established the Qing Dynasty, which also promoted gene flow between Manchus and Hans. The most well-known recent large-scale population migration event was the "Chuang Guandong"[8] (literally "Crashing into Guandong," with Guandong being an older name for Manchuria). Northeastern China was the hometown of Manchus, and the first emperor of the Qing Dynasty was born there. With the establishment of the Qing Dynasty (see text footnote 2), Manchuria did not allow people who were non-Eight Banner to enter. Manchuria was vast and rich in material but sparsely populated. Therefore, many Han Chinese who lived in Hebei and Shandong Provinces left their hometown and went to Manchuria for survival because of various reasons such as multi-year natural disasters and shortage of food (Reardon-Anderson, 2000; Reardon-Anderson, 2005). Within the last 300 years, at least 30 million immigrants traveled far away over the mountains across the seas to northeast China. Han Chinese from Shandong, Hebei Province, and other regions lived together with the native Manchu population in northeast China. Until 1840, Han Chinese had filled up most of Manchuria's cities and towns and left profound impacts on Manchu people in various aspects (Reardon-Anderson, 2000). For example, almost all Manchu people can speak Chinese; and in recent years, they had even abandoned the Manchu language. At the end of the Qing Dynasty, in order to consolidate its dominant position, the Qing government announced that the Manchus and Hans were one family and abandoned various restrictions on Han people to allow the Manchu nobility to marry Han people (Jones and Kuhn, 1978). However, intermarriages between Manchu people and Han people were in fact very common among ordinary people.

ALDER-based admixture time estimations revealed Mongolian and Even and Manchus have genetic admixture in ∼500 AD, which was the period of Southern and Northern Dynasties (see text footnote 6) (420–589 AD). It was a turbulent period of war and also a period of large-scale population admixture. At that time period, the Xianbei population controlled the Amur River Basin, Mongolian Plateau, and southern Siberia, and they became the largest ruling power in the North. In this study, we can also model Manchus deriving ancestry from the Xianbei people.

Previous studies have found strong associations between population genetic structure and linguistic similarity in Asia, and the populations from the same language group have a closer genetic affinity (Chen et al., 2019; He et al., 2020). Mongolic- and Tungusic-speaking populations were reported to be genetically similar (Ning et al., 2020; Wang et al., 2021a). In this study, we found that both Mongolic- and Tungusic-speaking populations have ancestry components related to ancient Mohe and Xianbei people. We proposed there were many cultural interactions and gene flows between Mongolic- and Tungusic-speaking populations.

[8]https://en.wikipedia.org/wiki/Chuang_Guandong

CONCLUSION

We reported the first research-based genome-wide SNP data of the Manchus from Liaoning Province. We used comprehensive population genetic analyses of PCA, ADMIXTURE, *qpAdm*, *qpWave*, *f*-statistic, F_{ST}, ALDER, IBD, fineSTRUCTURE, and TreeMix to investigate the complex genetic history and dynamic admixture process of northern Chinese populations. Our previous study documented a long-term genetic continuity from Neolithic hunter-gatherers to present-day Tungusic-speaking people in northeast Asia. It is therefore believed that the Manchu people, being members of an old branch of the Tungusic, should have a consistent genetic profile with other Tungusic groups. However, we found that Liaoning Manchus have a close genetic relationship and significant admixture signal with Han Chinese. The Manchu population was an exception to the coherent genetic structure of Tungusic-speaking people, probably due to the large-scale population migrations and genetic admixtures in the past hundred years.

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 Our study and sample collection were reviewed and approved by the Ethics Committee of Jinzhou Medical University (JZMULL2021010) and followed the recommendations provided by the revised Helsinki Declaration of 2000. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YoW and C-CW designed the study. XZ and GH analyzed the data and wrote the manuscript. WL, YuW, YiW, and HX carried out the sample collection. XL, YC, and QQ conducted the experiment. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by grants from the National Natural Science Foundation of China (No. 31571233), the scientific research project from the Education Department of Liaoning Province (No. JYTJCZR2020074), the Nanqiang Outstanding Young Talents Program of Xiamen University (X2123302), the Fundamental Research Funds for the Central Universities (ZK1144), the Major Project of the National Social Science Foundation of China (20&ZD248), and European Research Council (ERC) grant (ERC-2019-ADG-883700-TRAM).

ACKNOWLEDGMENTS

We thank Fudan University Taizhou Institute of Health Sciences for helping with DNA preservation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.754492/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

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

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., 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

Feng, Q., Lu, D., and Xu, S. (2018). AncestryPainter: a graphic program for displaying ancestry composition of populations and individuals. *Genomics Proteomics Bioinformatics* 16, 382–385. doi: 10.1016/j.gpb.2018.05.002

He, G., Wang, M., Zou, X., Chen, P., Wang, Z., Liu, Y., et al. (2021a). 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. L., Wang, M. G., Li, Y. X., Zou, X., Yeh, H. Y., Tang, R. K., et al. (2021c). Fine-scale north-to-south genetic admixture profile in Shaanxi Han Chinese revealed by genome-wide demographic history reconstruction. *J. Syst. Evol.* 00.1–00.18. doi: 10.1111/jse.12715

He, G., Wang, M., Zou, X., Tang, R., Yeh, H.-Y., Wang, Z., et al. (2021b). Genomic insights into the differentiated population admixture structure and demographic history of North East Asians. *bioRxiv* [Preprint] bioRxiv, 2021.2007.2019.452943, 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

He, J., and Guo, F. (2013). Population genetics of 17 Y-STR loci in Chinese Manchu population from Liaoning Province, Northeast China. *Forensic Sci. Int. Genet.* 7, e84–e85. doi: 10.1016/j.fsigen.2012.12.006

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. doi: 10.1038/s41559-019-0878-2

Jones, S. M., and Kuhn, P. A. (1978). "Dynastic decline and the roots of rebellion," in *The Cambridge History of China: Vol 10: Late Ch'ing 1800–1911*, ed. J. K. Fairbank (Cambridge: Cambridge University Press), 107–162. doi: 10.1017/CHOL9780521214476.004

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

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

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, J., Guo, L., Qi, R., Li, S. Y., Yin, J. Y., Zhang, W., et al. (2013). Allele frequencies of 19 autosomal STR loci in Manchu population of China with phylogenetic structure among worldwide populations. *Gene* 529, 282–287. doi: 10.1016/j.gene.2013.07.033

Liu, Y., Wang, M., Chen, P., Wang, Z., Liu, J., Yao, L., et al. (2021a). Combined low-/high-density modern and ancient genome-wide data document genomic admixture history of high-altitude East Asians. *Front. Genet.* 12:582357. doi: 10.3389/fgene.2021.582357

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

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.e–3266.e. doi: 10.1016/j.cell.2021.04.040

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

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

Reardon-Anderson, J. (2000). Land use and society in Manchuria and Inner Mongolia during the Qing Dynasty. *Environ. Hist.* 5, 503–530. doi: 10.2307/3985584

Reardon-Anderson, J. (2005). *Reluctant Pioneers: China's Expansion Northward, 1644–1937*. Stanford, CA: Stanford University Press, 40–45.

Wall, J. D., and Yoshihara Caldeira Brandt, D. (2016). Archaic admixture in human history. *Curr. Opin. Genet. Dev.* 41, 93–97. doi: 10.1016/j.gde.2016.07.002

Wang, M., Yuan, D., Zou, X., Wang, Z., Yeh, H.-Y., Liu, J., et al. (2021d). 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, M., He, G., Gao, S., Jia, F., Zou, X., Liu, J., et al. (2021b). Molecular genetic survey and forensic characterization of Chinese Mongolians via the 47 autosomal insertion/deletion marker. *Genomics* 113, 2199–2210.

Wang, M., He, G., Zou, X., Liu, J., Ye, Z., Ming, T., et al. (2021c). Genetic insights into the paternal admixture history of Chinese Mongolians via high-resolution customized Y-SNP SNaPshot panels. *Forensic Sci. Int. Genet.* 54, 102565. doi: 10.1016/j.fsigen.2021.102565

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.

Wang, C. Z., Wei, L. H., Wang, L. X., Wen, S. Q., Yu, X. E., Shi, M. S., et al. (2019). Relating Clans Ao and Aisin Gioro from northeast China by whole Y-chromosome sequencing. *J. Hum. Genet.* 64, 775–780.

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

Weissensteiner, H., Pacher, D., Kloss-Brandstatter, A., Forer, L., Specht, G., Bandelt, H. J., et al. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* 44, W58–W63.

Wu, Y. (2020). Inference of population admixture network from local gene genealogies: a coalescent-based maximum likelihood approach. *Bioinformatics* 36, i326–i334.

Xing, J., Adnan, A., Rakha, A., Kasim, K., Noor, A., Xuan, J., et al. (2019). Genetic analysis of 12 X-STRs for forensic purposes in Liaoning Manchu population from China. *Gene* 683, 153–158.

Xu, X. M., Zheng, J. L., Lou, Y., Wei, X. H., Wang, B. J., and Yao, J. (2019). Population genetics of 24 Y-STR loci in Chinese Han population from Jilin Province, Northeast China. *Mol. Genet. Genomic Med.* 7:e984.

Xue, Y., Zerjal, T., Bao, W., Zhu, S., Lim, S. K., Shu, Q., et al. (2005). Recent spread of a Y-chromosomal lineage in northern China and Mongolia. *Am. J. Hum. Genet.* 77, 1112–1116.

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, 295–298.

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.

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

Yao, J., and Wang, B. J. (2016). Genetic Variation of 25 Y-Chromosomal and 15 Autosomal STR Loci in the Han Chinese Population of Liaoning Province, Northeast China. *PLoS One* 11:e0160415. doi: 10.1371/journal.pone.0160415

Zhao, Y. B., Sun, W. Y., Zhan, Y., Di, W., and Yu, C. C. (2011). Mitochondrial DNA evidence of southward migration of Manchus in China. *Mol. Biol. (Mosk.)* 45, 825–830. doi: 10.1134/S0026893311050153

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: 15 October 2021
doi: 10.3389/fgene.2021.705753

Comprehensive Insights Into Forensic Features and Genetic Background of Chinese Northwest Hui Group Using Six Distinct Categories of 231 Molecular Markers

*Chong Chen[1,2†], Xiaoye Jin[1†], Xingru Zhang[1,2], Wenqing Zhang[1], Yuxin Guo[1], Ruiyang Tao[3], Anqi Chen[3,4], Qiannan Xu[3,5], Min Li[3,5], Yue Yang[3,6] and Bofeng Zhu[1,2,7]**

[1]Key Laboratory of Shaanxi Province for Craniofacial Precision Medicine Research, College of Stomatology, Xi'an Jiaotong University, Xi'an, China, [2]Guangzhou Key Laboratory of Forensic Multi-Omics for Precision Identification, School of Forensic Medicine, Southern Medical University, Guangzhou, China, [3]Shanghai Key Laboratory of Forensic Medicine, Shanghai Forensic Service Platform, Academy of Forensic Sciences, Ministry of Justice, Shanghai, China, [4]Department of Forensic Medicine, Shanghai Medical College of Fudan University, Shanghai, China, [5]Institute of Forensic Medicine, West China School of Basic Medical Sciences and Forensic Medicine, Sichuan University, Chengdu, China, [6]School of Basic Medicine, Inner Mongolia Medical University, Hohhot, China, [7]Department of Forensic Genetics, Multi-Omics Innovative Research Center of Forensic Identification, School of Forensic Medicine, Southern Medical University, Guangzhou, China*

OPEN ACCESS

Edited by:
Maxat Zhabagin, National Center for Biotechnology, Kazakhstan

Reviewed by:
Eaaswarkhanth Muthukrishnan, Dasman Diabetes Institute, Kuwait
Guanglin He, Nanyang Technological University, Singapore

***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: *06 May 2021*
Accepted: *07 September 2021*
Published: *15 October 2021*

Citation:
Chen C, Jin X, Zhang X, Zhang W, Guo Y, Tao R, Chen A, Xu Q, Li M, Yang Y and Zhu B (2021) Comprehensive Insights Into Forensic Features and Genetic Background of Chinese Northwest Hui Group Using Six Distinct Categories of 231 Molecular Markers. Front. Genet. 12:705753. doi: 10.3389/fgene.2021.705753

The Hui minority is predominantly composed of Chinese-speaking Islamic adherents distributed throughout China, of which the individuals are mainly concentrated in Northwest China. In the present study, we employed the length and sequence polymorphisms-based typing system of 231 molecular markers, i.e., amelogenin, 22 phenotypic-informative single nucleotide polymorphisms (PISNPs), 94 identity-informative single nucleotide polymorphisms (IISNPs), 24 Y-chromosomal short tandem repeats (Y-STRs), 56 ancestry-informative single nucleotide polymorphisms (AISNPs), 7 X-chromosomal short tandem repeats (X-STRs), and 27 autosomal short tandem repeats (A-STRs), into 90 unrelated male individuals from the Chinese Northwest Hui group to comprehensively explore its forensic characteristics and genetic background. Total of 451 length-based and 652 sequence-based distinct alleles were identified from 58 short tandem repeats (STRs) in 90 unrelated Northwest Hui individuals, denoting that the sequence-based genetic markers could pronouncedly provide more genetic information than length-based markers. The forensic characteristics and efficiencies of STRs and IISNPs were estimated, both of which externalized high polymorphisms in the Northwest Hui group and could be further utilized in forensic investigations. No significant departure from the Hardy–Weinberg equilibrium (HWE) expectation was observed after the Bonferroni correction. Additionally, four group sets of reference population data were exploited to dissect the genetic background of the Northwest Hui group separately from different perspectives, which contained 26 populations for 93 IISNPs, 58 populations for 17 Y-STRs, 26 populations for 55 AISNPs (raw data), and 109 populations for 55 AISNPs (allele frequencies). As a result, the analyses based on the Y-STRs indicated that the Northwest Hui group primarily exhibited intimate genetic relationships with reference Hui

groups from Chinese different regions except for the Sichuan Hui group and secondarily displayed close genetic relationships with populations from Central and West Asia, as well as several Chinese groups. However, the AISNP analyses demonstrated that the Northwest Hui group shared more intimate relationships with current East Asian populations apart from reference Hui group, harboring the large proportion of ancestral component contributed by East Asia.

Keywords: Chinese Hui minority, population genentics, forensics, molecular markers, MPS

INTRODUCTION

For recent years, the extensive applications of DNA analysis technologies have made it a crucial tool in forensic investigations. To date, the genomic DNA from biological samples is predominantly dissected by PCR and capillary electrophoresis (CE)-based method to reveal the length variations in genetic markers, such as short tandem repeats (STRs). Further, the DNA sequencing technology also serves as an important role in providing comprehensive information of the target DNA in forensic applications. Conventional Sanger sequencing, initially introduced in the 1970s, has enabled enormous progress in the fields of molecular biology, genomics, and genetics and been regarded as the standard sequencing technology in the forensic investigations (Sanger et al., 1977). Yet, the shortcomings of Sanger sequencing technology, such as low throughput and sensitivity, have hampered its utilization in the more in-depth and intricate genome analyses, which facilitate the exploration of other high-throughput DNA technologies in forensic studies (Fullwood et al., 2009). Massively parallel sequencing (MPS), or next-generation sequencing (NGS), has become an emerging tool commonly utilized in forensic genetic fields (Børsting and Morling, 2015; Li et al., 2017). MPS technology has advantages such as simultaneous sequencing of multiple types of genetic molecular markers and detecting samples at an extraordinarily high throughput capacity, which make it possible to yield forensic data containing more information in a single reaction (England et al., 2020). In addition, the polymorphisms of sequence variations contained in different genetic molecular markers which are undetectable by traditional CE technology, like STRs, are easily identified using MPS technology platforms, thus increasing the possibility of discovering new STR alleles (Gettings et al., 2015; Churchill et al., 2016; Novroski et al., 2016). MPS platforms also maintain the conventional abilities to identify STRs length polymorphisms, facilitating the compatibility of currently forensic DNA data generated by PCR and CE-based platform (Parson et al., 2016; Bruijns et al., 2018).

The ForenSeq™ DNA Signature Prep Kit (Verogen Inc., San Diego, CA, United States), a newly developed MPS-based commercial kit, is a length and sequence polymorphisms-based typing system for three kinds of STRs and three types of single nucleotide polymorphisms (SNPs), which can simultaneously detect 231 genetic molecular markers in a single reaction on MiSeq FGx™ Forensic Genomics System (Verogen Inc., San Diego, CA, United States). This kit provides two different primer mixes, including primer mixes A and B. Primer mix A is designed to detect amelogenin, 27 autosomal STRs (A-STRs), 24 Y-chromosomal STRs (Y-STRs), 7 X-chromosomal STRs (X-STRs), and 94 identity informative SNPs (IISNPs); and primer mix B contains primer mix A plus the primers for 22 phenotypic informative SNPs (PISNPs) and 56 biogeographical ancestry informative SNPs (AISNPs) (Jäger et al., 2017). To date, the system performance of the ForenSeq™ DNA Signature Prep Kit has been comprehensively evaluated by investigating the reproducibility, sensitivity, concordance, casework-type sample, and inter-laboratory comparison and validation, which has proven it to be a promising tool in forensic applications over recent years (Just et al., 2017; Xavier and Parson, 2017; Köcher et al., 2018). Further, several studies have demonstrated that this kit is suitable for challenging samples such as forensically degraded and mixed samples from crime scenes (Fattorini et al., 2017; Sharma et al., 2020), and also performs well in the kinship analyses (Li et al., 2018), phenotypic and biogeographical ancestry predictions (Sharma et al., 2019), and in the exploratory studies of population genetics (Delest et al., 2020).

The Chinese Hui ethnic group is one of the national minorities officially recognized by the People's Republic of China and widespread throughout China, including 34 provincial-level administrative regions. According to the population distribution, the Chinese Hui group is more concentrated in Northwest China. Intriguingly, the Hui group occupies a special existence among the Chinese ethnic minorities, of which the individuals are portrayed as Muslims who appear culturally and linguistically similar to the Han population. Currently, population genetic studies concentrated on limited kinds of genetic markers have been conducted on Hui groups from Chinese different regions. For example, Zou et al. detected the genetic polymorphisms of 30 insertion/deletion (InDels) in the Guangxi Hui group (Zou et al., 2020). HLA class I polymorphisms were investigated in the Hui group from Chinese Qinghai province by Hong et al. (2007). The genomic makeup and ancestry background of the Hui group from Sichuan province were explored using over 700K SNPs by Liu et al. (2021). Guo et al. investigated the population structure of the Hui group from Chinese Liaoning province based on 17 Y-STR loci (Guo, 2017). Yao et al. and Liu et al. explored the genetic background of the Hui group from Chinese Gansu province using 15 autosomal STRs (Yao et al., 2016) and 27 Y-STRs (Liu et al., 2018) loci, respectively. The genetic polymorphisms of the Hui group from Chinese Ningxia Hui (NXH) autonomous region were estimated

by various research teams using 15 STRs (Ma et al., 2017), 24 Y-STRs (Zhu et al., 2014), 30 InDels (Zhou et al., 2020), 12 X-STRs (Meng et al., 2014), and 17 Y-STR loci (Guo et al., 2008), respectively. Thirty InDels (Xie et al., 2018), 30 AISNPs (Jin et al., 2020), 22 STRs (Fang et al., 2018), and 39 ancestry-informative marker (AIM) InDel loci (Xie et al., 2020) were also separately applied in other Hui groups from Northwest China.

In fact, different opinions on the genetic origin and ancestral history of the Chinese Hui group have existed for years. Two kinds of hypotheses have been proposed to infer the historical formation of the Chinese Hui group, which are demic diffusion and cultural diffusion hypotheses (Liu et al., 2021). In addition, due to the limited types of genetic markers utilized separately in previous studies mentioned above, it might be more necessary to simultaneously enroll various types of genetic markers to provide more extensive information for exploring the genetic structure and ancestral origin of the Chinese Hui group. Thus, in the present study, we implemented 231 genetic markers (amelogenin, 27 A-STRs, 24 Y-STRs, 7 X-STRs, 94 IISNPs, 22 PISNPs, and 56 AISNPs) from the ForenSeq™ DNA Signature Prep Kit into the Hui group from Northwest China. Both length- and sequence-based polymorphisms of genetic markers were utilized into 90 unrelated male individuals recruited from the Northwest Hui group, with the intention of comprehensively detecting the forensic characteristics and genetic background of the Northwest Hui group and subsequently enriching the genetic information of the Chinese populations.

MATERIALS AND METHODS

Sample Information

In total, 90 blood samples from unrelated healthy Hui male individuals from Northwest China were collected. All the participants provided their written informed consents prior to sample collection. The present research was conducted according to the ethical guidelines of the Xi'an Jiaotong University Health Science Center and further authorized by the Ethical committees of the Xi'an Jiaotong University Health Science Center (approval number: 2019-1039; 2020-1382).

Library Preparation

Library preparation was performed in accordance with the recommendation of the ForenSeq™ DNA Signature Prep Kit (Illumina Inc., CA, United States) (Illumina, 2015). The first stage was to amplify and tag the DNA targets. One disc containing a 1.2 mm-diameter blood sample was directly amplified without DNA extraction in the ForenSeq™ Sample Plate. PCR was conducted on the GeneAmp PCR System 9700 Thermal Cycler (Applied Biosystems, CA, United States) according to the following parameters: 98°C for 3 min; eight cycles of 96°C for 45 s, 80°C for 30 s, 54°C for 2 min, and 68°C for 2 min; 10 cycles of 96°C for 30 s, 68°C for 3 min; 68°C for 10 min; and hold at 10°C. The first-round PCR products were subsequently utilized in the second-stage PCR to enrich the targets. The index adapters and sequences required for cluster amplification were also added in the second-stage PCR. The detailed PCR parameters were as follows: 98°C for 30 s; 15 cycles of 98°C for 20 s, 66°C for 30 s, and 68°C for 90 s; 68°C for 10 min; and hold at 10°C. All samples were amplified with DNA primer mix B. The amplified DNA libraries were then purified from the remaining reaction components using purification beads, and the concentrations of the libraries were normalized to ensure a consistent cluster density. The normalized libraries required to sequence on the same flow cell were then mixed in equal volumes during the library pooling stage. The mixed libraries were then diluted in a hybridization buffer, added the human sequencing control and denatured at 96°C in preparation for sequencing. Finally, after denaturation and dilution, the mixed libraries were sequenced on the MiSeq Desktop Sequencer using the MiSeq FGx Forensic Genomics System (Illumina Inc., CA, United States) (Illumina, 2015).

Data Analyses and Interpretation

The generated data were processed using the ForenSeq™ Universal Analysis Software based on the default thresholds. In detail, the analytical threshold (AT), interpretation threshold (IT), stutter filter (SF), and intra-locus balance (IB) were utilized to estimate the data quality (Illumina, 2015; Köcher et al., 2018). The optimized AT values were 1.5% for all loci except for DYS635 (3.3%), DYS389II (5%), and DYS448 (3.3%) loci, which represented the lower boundary for the valid results accounting for the entire read count per locus. The default IT values were 4.5% for all loci apart from DYS635 (10%), DYS389II (15%), and DYS448 (10%) loci, indicating the upper boundary of the uncertainty range. When the resulting values are between the AT and IT, the user should identify whether an actual variant has occurred. The SF values for all STR loci were extended from 7.5% to 50% with the average value as 19%. The general settings of IB values were 60% for STRs and 50% for SNPs. As for the valid sequencing depth, the minimum depths were set as 10× for STRs and 5× for SNPs. For A-STRs, the parameters for both sequence- and length-based polymorphisms were calculated using STRAF software v1.0.5 (Gouy and Zieger, 2017), including observed heterozygosity (H_{obs}), expected heterozygosity (H_{exp}), polymorphism information content (PIC), power of discrimination (PD), exclusion probability (PE), match probability (PM), typical paternity index (TPI), and the Hardy–Weinberg equilibrium (HWE). Specifically, the H_{obs} was utilized to measure the genetic polymorphic degree of a specific loci, which is described as the observed proportion of heterozygous genotypes detected in a population (Sheriff and Alemayehu, 2018). The H_{exp}, a fundamental statistical parameter used to estimate genetic diversity within populations, is referred to as the expected proportion of heterozygotes under HWE (Nei, 1973). The PIC has been reported as a statistical indicator for measuring the polymorphisms of genetic markers in a population (Botstein et al., 1980; Shete et al., 2000), which is actually determined using heterozygosity and number of alleles (Al-jumaah et al., 2012). The PM is defined as the probability of a match between two unrelated individuals selected at random (Fisher, 1951; Malaspinas et al., 2011). Closely related to PM, the PD is the probability of distinguishing two unrelated individuals (Tillmar, 2010). The PE can be treated as the probability of

excluding a man falsely indicated as the biological father, which is a measure of efficiency in paternity testing (Cifuentes et al., 2006; Tillmar, 2010; Vandeputte, 2012). For X-STRs and Y-STRs, the gene diversity (GD), PM, haplotype diversity (HD), and haplotype discrimination capacity (DC) were calculated based on the methods of previous reports (Nei, 1987; Liu et al., 2020). In terms of the IISNPs, the STRAF software v1.0.5 was again utilized to estimate the forensic parameters, including H_{obs}, H_{exp}, PIC, PD, PE, PM, TPI, and HWE.

Population Genetic Analyses

A total of four group sets of previously published population data were exploited as references to perform population genetic analyses and further dissect the ancestral components of the Northwest Hui group. The employed data are as follows: raw data of 26 populations for 93 IISNPs enlisted from the 1000 Genomes Project (https://www.internationalgenome.org/data), raw data of 17 Y-STRs for 58 populations collected from previous studies or the Y Chromosome Haplotype Reference Database (YHRD) database (https://yhrd.org/), raw data of 26 populations for 55 AISNPs gathered from the 1000 Genomes Project, and allele frequencies of 109 populations for 55 AISNPs assembled from previously published studies (Kidd et al., 2014; Pakstis et al., 2015; Pakstis et al., 2017; Pakstis et al., 2019a). The detailed reference population information and citations are listed in **Supplementary Table S1**.

Initially, population genetic analyses of the Northwest Hui group and the 26 reference populations (Auton et al., 2015) were performed based on the 93 overlapping IISNPs. A heatmap of allele frequencies for these 93 IISNPs in the Northwest Hui and 26 reference populations was plotted using the *R* software v3.3 (Team, 2016), which was hierarchically clustered based on the Euclidean distances. Principal component analysis (PCA) of these populations was conducted using the XLSTAT program (https://www.xlstat.com/en/) based on the allele frequencies of these 93 IISNPs. We also conducted a PCA plot of these populations at the individual level using the PLINK software v1.9 based on the raw data of these 93 IISNPs (Chang et al., 2015). The pairwise D_A genetic distances of these populations were calculated using the DISPAN program (http://www.personal.psu.edu/nxm2/dispan2.htm), and a rooted neighbor-joining (NJ) tree was further constructed using the MEGA software v6.0 (Tamura et al., 2013) based on the D_A genetic distances. Moreover, the population-specific divergence values of each loci in different intercontinental population sets were estimated using the informativeness for assignment (I_n) statistic based on the allele frequencies of 93 IISNPs (Phillips, 2015). The I_n statistical analysis was originally introduced by Rosenberg et al. (2003), with the purpose of determining the amount of ancestry information provided by biallelic or multiallelic markers, which could be further utilized to evaluate the effectiveness of genetic markers in differentiating various populations. Next, the population genetic relationships between the Hui group and 58 worldwide populations were revealed from a paternal perspective based on 17 overlapping Y-STRs. The pairwise R_{st} values based on Y-STR haplotypes were generated using an online AMOVA tool (https://yhrd.org/amova). R_{st} is developed based on a stepwise mutation model to assess the genetic differentiations among populations (Balloux and Lugon-Moulin, 2002). Both heatmaps and histograms of the R_{st} values were performed using *R* software v3.3 (Team, 2016). Multidimensional scaling (MDS) was yielded based on R_{st} values using the Statistical Package for the Social Sciences (SPSS) 16.0 software. Eventually, the ancestral structure of the Northwest Hui group based on 55 AISNPs was explored using two sets of population data as the references. The PCA plot on individual level and STRUCTURE analysis were performed using the XLSTAT program (https://www.xlstat.com/en/) and ADMIXTURE software v1.3 (Alexander et al., 2009), respectively, based on the 55 AISNPs genotypes of the Northwest Hui group and the 26 reference populations. The ancestral component analysis of the Northwest Hui group was performed using ADMIXTURE software v1.3 (Alexander et al., 2009) on the basis of 55 AISNPs genotypes. The PCA, pairwise D_A genetic distances, and the rooted NJ tree of the Northwest Hui and other 109 reference populations were conducted using SPSS 16.0 software, the DISPAN program, and MEGA software v6.0 (Tamura et al., 2013), respectively, using the allele frequencies of 55 overlapping AISNPs.

RESULTS

MPS Results for Three Genres of STRs and Three Kinds of SNPs

The MPS genetic data of the 90 male Hui individuals were genotyped using the primer mix B reagent of the ForenSeq™ DNA Signature Prep Kit. The genotyping results for three genres of STRs (27 A-STRs, 24 Y-STRs, and 7 X-STRs) based on both length- and sequence-based polymorphisms are presented in **Supplementary Table S2**. The total success rates of 27 A-STRs, 24 X-STRs, and 7 Y-STRs were 99.96%, 99.84%, and 99.81%, respectively. A total of 54 out of 58 STR loci were efficiently genotyped with a success rate of 100%, while the success rates of the remaining PentaE, DYS392, DYS448, and DXS7132 loci were 98.89, 98.89, 96.67, and 98.89%, respectively. The genotyping results for 94 IISNPs, 56 AISNPs, and 22 PISNPs are shown in **Supplementary Table S3** with the success rates of 100%. As presented in **Supplementary Table S4**, for 230 genetic markers without an amelogenin locus, the total sequencing depth of 90 samples was 2,530,188; the average sequencing depths for each sample and each locus were 28,113.20 and 11,000.82, respectively; further, the average sequencing depth per locus for each sample was 122.23. In detail, as shown in **Supplementary Figure S1**, the sequencing depths of the A-STRs fluctuated from 209.97 (SD ± 166.79) at PentaE locus to 12,535.64 (±3,565.09) at TH01 locus. The minimum sequencing depth of the X-STRs was 91.13 (±30.97) at DXS10103 locus, whereas the maximum sequencing depth, observed at DXS10074 locus, was 2,540.82 (±827.35). For the Y-STRs, the DYS438 locus provided the highest sequencing depth as 10,176.43 (±3,039.40), while the DYS460 locus offered the lowest sequencing depth as 221.01 (±106.94). In terms of the 94 IISNP loci, the sequencing depths ranged from 77.14 (±21.23) at

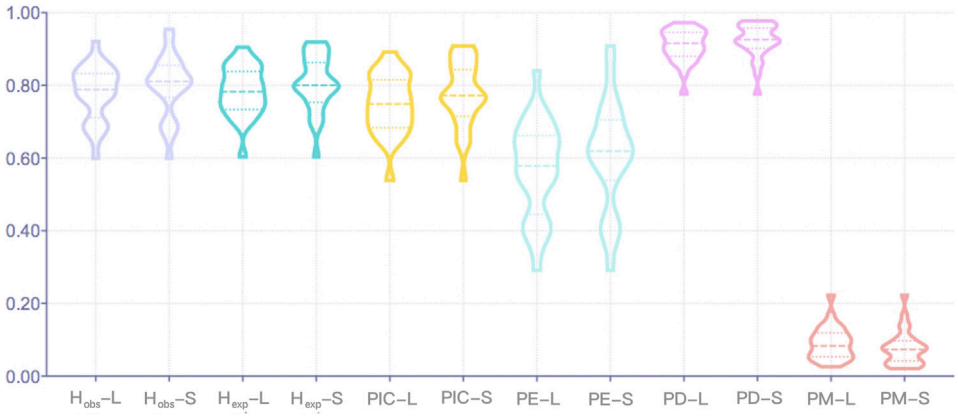


FIGURE 1 | The forensic parameters of 27 A-STR loci at the length (L)- and sequence (S)-based levels. A-STR, autosomal short tandem repeat.

rs1736442 locus to 5,700.47 (±2,072.63) at rs8037429 locus. The maximum and minimum sequencing depths of 56 AISNPs were detected as 6,432.28 (±1,907.24) at rs7997709 locus and 208.23 (±57.93) at rs310644 locus, respectively. The sequencing depths of 22 PISNP loci were in the range from 430.57 (±120.85) at rs12821256 locus to 2,456.13 (±628.14) at rs2402130 locus.

Forensic Parameters for Various Genres of Genetic Markers

For the Northwest Hui group, the allelic polymorphisms and forensic parameters of 27 A-STR loci calculated on the basis of sequence- and length-based polymorphisms are presented in **Supplementary Table S5**. The number of length-based alleles fluctuated from 5 at D3S1358 and D4S2408 loci to 15 at D18S51, FGA, and Penta E loci, whereas the number of sequence-based alleles varied from 6 at D22S1045, TH01, and TPOX loci to 33 at D12S391 locus. HWE tests were conducted separately on sequence- and length-based A-STR loci. Before Bonferroni correction, it was observed that the TH01 locus deviated from the HWE for both sequence- and length-based polymorphisms; the vWA locus departed from the HWE for length-based polymorphisms; and the remaining 25 A-STR loci were in accordance with the HWE. The other forensic parameters are graphically presented in **Figure 1**. The H_{obs} values for length- and sequence-based STRs ranged from 0.6000 to 0.9222 and from 0.6000 to 0.9556, respectively, of which the lowest values were all observed at TPOX locus and the highest at D12S391 locus. The H_{exp} values for the length-based STRs spanned from 0.604 at TPOX locus to 0.906 at Penta E locus, and for sequence-based STRs varied from 0.604 at TPOX locus to 0.9200 at D12S391 locus. The average H_{obs} and H_{exp} values for the length-based polymorphisms were 0.7798 (±0.0744) and 0.7845 (±0.0670), respectively. For the sequence polymorphism level, the average H_{obs} and H_{exp} values were 0.7999 (±0.0829) and 0.8073 (±0.0747), respectively. The lowest PIC values were detected at TPOX locus with the value of 0.5387 for both length- and sequence-based polymorphisms, whereas the highest values of length- and sequence-based polymorphisms were 0.8928 at PentaE locus and 0.9095 at D12S391 locus, respectively. The TPOX locus revealed the lowest PE value for both length- and sequence-based STRs with the value of 0.2909, and the D12S391 locus exhibited the highest PE values for both length- and sequence-based STRs with values of 0.8410 and 0.9096, respectively. The combined PE values for the length- and sequence-based genotypes were $1–3.4332 \times 10^{-11}$ and $1–1.0266 \times 10^{-12}$, respectively. The PD values of all A-STR loci were larger than 0.7760 at TPOX locus for both length- and sequence-based polymorphisms, while all the PD values were less than 0.9741 at PentaE locus based on length polymorphisms and less than 0.9788 at D21S11 locus based on sequence polymorphisms. Additionally, the combined PD value observed at the length and sequence levels were $1–3.0634 \times 10^{-30}$ and $1–8.0118 \times 10^{-33}$, respectively. The PM values were in the range from 0.0259 (PentaE) to 0.2240 (TPOX) for length-based STRs and from 0.0212 (D21S11) to 0.2240 (TPOX) for sequence-based STRs. The TPOX locus showed the lowest TPI value as 1.2500 for both length- and sequence-based STRs, and the D12S391 locus presented the highest TPI values as 6.4286 for length-based STRs and 11.25 for sequence-based STRs.

Further, the haplotypic results and forensic parameters of 24 Y-STRs revealed from the perspective of length and sequence levels are presented in **Supplementary Table S6**. The GD values for length-based polymorphisms spanned from 0.4793 at DYS391 to 0.9718 at DYS385a-b loci, and for sequence-based polymorphisms from 0.4793 at DYS391 to 0.9903 at DYF387S1 locus. Eleven of the 24 Y-STRs exhibited discrepancies in GD values between length- and sequence-based Y-STRs, of which the GD values increased from 0.6324 to 0.6425 at DYS389I, 0.7813 to 0.9311 at DYS389II, 0.7610 to 0.7893 at DYS390, 0.5181 to 0.6774 at DYS437, 0.5912 to 0.6372 at DYS438, 0.7279 to 0.7391 at DYS439, 0.7498 to 0.8656 at DYS448, 0.8209 to 0.8272 at DYS570, 0.8599 to 0.8754 at DYS612, 0.8237 to 0.8569 at DYS635, and 0.6857 to 0.6932 at Y-GATA-H4 locus. No discrepancies were observed between the length- and sequence-based results for the DC, PM, and HD values, which were 1.0000, 0.0111, and 1.000, respectively. The length- and

sequence-based haplotypes and forensic parameters for seven X-STRs in males are displayed in **Supplementary Table S7**. The GD values fluctuated from 0.5316 at DXS7423 to 0.8986 at DXS10135 locus for length-based polymorphisms and from 0.5316 at DXS7423 to 0.9176 at DXS10135 locus for sequence-based polymorphisms. For seven X-STR loci, different GD values for length- and sequence-based STRs were detected at four loci, including DXS8378, DXS7132, DXS10103, and DXS10135. The length-based polymorphisms were disclosed the same DC, PM, and HD values as the sequence-based polymorphisms, which were 1.0000, 0.0111, and 1.0000, respectively.

The allele frequencies and forensic parameters of 94 IISNPs are presented in **Supplementary Table S8**. The minor allele frequencies (MAF) approximately ranged from 0.1056 (rs2040411, T; rs733164, G) to 0.4944 (rs1028528, A). The MAF emerged 32.98% of the total loci for A alleles, 22.34% for C alleles, 20.21% for G alleles, and 24.47% for T alleles. The H_{obs} and H_{exp} values varied from 0.1889 (rs1355366, rs2056277, rs2107612, and rs733164) to 0.5778 (rs722290) and from 0.1900 (rs2056277 and rs740910) to 0.5030 (rs3780962, rs891700, rs2269355, and rs907100) with the average values of 0.4249 (±0.0958) and 0.4330 (±0.0842), respectively. The PIC values varied from 0.1710 at rs2056277 and rs740910 loci to 0.3750 at rs891700 locus with an average value of 0.3343 (±0.0532). The PD values spanned from 0.3242 (rs2056277) to 0.6649 (rs214955) with average value of 0.5708 (±0.0795), and the combined PD value for the 94 IISNPs was $1–7.3652 \times 10^{-36}$. The PE values ranged from 0.0268 (rs1355366, rs2056277, rs2107612, and rs733164) to 0.2651 (rs722290) with average value of 0.1403 (±0.0579). The lowest PM value was 0.3351 at rs214955 locus, whereas the highest value was 0.6758 at rs2056277 locus; and the average value was 0.4292 (±0.0795). The TPI values were in the range from 0.6164 (rs1355366, rs2056277, rs2107612, and rs733164) to 1.1842 (rs722290), and the average value was 0.8916 (±0.1354).

STR Sequence Variations Observed Using the MPS Method

The sequence- and length-based alleles and corresponding frequencies of 58 STRs (27 A-STRs, 24 Y-STRs, and 7 X-STRs) for the Northwest Hui group are listed in **Supplementary Table S9**. Additional alleles with sequence variants were revealed by using the MPS platform, which were indistinguishable using the CE method. In contrast to the length-based alleles, increasing rates of detected alleles on the sequence level were observed, which are presented in **Supplementary Table S10**. On the basis of length polymorphism alone, 451 distinct alleles were identified from 58 STRs in 90 unrelated Hui individuals. When the sequence variants were taken into consideration, the allelic diversities of 31 STRs, including 15 A-STRs, 12 Y-STRs, and 4 X-STRs, increased obviously, leading to a total allele number of 652 for 58 STRs in the Northwest Hui group. In detail, the allele numbers for A-STRs in the Hui group ranged from 5 (D3S1358 locus) to 15 (D18S51, FGA and PentaE loci) on the length polymorphism level, whereas 6 (D22S1045, TH01 and TPOX loci) to 33 (D12S391 locus) alleles were observed on the sequence level, which contributed 1 to 23 additional alleles. The D12S391 locus exhibited the most diversity with 230% increase in ratio. The allele numbers in four STRs, including D2S1338 (170.00%), D21S11 (158.33%), D13S317 (125.00%), and D3S1358 (120.00%) loci, increased by more than double. For the Y-STRs loci, a total of 155 and 231 different alleles were identified on the basis of length and sequence polymorphisms, respectively. The highest allelic numbers were 12 at DYS385a-b loci for the length-based level and 33 at DYF387S1 locus for the sequence-based level. The increased ratios of allelic numbers in three Y-STRs exceeded 100%, including DYF387S1 (266.67%), DYS389II (242.86%), and DYS448 (150.00%) loci. Nine Y-STRs displayed 16.67–83.33% extra sequence-based alleles compared with length-based alleles. Four X-STRs, DXS10135, DXS10103, DXS8378, and DXS7132 loci, showed both length and sequence polymorphisms with the increasing ratios ranging from 14.29% to 57.89%. No additional alleles were identified based on sequence polymorphisms in 12 A-STRs, 12 Y-STRs, and 3 X-STRs compared with the length-based polymorphisms, thus providing the same recognition capability as the traditional CE approach.

Phenotypic Predictive Analysis

The phenotypic SNPs utilized in this study mainly focus on pigmentation, including hair and eye colors. The predicted phenotype results for male individuals from the Northwest Hui group are displayed in **Supplementary Table S11** and **Supplementary Figure S2**. The hair colors contained four possible traits, including brown, red, black and blond colors, and the eye colors encompassed three possible traits, including brown, blue and intermediate colors. The tested 90 male individuals from the Northwest Hui group were predominantly revealed with black hair and brown eyes, and the percentages of black hair and brown eyes for all individuals were in the range from 52% to 96%, and 87% to 100%, respectively. The predicted results are roughly in accordance with the phenotypes of the studied Hui group.

Genetic Relationships Between the Northwest Hui Group and Reference Populations

Genetic Differentiation Analyses Between the Northwest Hui Group and Reference Populations Based on IISNPs

The genetic differentiations between the Northwest Hui group and worldwide reference populations were analyzed based on the overlapping IISNPs loci. The reference data were obtained from 2,504 individuals of 26 populations in the 1000 Genomes Project. In contrast with the 94 IISNPs provided by the ForenSeq™ DNA Signature Prep Kit, one locus named rs938283 was not found in the 1000 Genomes Project, thus leading to the use of 93 overlapping IISNPs in the following analyses. Serial plots, including a heatmap of ancestral allele frequencies, a NJ tree, two PCA plots, and a plot of population-specific divergences based on I_n statistic, were constructed, with the intention of determining the genetic discrimination abilities of these 93

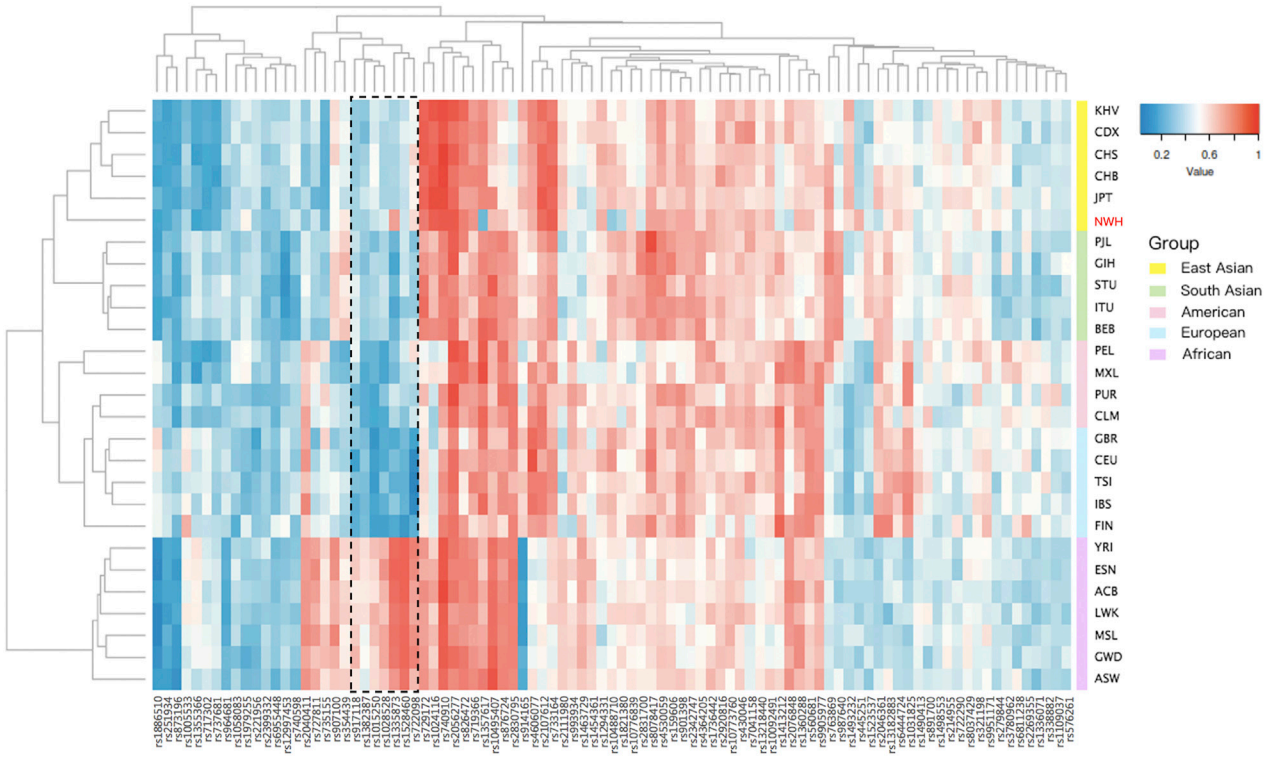


FIGURE 2 | Cluster heatmap constructed based on allele frequencies of 93 overlapping IISNP loci in the Northwest Hui and 26 reference populations. IISNP, identity-informative single nucleotide polymorphism.

IISNPs among the Northwest Hui group and reference populations, as well as the individual identification performances of these IISNPs in the Northwest Hui group.

The allele frequencies of 93 IISNPs were utilized to generate a cluster heatmap, facilitating the intuitive visualization of polymorphic distributions for these 93 same IISNPs in the Northwest Hui group and reference populations. As presented in **Figure 2**, the deeper blue indicated lower allelic frequency values, whereas the deeper red represented higher-frequency values. Populations were divided by different continental origins, including African, American, East Asian, European, and South Asian clusters, whereas some clusters exhibited different patterns of allelic frequency distributions. In particular, the frequency distributions in African populations were evidently distinct from those in the other four intercontinental clusters. For example, one set of IISNP loci marked using the dotted box in **Figure 2**, including the rs917118, rs1382387, rs1015250, rs1028528, rs1335873, rs1528460, and rs722098 loci, overwhelmingly displayed red in African populations, while they presented blue in the remaining populations from the other four continents. Additionally, the allelic frequency distribution in the Northwest Hui group showed the most similar pattern to East Asian populations, in accordance with the dendrogram classifications based on Euclidean distances in **Figure 2**. The MAF values of the Northwest Hui group and reference populations are plotted based on the allelic frequencies of these IISNPs in **Supplementary Figure S3**.

To further inspect whether the IISNP genetic markers could reveal the genetic relationships between the Northwest Hui group and worldwide reference populations, two PCA plots and one NJ tree were constructed. As presented in **Figure 3A**, the genetic relationships among populations were revealed by the first and second principal components accounted for 48.58% and 20.22%, respectively. According to the first discrimination component, all the populations were segregated into two group sets, i.e., the African group (deep blue) and the non-African group. From the perspective of the second discrimination component, the non-African populations were further categorized into four subgroups. One subgroup was distributed in the top right corner of the first quadrant, including populations from East Asia (yellow) and the Northwest Hui group (red); South Asian populations (green) belonging to another subgroup were located directly beneath the East Asian populations in the bottom right corner of the first quadrant; European populations (pink) were clustered as a subgroup in the lower right corner of the fourth quadrant; and American populations (purple) were scattered between the South Asian and European subgroups in the fourth quadrant. In addition, to dissect the genetic relationships more deeply, a PCA plot on the individual level was constructed. As shown in **Figure 3B**, the dots in various colors represented the individuals deriving from different biogeographical regions. All individuals were divided into two large cluster groups on the first principal component. Africans in deep blue occupied the right-hand side of the plot as the first

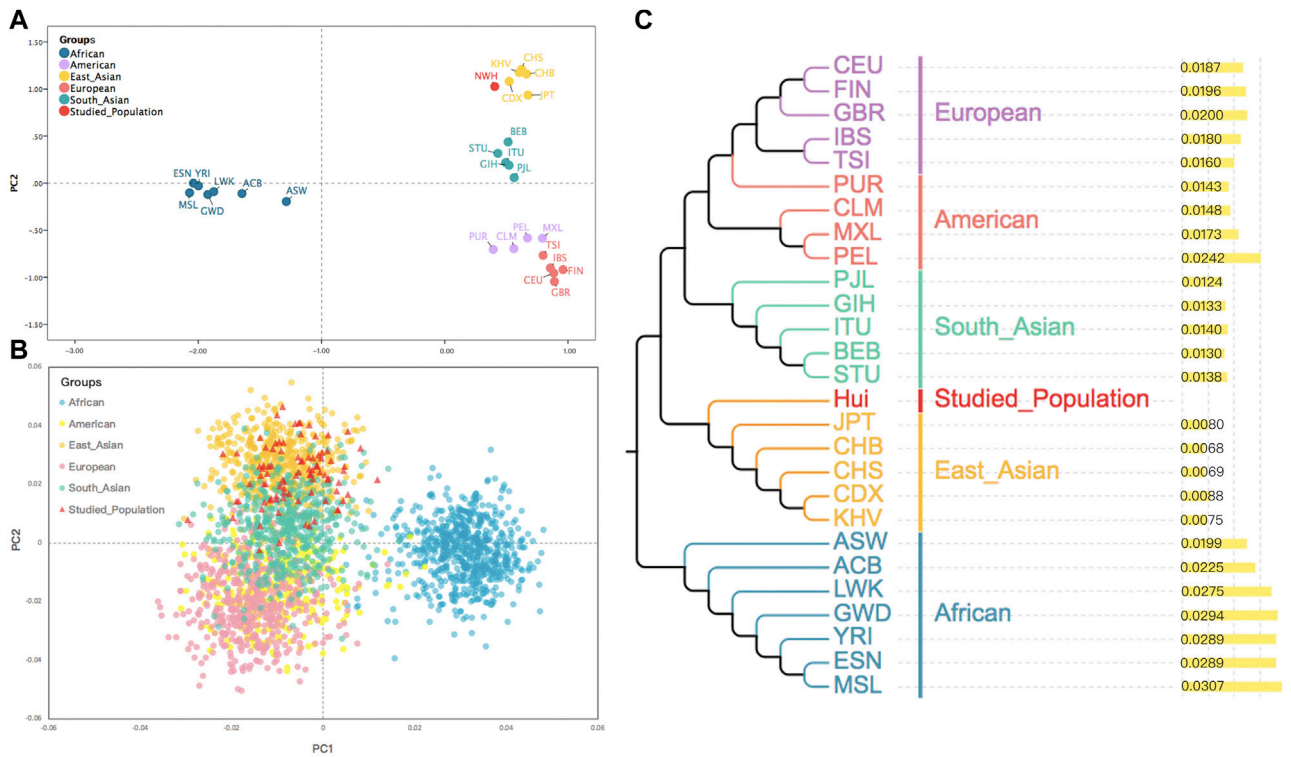


FIGURE 3 | (A) A PCA plot of the Northwest Hui group and the 26 reference populations based on the allele frequencies of these 93 same IISNPs. **(B)** A PCA plot for individuals from the Northwest Hui group and the 26 reference populations calculated based on the raw data of these 93 same IISNPs. **(C)** A neighbor-joining tree constructed based on the D_A genetic distances of the Northwest Hui group and the 26 reference populations. PCA, principal component analysis; IISNPs, identity-informative single-nucleotide polymorphisms.

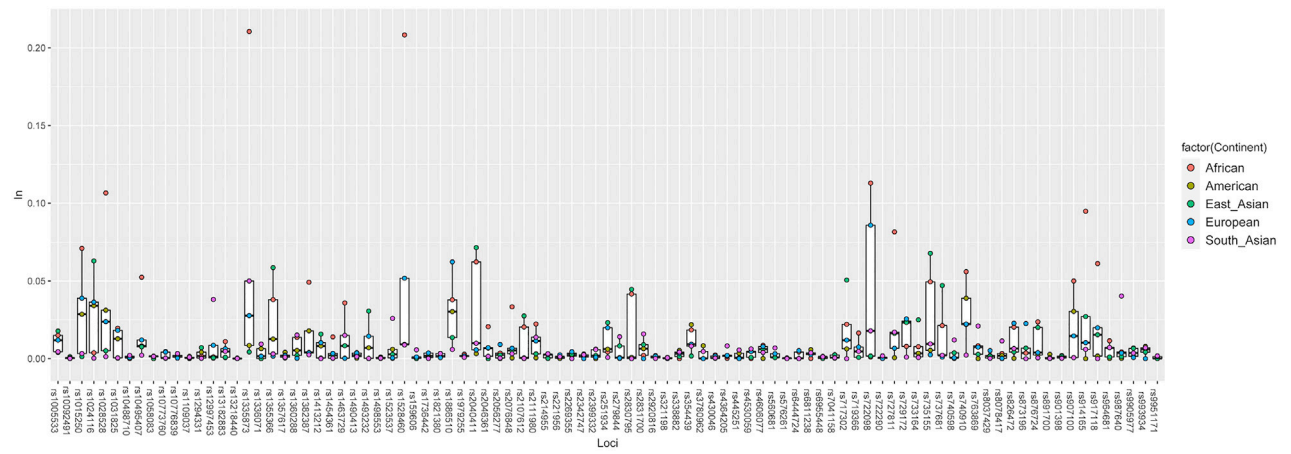


FIGURE 4 | A plot of population-specific divergences for five intercontinental population groups performed based on I_n statistic.

cluster. The other non-African cluster was further dispersed into three sub-clusters on the second principal component, namely, East Asians in orange, South Asians in green, and Europeans in pink, which superimposed without a clear boundary. The Americans in yellow were mainly clustered with South Asians and Europeans. Individuals from the Northwest Hui group (red) predominantly overlapped with East Asians, while sporadic Hui individuals were scattered among the South Asians. A NJ tree coupled with a histogram was subsequently constructed based on the D_A values in **Figure 3C**. All populations were divided into four distinct sub-branches, which were the European and American populations for the first, South Asian populations

for the second, the East Asian populations for the third, and the African populations for the fourth sub-branch. The cluster distributions of all populations were roughly concordant with their geographical regions, except for four American populations. Further, the Northwest Hui group was evidently clustered with East Asian populations and displayed the lowest divergencies with Han populations, including the CHB ($D_A = 0.0068$) and CHS ($D_A = 0.0069$) populations. Obviously, populations from Africa revealed the largest genetic distances from the Northwest Hui group with the D_A values ranging from 0.0199 to 0.0307.

Eventually, to determine the ability of the 93 IISNP loci on distinguishing populations, a plot of population-specific divergences for the five intercontinental clusters was constructed based on the I_n statistic. In **Figure 4**, populations from one continent had one corresponding I_n value at each IISNP locus, thus, there were five intercontinental I_n values at each IISNP locus. African populations generated the largest I_n values at 30 IISNP loci, ranging from 0.0200 at rs901398 locus to 0.2105 at rs1335873 locus, in comparison with the remaining intercontinental populations. African populations with relatively high I_n values always tended to separate from American, East Asian, European, and South Asian populations, especially at four loci (rs1335873, rs1528460, rs722098, and rs1028528) revealing the highest discrepancy I_n values ($I_n > 0.1$). Additionally, American populations encompassed 13 loci showing the largest discrepancy values, and the I_n values of American populations at these 13 loci varied from 0.0004 at rs13218440 locus to 0.0219 at rs354439 locus. A total of 19 loci showed the largest discrepancy values in East Asian populations, ranging from 0.0005 at rs576261 locus to 0.0715 at rs2040411 locus. European and South Asian populations exhibited the largest discrepancy values at 11 and 20 loci, respectively, of which the I_n values were in the range from 0.0015 (rs6955448) to 0.0623 (rs1886510) and from 0.0018 (rs722290) to 0.0403 (rs987640), respectively. The I_n values in American populations were in the range from 0.0000 at six loci (rs873196, rs914165, rs2399332, rs10092491, rs763869, and rs2056277) to 0.0389 at rs740910 locus with an average value of 0.0052 (±0.0084). East Asian populations generated I_n values ranging from 0.0000 at seven loci (rs8078417, rs2046361, rs13218440, rs1058083, rs1336071, rs1109037, and rs9951171) to 0.0715 at rs2040411 locus with an average value of 0.0091 (±0.0157). Nine loci, including rs2831700, rs2107612, rs445251, rs2342747, rs13218440, rs321198, rs993934, rs3780962, and rs2830795, presented the lowest discrepancy values ($I_n = 0.0000$) in European populations, while rs722098 locus yielded the highest discrepancy value ($I_n = 0.0858$), and the average I_n value was 0.0245 (±0.0190). The I_n values in South Asian populations varied from 0.0000 at eight loci (rs6444724, rs873196, rs917118, rs1413212, rs2830795, rs2269355, rs576261, and rs717302) to 0.0500 at rs1335873 locus with an average value of 0.0140 (±0.0118). The African populations offered the lowest discrepancy value ($I_n = 0.0000$) at eight loci (rs10488710, rs891700, rs338882, rs1493232, rs576261, rs445251, rs159606, and rs13218440), whereas the highest value was observed at rs1335873 locus ($I_n = 0.2105$), and the average value was 0.0198 (±0.0373).

The Comparisons of Y-STR Haplotype Polymorphisms Between the Northwest Hui Group and Reference Populations

The patrilineal genetic data of 59 populations, the Northwest Hui group and 58 reference populations, were employed in this study, which could be classified into six different biogeographical group sets, including 44 populations from East Asia, seven populations from West Asia, three populations from Central Asia, two populations from Africa, two populations from Europe, and one population from Australia. A total of 17 shared Y-STR loci were employed from these 59 populations, including DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385a, DYS385b, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, and YGATAH4. It is worth mentioning that three loci (DYS393, DYS456, and DYS458) from the abovementioned 17 Y-STR loci were absent in the 24 Y-STR markers from the ForenSeq™ DNA Signature Prep Kit. The genotype profiles of these three loci in the Northwest Hui group were replenished using the PCR and CE-based method.

As shown in **Figure 5**, a heatmap was constructed using the pairwise R_{st} values calculated based on the raw Y-STR data from the worldwide populations. The gradually deepening yellow indicated the increasing R_{st} values, whereas the deepening blue indicated the decreasing R_{st} values. In the heatmap, populations were labeled according to different linguistic families, containing eight populations from the Altaic family in yellow, two populations from the Afro-Asiatic family in grey, five populations from the Indo-European family in red, 42 populations from the Sino-Tibetan family in blue, and two populations from the Niger-Congo family in pink. All the Hui groups from Chinese different regions, including the studied Northwest Hui group (NWH), Ningxia Hui (NXH), Shaanxi Hui (SAXH), Henan Hui (HNH), Gansu Hui (GSH), Qinghai Hui (QHH) and Yunnan Hui (YNH), except for Sichuan Hui (SCH) were primarily clustered together, then gathered with the populations from East Asia, and subsequently grouped with Central and West Asian populations in deeper blue color.

To display the genetic relationships more clearly, a histogram of pairwise R_{st} values between the Northwest Hui group and the reference populations was plotted on the right-hand side of the heatmap. The Northwest Hui group retained extremely low genetic distances ($R_{st} < 0.01$) with other reference Hui groups from Chinese different regions, with the exception of SCH. In detail, four reference Hui groups from Chinese different regions, including GSH ($R_{st} = -0.0068$), HNH ($R_{st} = -0.0022$), QHH ($R_{st} = -0.0082$), and YNH ($R_{st} = -0.0081$), displayed negative R_{st} values; two reference Hui groups, NXH ($R_{st} = 0.0041$) and SAXH ($R_{st} = 0.0062$), showed extremely low genetic distances; and one reference Hui group, SCH ($R_{st} = 0.0371$), exhibited the lower genetic differentiation ($0.01 < R_{st} < 0.05$) from the Northwest Hui group. Intriguingly, the studied Hui group revealed clearly lower genetic differentiations from the Central and West Asian populations, especially for East Anatolian Turkey (EAT) ($R_{st} = 0.0038$). It is worth mentioning that one Chinese group named

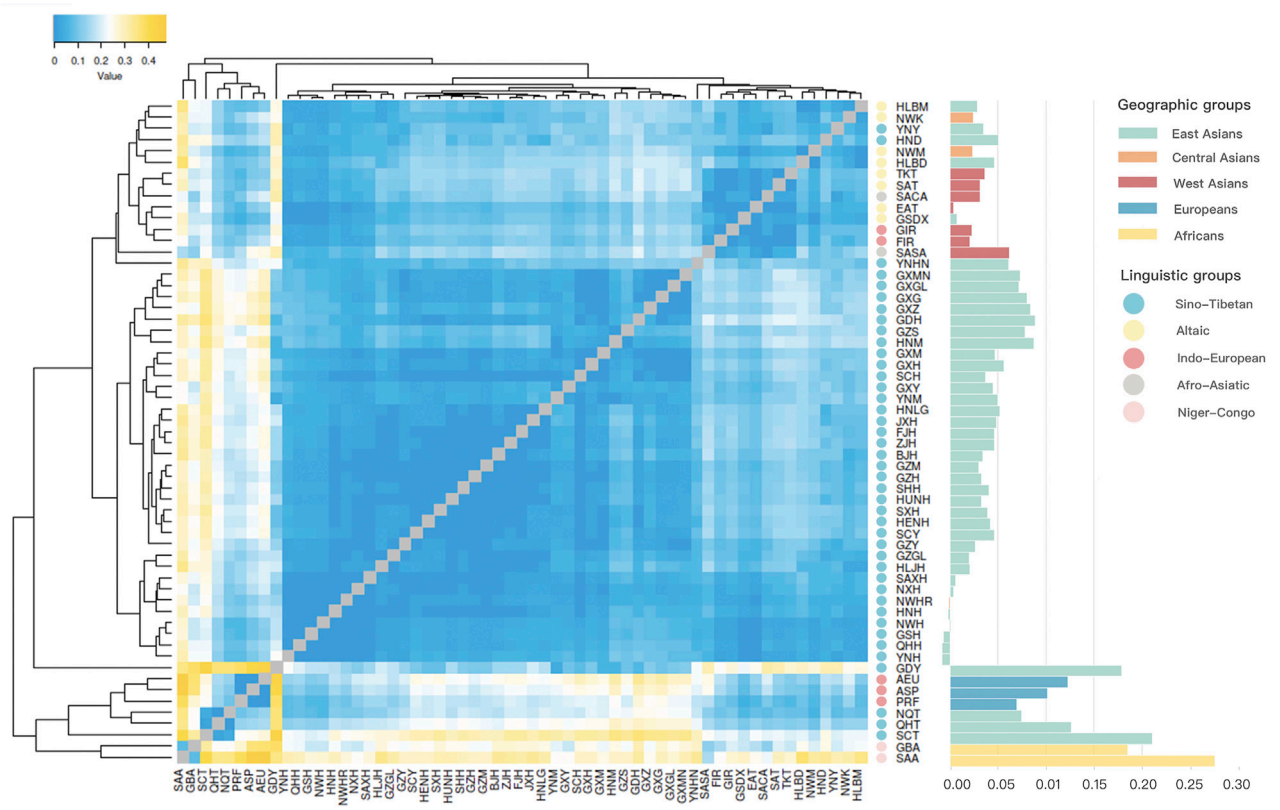


FIGURE 5 | A cluster heatmap constructed based on the pairwise R_{st} values of the Northwest Hui group and 58 reference populations, in combination with a histogram of pairwise R_{st} values between the Northwest Hui group and other reference populations in the right side.

Gansu Dongxiang (GSDX) (R_{st} = 0.0072) also exhibited extremely low genetic distance with the Northwest Hui group.

To further verify the genetic relationships between the Northwest Hui group and the reference populations, we conducted a MDS analysis for all the populations without the reference Hui groups based on pairwise R_{st} values. In **Figure 6**, the populations were roughly divided into several clusters as follows. Two African populations (blue) were scattered in the bottom of the plot, three European populations (yellow) were in the upper left corner, and the majority of the East Asian populations (green) were gathered in the upper right side. Evidently, the Northwest Hui group was located between East and West Asian populations, but was more prone to blend with West Asian (purple), Central Asian (red), and several East Asian (green) populations, including the Arab (SACA and SASA), Turkish (EAT, TKT, and SAT), and Iranian (FIR and GIR) populations from West Asia; the Northwest Kazakh (NWK) and Northwest Mongolian (NWM) groups from Central Asia; and Hulunbuir Mongolian (HLBM), Hulunbuir Daur (HLBD), GSDX, and other populations from East Asia.

Population Genetic Analyses Based on AISNPs

In the present study, we employed two sets of population data on the basis of the 55 overlapping AISNPs to uncover biogeographical ancestral information for the Northwest Hui group, including 26 reference populations from the 1000 Genomes Project and 109 reference populations from previous studies.

The PCA analysis of the Northwest Hui group and the 26 reference populations was constructed, from which the raw population data of 55 same AISNPs were available. To obtain a more intuitive display, the first two principal components were adopted to perform the PCA plot. As presented in **Figure 7**, the clustering pattern of individuals from five continents were revealed, preliminarily clarifying the ancestral determination performance of these 55 AISNPs. Africans (blue) located on the right-hand side of the plot were clearly distinguished from the other populations based on the first principal component. When the second principal component was taken into consideration, three clusters of individuals belonging to different continental regions were observed, which were East Asians (orange), South Asians (pink), and Europeans (purple). However, individuals from America (green) were largely overlapped with individuals from South Asia and Europe. The studied Hui individuals (red) were mainly superimposed with the East Asians.

To explicitly demonstrate the ancestral admixture patterns, we executed a STRUCTURE algorithm from $K = 2$ to $K = 6$ for the 26 reference populations and the Northwest Hui group based on the raw population data of 55 AISNPs in **Figure 8**. The different K values were equivalent to the number of predicted ancestral

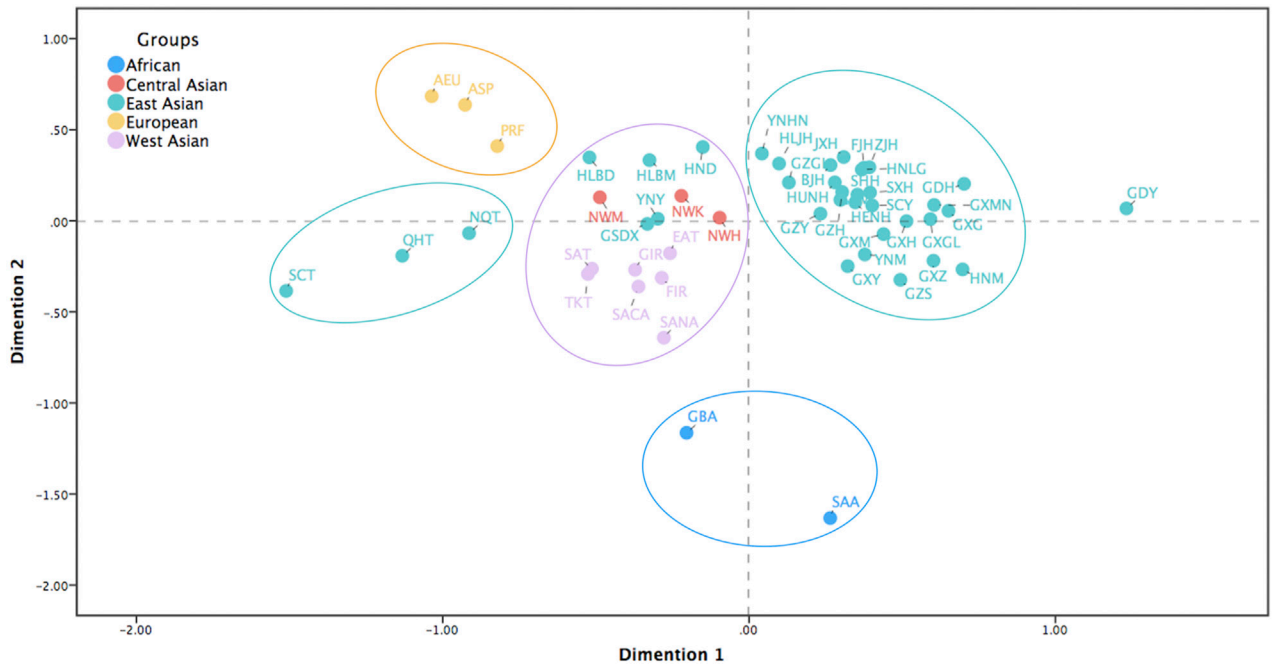


FIGURE 6 | MDS analysis for the Northwest Hui group and the reference populations without the reference Hui groups from Chinese different regions based on pairwise R_{st} values. MDS, multidimensional scaling.

FIGURE 7 | A PCA plot for individuals from the Northwest Hui group and the 26 reference populations based on the raw data for the same 55 AISNPs. PCA, principal component analysis; AISNPs, ancestry-informative single-nucleotide polymorphisms.

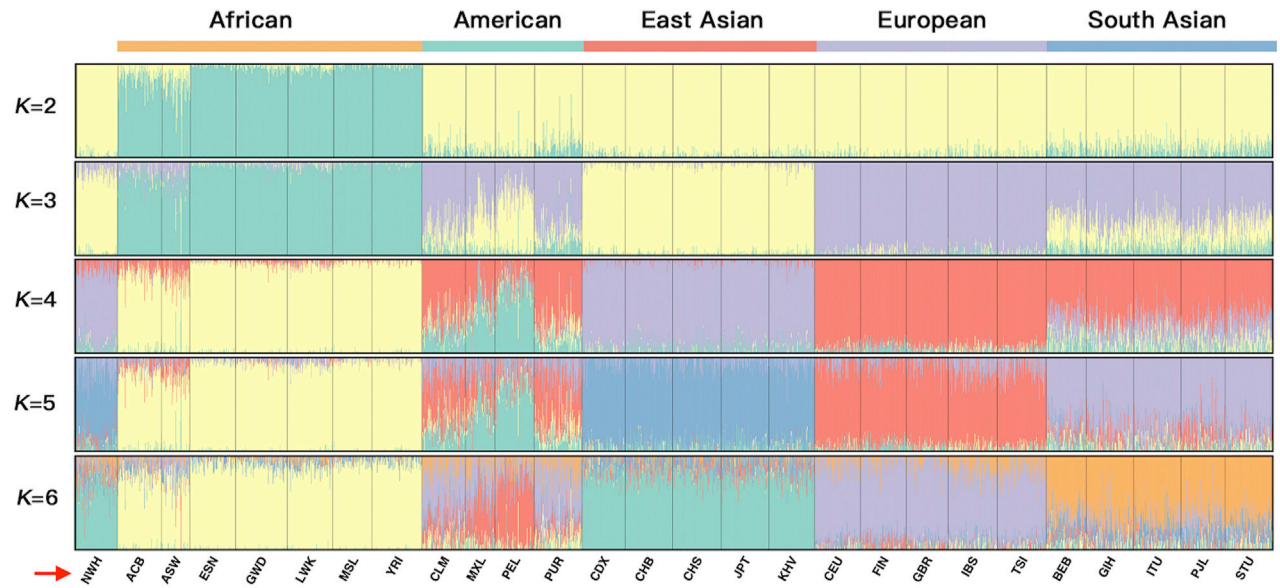


FIGURE 8 | A STRUCTURE plot for the 26 reference populations and the Northwest Hui group based on the raw population data for 55 AISNPs from $K = 2$ to $K = 6$. The K values are equivalent to the predicted ancestral components shown in different colors. AISNPs, ancestry-informative single-nucleotide polymorphisms.

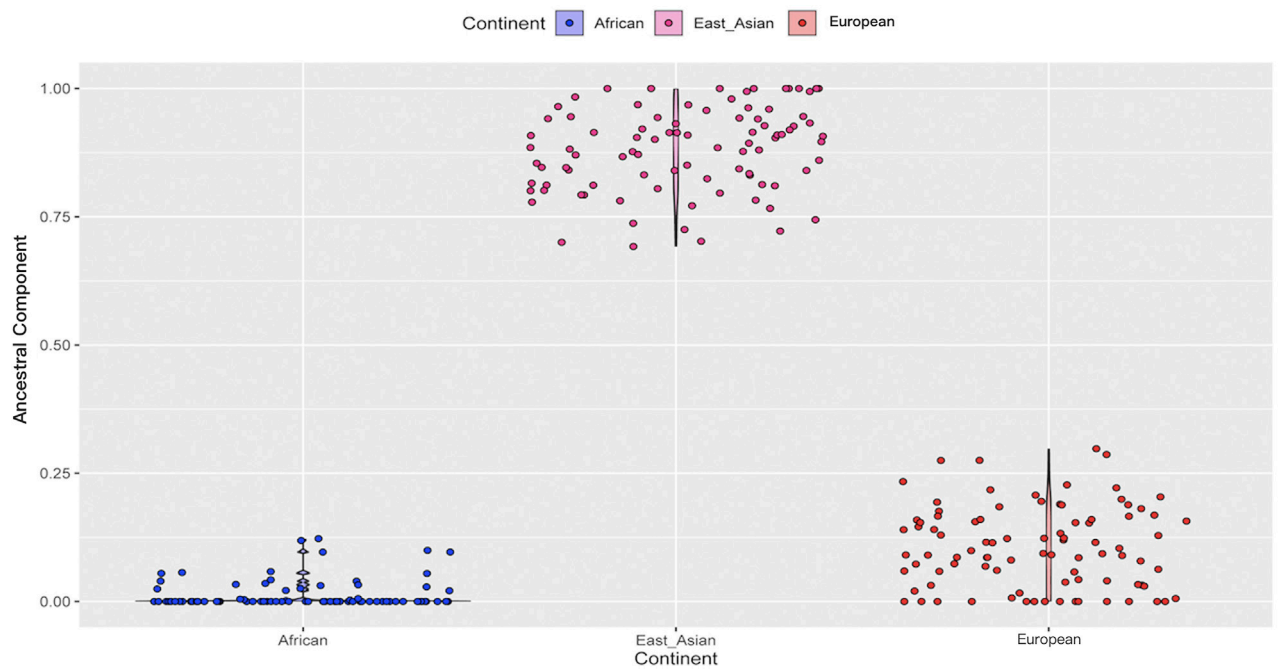


FIGURE 9 | The prediction of ancestral components for the Northwest Hui group on the basis of 55 AISNPs genotypes. AISNPs, ancestry-informative single-nucleotide polymorphisms.

components in different colors. Each individual was denoted by a vertical line partitioned into several segments corresponding to the contributions of different ancestral components. For example, when $K = 2$ (top line, **Figure 8**) was taken into consideration, all individuals were composed of two presumed ancestral components (yellow and green). Only populations from Africa, mainly covered by the green component, were distinctly separated from the non-African populations indicated by the large proportion of yellow. An additional ancestral component in purple was added at $K = 3$, which was the optimal number of

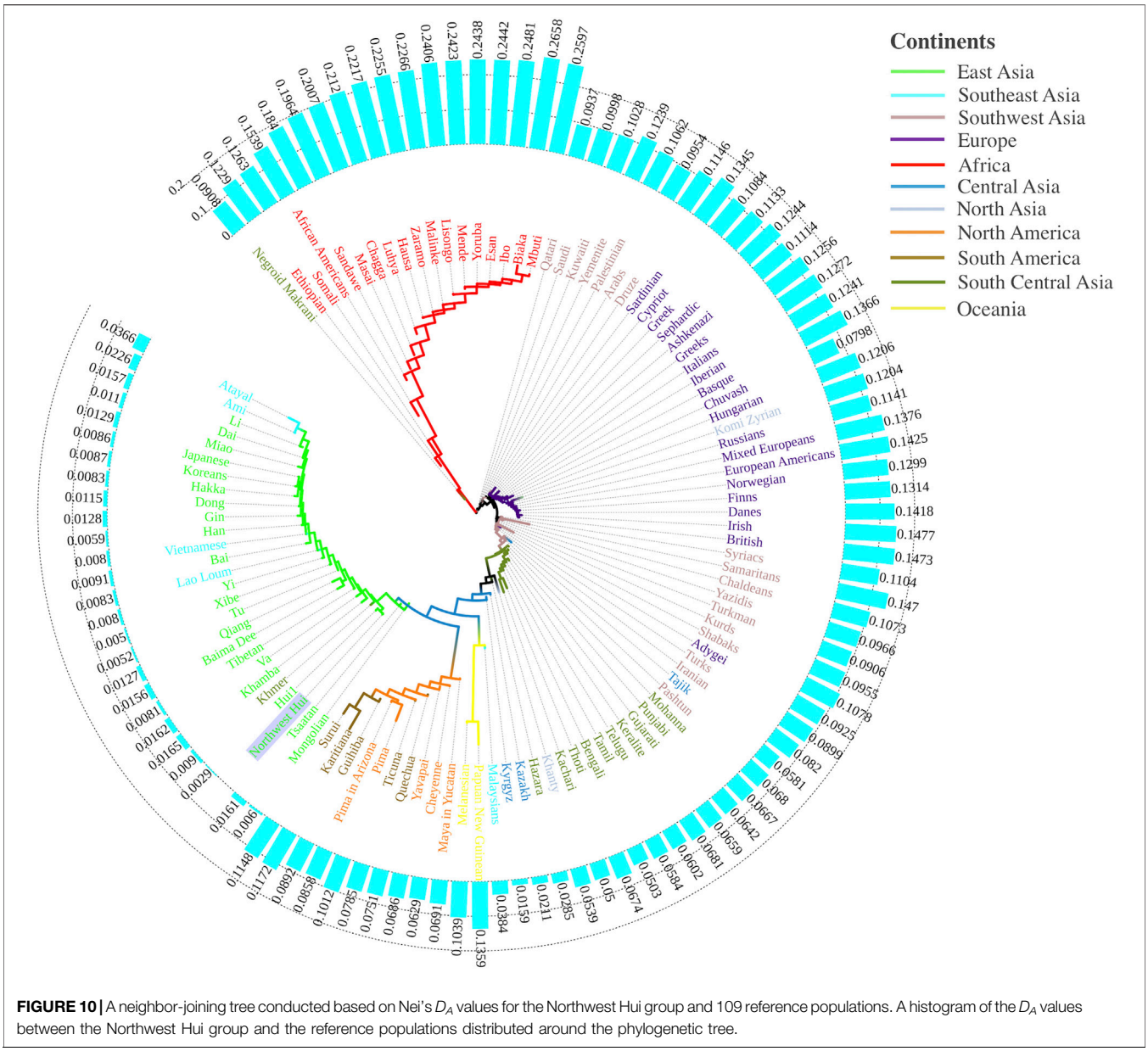


FIGURE 10 | A neighbor-joining tree conducted based on Nei's D_A values for the Northwest Hui group and 109 reference populations. A histogram of the D_A values between the Northwest Hui group and the reference populations distributed around the phylogenetic tree.

assumed ancestral components. In this case, the population clusters dominated by green, yellow, and purple appertained to Africa, East Asia, and Europe, respectively. Populations from America and South Asia exhibited similar ancestral components, the mixture of purple, yellow, and green, yet the proportions of the ancestral components were discrepant between the American and South Asian populations. At $K = 3$, the Northwest Hui group was dominated by yellow component and less purple component, unveiling an analogous admixture pattern with East Asian populations. Further, at $K = 4$, three population clusters from Africa, Europe, and East Asia were mainly distinguished by yellow, red, and purple components, respectively; an extra ancestral component in green could readily distinguish American populations from South Asian populations.

Based on the results of the STRUCTURE analyses, we subsequently conducted the prediction of ancestral components for the Northwest Hui group in **Figure 9**. A total of 90 Hui individuals represented by blue dots exhibited the lowest African ancestral component. On the contrary, 90 Hui individuals (pink dots) revealed the highest East Asian ancestral component, and the percentages of East Asian components exceeded 75% in the vast majority of Hui individuals. The Hui individuals shown with red dots disclosed small proportions of European ancestral components, with the percentages generally below 25%.

To comprehensively dissect the genetic structure of the Northwest Hui group, we subsequently implemented the data from 109 worldwide populations as the references, from which the population frequency data of 55 AISNPs were available. The Nei's D_A values among the studied Hui group and the 109 reference

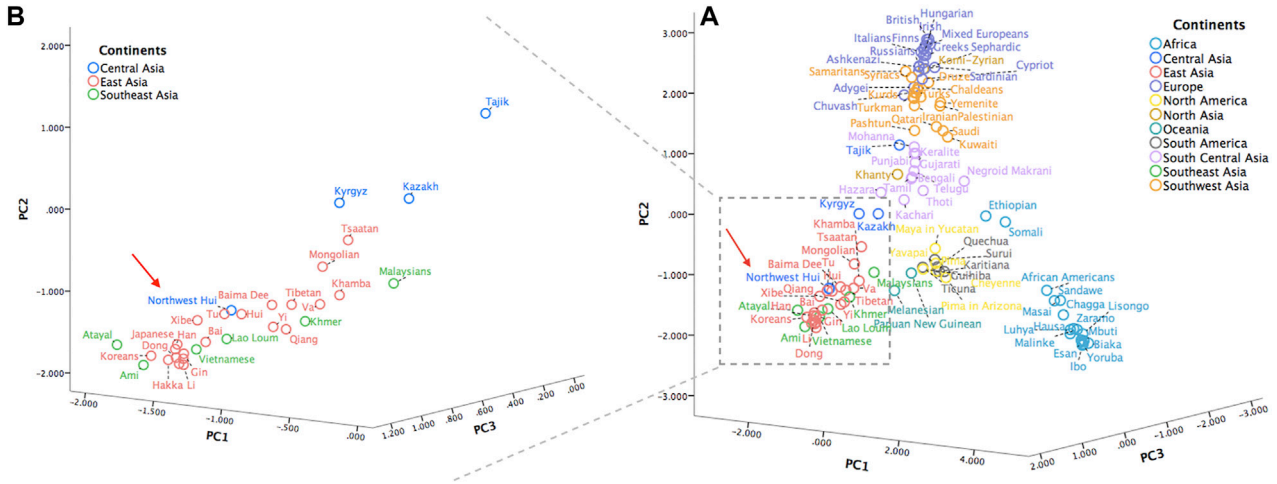


FIGURE 11 | (A) A PCA plot of the Northwest Hui group and 109 reference populations carried out based on the frequency data of 55 AISNPs. **(B)** A PCA plot of the Northwest Hui group and populations from East, Southeast, and Central Asia. PCA, principal component analysis; AISNPs, ancestry-informative single-nucleotide polymorphisms.

populations were evaluated and utilized to construct a NJ tree. As shown in **Figure 10**, the NJ tree was encircled by the histogram of D_A values between the Northwest Hui group and the reference populations. For the NJ plot, populations from different continents tended to congregate corresponding to their biogeographical regions, mainly including populations from East Asia (green sub-branches), America (orange and brown sub-branches), South Asia (dark green, pink and light blue sub-branches), Europe (purple sub-branches), and Africa (red sub-branches). The Northwest Hui group was chiefly assembled with populations from East Asia. In detail, many Chinese populations exhibited extremely low genetic differentiations ($D_A < 0.01$) with the Northwest Hui group. The smallest genetic distance was observed between the studied Hui group and the reference Hui group with $D_A = 0.0029$, followed by the Xibe group with $D_A = 0.0050$, the Tu group with $D_A = 0.0052$, the Southern Han population with $D_A = 0.0059$, the Mongolian group with $D_A = 0.0060$, the Yi group with $D_A = 0.0080$, the Tibetan group with $D_A = 0.0081$, the Hakka group with $D_A = 0.0083$, and the Bai group with $D_A = 0.0091$. In addition, three Southeast Asian populations (Vietnamese, Lao Loum, and Khmer) and two East Asian populations (Japanese and Korean) also displayed extremely low genetic distances with the Northwest Hui group. Populations from North America, South America, North Asia, and Europe predominantly manifested moderate differentiations ($0.05 < D_A < 0.15$) from the Northwest Hui group, whereas most populations from Africa primarily differentiated from the Northwest Hui group with large genetic distances ($0.15 < D_A < 0.25$).

To gain more insight into the population clustering pattern, a PCA plot of the Northwest Hui group and 109 reference populations was performed based on the frequency data of 55 identical AISNPs? (**Figure 11A**). The first principal component occupied 40.60% of the total variations, followed by the second principal component accounting for 37.97% and the third principal component accounting for 8.48%. Five population clusters, including populations from Africa (light blue), South Central Asia (light purple), Southwest Asia (orange), Europe (deep purple) America (yellow and grey), were clearly distinguished from each other. However, populations from East Asia (red), Southeast Asia (green), and Central Asia (blue) partially overlapped in the bottom left corner. To obtain clear insights of the relationships between the Northwest Hui group and the neighboring populations in the plot, another PCA plot was constructed (**Figure 11B**), concentrating on the populations from East, Southeast, and Central Asia. As anticipated, the studied Hui group was clustered with the majority of the East Asian populations and, more specifically, with the reference Hui and Tu groups.

DISCUSSION

Sequencing Performance

Considering all 231 genetic markers, the sequencing depths for Northwest Hui individuals roughly fluctuated from 77.14 (±21.23) at rs1736442 IISNP locus to 12,535.64 (±3,565.09) at TH01 locus. A disequilibrium in sequencing depth was clearly observed among the different kinds of genetic markers. These sequencing results were within the acceptable criteria defined by the manufacturer and approximately consistent with previous MPS-related studies (Churchill et al., 2017; Guo et al., 2017; Köcher et al., 2018; Hollard et al., 2019; Hussing et al., 2019; Fan et al., 2020). Specifically, partial genetic markers presenting the lowest sequencing depths in this study, such as DYS460, DXS10103, and rs1736442, were also reported to perform poorly in other studies (Almalki et al., 2017; Guo et al., 2017; Jäger et al., 2017; Silvia et al., 2017; Xavier and Parson, 2017; Hollard et al., 2019; Fan et al., 2020). It was speculated that the constant low detection of these markers might be in relation to amplicon length, low (below recommendations) copy number

input DNA, inaccurate library quantification, or overmultiplexing of samples (Guo et al., 2017).

Forensic Parameter Efficiencies for Various Genres of Genetic Markers

No significant departure from the HWE expectation was observed following the Bonferroni correction of 27 A-STR loci. In terms of the other forensic parameters of 27 A-STRs shown in **Figure 1**, the sequence-based polymorphisms provided relatively higher median values in comparison with the length-based polymorphisms, except for the PM parameter, which indicated that the A-STR loci detected at the sequence level might contain more genetic information compared to those at the length level. Previous reports in the literature corroborate that genetic markers with heterozygosity values over 0.5 within populations are appropriate for genetic diversity studies (Dávila et al., 2009; Sheriff and Alemayehu, 2018), illustrating that these 27 A-STR loci with high H_{obs} and H_{exp} values at both length and sequence levels were suitably informative for genetic studies. Additionally, we also discovered that the average H_{obs} values were close to, although slightly lower than, the average H_{exp} values from either the length or sequence perspective, possibly indicating no overall loss in heterozygosity (Araújo et al., 2006). The PIC values at the length and sequence levels were both greater than 0.5; according to Marshall et al., a locus revealing a PIC >0.5 could be treated as highly informative and as a polymorphic marker for genetic characterization and diversity studies (Marshall et al., 1998). The average PM value generated using the sequence-based method was lower than that generated using the length-based method, indicating that sequence-based genotypes might contain more genetic polymorphisms than length-based genotypes. The combined PD value observed at the sequence level was higher than that at the length level, denoting that sequence-based STRs could provide more discrimination power for individual identifications. The combined PE value for the sequence-based genotypes was greater than that for the length-based genotypes, indicating that more genetic variations revealed by sequencing increased the exclusive ability for parentage identification.

When analyzing 94 IISNPs in the Northwest Hui group, only three loci, rs1360288, rs2107612, and rs214955, significantly deviated from the HWE expectations, which were further conformed to the HWE expectations after Bonferroni corrections. The average values of other forensic parameters, including H_{obs}, H_{exp}, PIC, PD, PE, and TPI, were overwhelmingly lower than those of A-STRs. However, the combined PD value for the 94 IISNPs was higher than those of the 27 A-STRs at both the length and sequence levels. Previous reports also claimed that these 94 IISNPs could be comparable with the 27 A-STRs considering the discrimination power for individual identification (Wendt et al., 2016; Li et al., 2018; Delest et al., 2020; Fan et al., 2020).

STR Sequence Variations Observed Using the MPS Method

Compared with STR alleles characterized by the lengths of repeat regions, the detection of possible sequence variations in these repeat regions can effectively increase the polymorphism level of these alleles. In this study, the Northwest Hui group presented plentiful sequence variations on the basis of the MPS data. Specifically, the numbers of sequence-based alleles at eight STR loci were more than doubled in comparison with those of the length polymorphic alleles, including D3S1358, D13S317, D21S11, D2S1338, D12S391, DYS448, DYS389II, and DYF387S1, which were partially consistent with previously published studies (Gettings et al., 2015; Gettings et al., 2016; Novroski et al., 2016; Delest et al., 2020). For example, Novroski et al. claimed that D2S1338, D12S391, and D21S11 loci provided the most significant contribution to increasing allele diversity via sequence variations in repeat regions, focusing on African American, Caucasian, Hispanic, and Chinese populations (Novroski et al., 2016). D12S391 and DYF387S1 were determined to be the most polymorphic loci for the A-STRs and Y-STRs in French populations, respectively (Delest et al., 2020). These sequence variations enhanced the forensic efficiencies of these STR markers in the forensic applications, such as individual identification and kinship testing, which were generally consistent with previous studies (Hussing et al., 2019; Khubrani et al., 2019; Delest et al., 2020).

Genetic Relationships Between the Northwest Hui Group and Reference Populations

The Performances of IISNPs in Population Genetics

As depicted in **Figure 2**, certain discrepancies the allelic frequency distributions among the intercontinental populations were observed, especially for partial IISNP loci marked by the dotted box. However, the SNP loci utilized for forensic individual identifications are reported to display little allele frequency differentiation and high heterozygosity among the applied populations (Kidd et al., 2006; Yousefi et al., 2018). Further, MAF can also be considered when screening the SNP loci for individual identifications, which is defined as the frequency of the least common allele for each genetic marker in a given population (Huang et al., 2018; Yousefi et al., 2018). Various MAF thresholds were implemented for the selection of optimal IISNPs in previously published research. For example, Yousefi et al. chose the SNP loci with MAF value of more than 0.2 as one of the criteria for the individual identification (Yousefi et al., 2018). Huang et al. conducted a genome-wide SNP screening based on the HapMap and 1000 Genomes databases for forensic individual identifications, of which the MAF values were in the range from 0.35 to 0.43 (Huang et al., 2018). Briefly, SNP loci with higher MAF values are recommended for individual identifications, which also tend to produce higher heterozygosity values. Thus, for the purpose of individual identifications, SNP loci showing relatively homogeneous frequency distributions as well as high MAF values among tested populations are recommended.

In order to determine whether these studied IISNPs have abilities of population differentiations, we delineated two PCA plots based on the population and individual levels. The

population discrepancies on the PCA plots were readily discerned, particularly for African populations that were always clearly distinguished from the other four intercontinental populations. Conversely, populations from East Asia, America, South Asia, and Europe were prone to exhibit relatively close genetic relationships in comparison with the African populations based on these 93 IISNPs. This speculation is clearly verified by **Figure 3B**, where all the individuals other than the Africans formed inconspicuous clusters. For the NJ tree presented in **Figure 3C**, the Northwest Hui group was initially grouped with East Asian populations and exhibited the furthest genetic relationships from African populations. The abovementioned results indicated that the population genetic structures revealed by these 93 shared IISNPs conveyed certain disparities, which might be attributed to the uneven frequency distributions of some IISNP loci among intercontinental populations, especially between African and non-African populations.

To validate the notion mentioned above, we executed an I_n statistical analysis to evaluate the distinguishing abilities of 93 IISNPs across five intercontinental population groups. In the present study, African populations showed the largest discrepancy values at 30 of the 93 IISNP loci in comparison with the other four intercontinental population groups, in which the loci rs1335873, rs1528460, rs722098, and rs1028528 ranked in the top four with the discrepancy I_n values greater than 0.1. According to a previous report, molecular markers with high I_n values are more liable to infer population structure than those with low I_n values (Rosenberg et al., 2003). However, when the I_n statistic was utilized to screen ancestral information loci, the explicit threshold value was not clearly unified across different studies (Rosenberg et al., 2003; Xu et al., 2008; Phillips, 2015; Zeng et al., 2016; Jin et al., 2020). The recommended I_n threshold value ($I_n > 0.1$) for selecting ancestral markers could be acquired from a previous study (Jin et al., 2020), of which the criteria indicated that the four loci mentioned above might be suitable for ancestral information inferences. In contrast with the African populations, these 93 IISNPs provided overwhelmingly lower I_n values in the American, European, East Asian, and South Asian populations. The smaller I_n values denote that the tested genetic markers exhibit more similar allelic frequency distributions in the applied populations. In combination with the ubiquitously higher MAF values among the non-African populations in **Supplementary Figure S3**, we thus speculated that these IISNPs might yield a slightly better performance in the non-African populations on the purpose of individual identifications. Eventually, **Figure 3B** showed that the Hui individuals predominantly overlapped with East Asians; in combination with the high combined PD value of these IISNPs in the Northwest Hui group, these 93 IISNPs could perform well in identifying Northwest Hui individuals.

The Genetic Differentiations Assessed by Y-STRs Between the Northwest Hui and Reference Populations

To comprehensively dissect the patrilineal genetic landscape of the Hui group, we enrolled the reference population data of 17 shared Y-STRs from previously published research and the YHRD database to assess the population differentiations among the Northwest Hui group and reference populations.

A heatmap of pairwise R_{st} values among the applied 59 populations in **Figure 5** illustrated that all Hui groups from Chinese different regions displayed relatively close genetic relationships with East Asian populations, as well as some Central and West Asian populations. To be more specific, the histogram of pairwise R_{st} between the Northwest Hui group and the reference populations demonstrated that the Northwest Hui group revealed the lowest genetic differentiations from the reference Hui groups, except for SCH, which was congruous with the previous study (Xie et al., 2019). It was reported that the separated distribution of SCH and the Northwest Hui group was largely attributed to the different frequency distributions of the D, E, D, G, H, J1, R1a, R1b, and R2 sub-haplogroups (Xie et al., 2019). Intriguingly, according to the histogram, relatively intimate genetic relationships might exist between the Northwest Hui group and populations from Central and West Asia, as well as several East Asian populations, like GSDX. In the light of the MDS plot, the Northwest Hui group might have close genetic relationships with some Central and West Asian populations as well as several East Asian populations, such as the Dongxiang, Kazakh, Mongolian, Arab, Turkish, and Iranian populations. From the linguistic perspective, the abovementioned Arab populations belong to the Afro-Asiatic language family; Iranian populations belong to the Indo-European language family; and Turkish, Dongxiang, Kazakh, and Mongolian populations belong to the Altaic language family. The formation of a population is always accompanied by the corresponding establishment of its language family. Although the language family may not be able to exhaustively reflect the exact formation history of a population, from cultural and historical perspectives, it can still provide certain evidence as to the eventual formation history of the Hui group. In the present study, the Northwest Hui group appertained to the Sino-Tibetan family with Chinese as the predominant language (Gladney, 2020), which was more likely to cluster with populations from the Altaic and Indo-European language families based on the Y-STR analyses. Another study also reported that Chinese Hui groups maintained several Arabic and Persian phrases in their language (Dillon, 1999).

In light of historical documents, various ancestries may have participated in the eventual formation of the Chinese Hui group. Partial Islamic adherents from Central and West Asia, such as Persians, Arabs, and Turks, were encouraged to migrate to China during several medieval dynasties, especially the Yuan (1,271–1,368 AD) Dynasty ruled by the Mongolians. It was believed that these Muslims entered China via two routes: from Northwest China via the Silk Road and from Southeastern coastal areas via the Maritime Silk Road. The Silk Road served as a series of extensive inland trade routes connecting East and West Eurasia and promoted the exchanges of economies, cultures, politics, and religions between civilizations during the 2nd century BCE to the 18th century (Elisseeff, 2000; Gan et al., 2009). It was reported that most of these Muslims, including soldiers and traders, migrated to China via the Silk

Road and intermarried with local indigenous peoples, facilitating the ultimate formation of the Chinese-speaking Hui group (Elisseeff, 2000; Gan et al., 2009). The opinions on the ethnic origin of the Chinese Dongxiang group are relatively divergent according to different historians. In short, the Dongxiang was commonly described as the descendants of Mongolian troops (1,162–1,227 AD) who settled in the Hezhou region and possibly mixed with Sarts (Arab traders and Turkic-speaking city dwellers from Central Asia) (Schwarz, 1984; Wang et al., 2018). Accordingly, it is reasonable that the Northwest Hui group externalized genetically intimate relationships with the populations from Central and West Asia, as well as Dongxiang and Mongolian groups from China.

Ancestral Components of the Northwest Hui Group Revealed Using AISNPs

During the formation of a population, the genetic structure might be changed due to different historical events, such as migrations or intermarriages (Mellars, 2006; Duda and Jan Zrzavý, 2016), and changes in the genetic landscape of one population might distinguish it from other populations. Therefore, partial genetic variations found in modern populations can be treated as ancestral informative markers, such as AIM-STRs (Pereira et al., 2011; Phillips et al., 2011), AISNPs (Phillips et al., 2012; Kidd et al., 2014; Phillips et al., 2014; Jin et al., 2020), and AIM-InDels (Romanini et al., 2015), shedding light on the population evolutionary process.

The raw data of 26 reference populations were initially utilized to estimate the performance of 55 AISNPs in ancestral estimation and further disclose the genetic structure and ancestral bioinformation of the Northwest Hui group. As presented in **Figure 7**, the PCA result at the individual level indicated that individuals from East Asia, South Asia, Europe, and Africa could evidently be separated, while individuals from America were substantially overlapped with South Asians and Europeans. As depicted in a previous study, American populations displaying affinities with European populations might be largely attributed to the immigration of Europeans to the American continent since the discovery of America by Christopher Columbus in 1492 (Jordan et al., 2019). Since the 1700s, a large proportion of South Asians have immigrated to America and intermarried with indigenous Americans; consequently, over 3.4 million Americans can trace their ancestry back to South Asia according to the 2010 census (Perez and Hirschman, 2009). These historical records might partially explain why the Americans overlapped the Europeans and South Asians in the PCA plot. Other studies also reported that American populations from the 1000 Genomes Project might retain ancestral components from European, African, and indigenous American populations, which posed challenges in the ancestral analyses of American populations (Pakstis et al., 2019b; Jin et al., 2020). STRUCTURE analyses in **Figure 8** supported the abovementioned opinion, in which the genetic components revealed in American populations partially coincided with those from European and South Asian populations at $K = 3$ or $K = 4$. Thus, the above results demonstrated that these 55 AISNPs could perform well in the estimation of the individual ancestral components, especially for individuals from East Asia, South Asia, Africa, and Europe.

In terms of the studied Hui group, the Hui individuals were largely assembled with East Asians in the PCA plot, demonstrating that the Northwest Hui group might retain a significant amount of genetic components from East Asia. This opinion was also congruent with the results of the STRUCTURE analyses. At $K = 3$, the studied Hui group shared the similar proportion of genetic components with East Asian populations with a large area of yellow component; a trace amount of genetic component in purple which was mainly found in European populations was also detected in the studied Hui group. We further performed ancestral component prediction of the 90 Hui individuals. In **Figure 9**, the Northwest Hui group demonstrated the major genetic component from East Asia and the minor component from Europe, which were roughly consistent with the genetic component distributions of the Northwest Hui group at $K = 3$ in the STRUCTURE analysis (**Figure 8**).

In order to have a more comprehensive genetic interpretation of the Northwest Hui group, a phylogenetic tree was further delineated based on the allelic frequency data of the Northwest Hui group and 109 reference populations. In the phylogenetic tree plot, the Northwest Hui group was initially clustered with East Asian populations, exhibiting extremely low genetic differentiations from the Tu, Xibe, Yi, Bai, Han, Hakka, and Mongolian groups, etc. The intimate genetic relationships between the Hui group and some abovementioned Chinese populations were also reported in previous studies (Hong et al., 2007; Fang et al., 2018; He et al., 2018; Lan et al., 2018; Xie et al., 2018; Wang et al., 2019; Xie et al., 2019; Chen et al., 2020; Zhou et al., 2020). For example, the close genetic relationships between Hui groups and Han populations have been commonly reported from different perspectives, including Y-STRs and Y-SNPs (Wang et al., 2019; Xie et al., 2019), STRs (Fang et al., 2018), AISNPs (He et al., 2018), InDels (Xie et al., 2018; Zhou et al., 2020), and mitochondrial DNA (Chen et al., 2020). Other populations, such as Mongolian (Hong et al., 2007) and Xibe (Lan et al., 2018), were also reported to possibly have undergone gene exchange with Hui groups to some extent.

The relatively intimate genetic relationships between the Northwest Hui and East Asian populations were subsequently certificated by two PCA plots. The genetic affinities between the Hui group and East Asian populations were further supported by previously published studies (Yao et al., 2016; He et al., 2018; Xie et al., 2018; Wang et al., 2019; Xie et al., 2019; Jin et al., 2020; Zhou et al., 2020). In detail, He et al. investigated the genetic background of the NXH group using 165 AISNPs and discovered that the genetic components of the NXH group was predominantly contributed by the East Asian ancestral component (He et al., 2018). Another study based on InDels loci also claimed that East Asian populations provided a large proportion of ancestral component for the NXH group (Zhou et al., 2020). The GSH group was also explored by researchers and was found to exhibit substantial genetic intimacy with East Asian populations (Yao et al., 2016; Xie et al., 2018; Jin et al., 2020). To be more precise, the studied Hui group might have closer genetic relationships with the reference Hui and the Tu groups. The Chinese Tu group, predominately dwelling in Northwestern China, was officially recognized as one of the 56

ethnic groups in 1953. The majority of the Tu people speak the Monguor language, which pertains to the family of Mongolic language, one of the largest sub-branches of the Altaic language family. There are different opinions on the ethnogenesis of the Tu group, and various studies indicated that its genetic origin could be traced back to Tuyuhun Xianbei, to the Mongolian troop that came to the current Qinghai-Gansu region during the Mongolian conquest, or to the Han population (Schwarz, 1984; Hu, 2010). The Chinese Hui group was once conquered by the Yuan Dynasty (Schwarz, 1984; Xie and Shan, 2002) as aforementioned, and according to previous studies, the Hui group might exchange a large proportion of genes with the Chinese Han population due to the long-term intermarriages (Schwarz, 1984; Chen et al., 2020). These historical records and genetic studies might partially explain the intimate relationship between the Northwest Hui and Tu group revealed by the PCA plot.

However, the genetic background of the Northwest Hui group estimated using AISNPs showed certain divergence from that assessed using Y-STRs. In addition to the reference Hui groups from Chinese different regions, the Northwest Hui group depicted using Y-STRs was more prone to display closer genetic relationships with populations from Central and West Asia, as well as several Chinese groups. Additionally, our previous research based on the complete mitochondrial genome illustrated that the Northwest Hui group typically exhibited closer genetic relationships with East Asian populations, roughly concordant with the genetic result provided by 55 AISNPs in this study (Chen et al., 2020). Briefly, our previous study indicated that the genetic component of the maternal lineage that appeared in the Northwest Hui group might be predominantly derived from East Asia (Chen et al., 2020); the Y chromosome strictly follows the paternal inheritance with little recombination during the genetic process, so the paternal lineage of the Northwest Hui group may still retain some ancient genetic imprints closely related to Central or West Asian populations (Gladney, 1998). Considering the ancestral component prediction, the Northwest Hui group revealed a large proportion of ancestral components from East Asia and relatively little from Europe. Thus, we speculated that a sex bias phenomenon might exist in the long-term intermarriage process of the Northwest Hui group, denoting that males from West or Central Asia might intermarry females from East Asia, and eventually facilitating the formation of the current Hui group. This perspective is also supported by previous research (Wang et al., 2019) and historical material (Schwarz, 1984).

CONCLUSION

In this study, a total of 231 genetic markers were originally applied in 90 Hui male individuals dwelling in Northwest China. According to the current research, the studied genetic markers displayed satisfactory forensic performances in the Northwest Hui group, especially for the application of sequence polymorphisms which significantly enhanced the genetic diversities of STR genetic markers. Both 27 A-STRs and 94 IISNPs were polymorphic enough and could yield high forensic efficiencies in the forensic applications, such as paternity testings and individual identifications. Four of 94 IISNP loci, the rs1335873, rs1528460, rs722098, and rs1028528 loci, exhibited relatively larger discrepancies in the distributions of allelic frequencies among intercontinental populations, demonstrating that these loci might contain certain potential for the ancestry inferences. Studies based on the Y-STRs illustrated that the Northwest Hui group initially presented closest genetic relationships with reference Hui groups from Chinese different regions except for SCH and also externalized closer genetic relationships with populations from Central and West Asia, as well as several Chinese groups. Yet, the genetic structure revealed using the AISNPs indicated that the Northwest Hui group evidently displayed genetic affinities with populations from East Asia rather than those from Central or West Asia. In combination with the ancestral component estimation, it was proven that the Northwest Hui group contained a large proportion of ancestral components from East Asia and relatively little from Europe. The aforementioned results possibly indicated that the sex bias phenomenon of intermarriages might have existed in the formation of the Northwest Hui group, involving more intermarriages between males from Central and West Asia and females from East Asia.

DATA AVAILABILITY STATEMENT

The data is available in NCBI PRJNA738655.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of Xi'an Jiaotong University Health Science Center (approval number: 2019-1039; 2020-1382). The 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

Conceptualization: CC, XJ, and BZ; data curation: CC, XJ, XZ, WZ, RT, QX, and YY; formal analysis: CC, XJ, XZ, YG, QX, and ML; funding acquisition: BZ; investigation: WZ, RT, YY, and ML; methodology: CC, XZ, YG, RT, and AC; resources: BZ; software: XJ and YG; supervision: BZ; validation: XZ, WZ, AC, and QX; visualization: CC, XJ, and AC; writing—original draft: CC; writing—review and editing: CC, XJ, XZ, WZ, YG, RT, AC, QX, ML, YY, and BZ.

FUNDING

This research was supported by the National Natural Science Foundation of China (81772031, 81930055).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.705753/full#supplementary-material

REFERENCES

Al-jumaah, R., Musthafa, M. M., Al-Shaikh, M., and Badri, O. M., (2012). J. o. B.*Hussein* 11, 16539–16545.

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

Almalki, N., Chow, H. Y., Sharma, V., Hart, K., Siegel, D., and Wurmbach, E. (2017). Systematic Assessment of the Performance of Illumina's MiSeq FGx Forensic Genomics System. *ELECTROPHORESIS* 38, 846–854. doi:10.1002/elps.201600511

Araújo, A. M. d., Guimarães, S. E. F., Machado, T. M. M., Lopes, P. S., Pereira, C. S., Silva, F. L. R. d., et al. (2006). Genetic Diversity between Herds of Alpine and Saanen Dairy Goats and the Naturalized Brazilian Moxotó Breed. *Genet. Mol. Biol.* 29, 67–74. doi:10.1590/s1415-47572006000100014

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., et al. (2015). *Nature* 526, 68–74.

Balloux, F., and Lugon-Moulin, N. (2002). The Estimation of Population Differentiation with Microsatellite Markers. *Mol. Ecol.* 11, 55–65. doi:10.1046/j.0962-1083.2001.01436.x

Børsting, C., and Morling, N. (2015). Next Generation Sequencing and its Applications in Forensic Genetics. *Forensic Sci. Int. Genet.* 18, 78–89. doi:10.1016/j.fsigen.2015.02.002

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms. *Am. J. Hum. Genet.* 32, 314–331.

Bruijns, B., Tiggelaar, R., and Gardeniers, H. (2018). Massively Parallel Sequencing Techniques for Forensics: A Review. *Electrophoresis* 39, 2642–2654. doi:10.1002/elps.201800082

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., Li, Y., Tao, R., Jin, X., Guo, Y., Cui, W., et al. (2020). The Genetic Structure of Chinese Hui Ethnic Group Revealed by Complete Mitochondrial Genome Analyses Using Massively Parallel Sequencing. *Genes* 11, 1352. doi:10.3390/genes11111352

Churchill, J. D., Novroski, N. M. M., King, J. L., Seah, L. H., and Budowle, B. (2017). Population and Performance Analyses of Four Major Populations with Illumina's FGx Forensic Genomics System. *Forensic Sci. Int. Genet.* 30, 81–92. doi:10.1016/j.fsigen.2017.06.004

Churchill, J. D., Schmedes, S. E., King, J. L., and Budowle, B. (2016). Evaluation of the Illumina Beta Version ForenSeq DNA Signature Prep Kit for Use in Genetic Profiling. *Forensic Sci. Int. Genet.* 20, 20–29. doi:10.1016/j.fsigen.2015.09.009

Cifuentes, L. O., Martínez, E. H., Acuña, M. P., and Jonquera, H. G. (2006). Probability of Exclusion in Paternity Testing: Time to Reassess. *J. Forensic Sci.* 51 (2), 349–350. doi:10.1111/j.1556-4029.2006.00064.x

Dávila, S. G., Gil, M. G., Resino-Talaván, P., and Campo, J. L. (2009). Evaluation of Diversity between Different Spanish Chicken Breeds, a Tester Line, and a White Leghorn Population Based on Microsatellite Markers. *Poult. Sci.* 88, 2518–2525. doi:10.3382/ps.2009-00347

Delest, A., Godfrin, D., Chantrel, Y., Ulus, A., Vannier, J., Faivre, M., et al. (2020). Sequenced-based French Population Data from 169 Unrelated Individuals with Verogen's ForenSeq DNA Signature Prep Kit. *Forensic Sci. Int. Genet.* 47, 102304. doi:10.1016/j.fsigen.2020.102304

Dillon, M. (1999). *China's Muslim Hui Community: Migration, Settlement and Sects*. London: Curzon Press.

Duda, P., and Jan Zrzavý, Z. (2016). Human Population History Revealed by a Supertree Approach. *Sci. Rep.* 6, 29890. doi:10.1038/srep29890

Elisseeff, V. (2000). *The Silk Roads : Highways of Culture and Commerce*.

England, R., Nancollis, G., Stacey, J., Sarman, A., Min, J., and Harbison, S. (2020). Compatibility of the ForenSeq DNA Signature Prep Kit with Laser Microdissected Cells: An Exploration of Issues that Arise with Samples Containing Low Cell Numbers. *Forensic Sci. Int. Genet.* 47, 102278. doi:10.1016/j.fsigen.2020.102278

Fan, H., Du, Z., Wang, F., Wang, X., Wen, S.-Q., Wang, L., et al. (2020). The Forensic Landscape and the Population Genetic Analyses of Hainan Li Based on Massively Parallel Sequencing DNA Profiling. bioRxiv. doi:10.1101/2020.03.27.011064

Fang, Y., Guo, Y., Xie, T., Jin, X., Lan, Q., Zhou, Y., et al. (2018). Forensic Molecular Genetic Diversity Analysis of Chinese Hui Ethnic Group Based on a Novel STR Panel. *Int. J. Leg. Med.* 132, 1297–1299. doi:10.1007/s00414-018-1829-1

Fattorini, P., Previderé, C., Carboni, I., Marrubini, G., Sorçaburu-Cigliero, S., Grignani, P., et al. (2017). Performance of the ForenSeqTMDNA Signature Prep Kit on Highly Degraded Samples. *Electrophoresis* 38, 1163–1174. doi:10.1002/elps.201600290

Fisher, R. A. (1951). Standard Calculations for Evaluating a Blood-Group System. *Heredity* 5, 95–102. doi:10.1038/hdy.1951.5

Fullwood, M. J., Wei, C.-L., Liu, E. T., and Ruan, Y. (2009). Next-generation DNA Sequencing of Paired-End Tags (PET) for Transcriptome and Genome Analyses. *Genome Res.* 19, 521–532. doi:10.1101/gr.074906.107

Gan, F., Brill, R. H., and Tian, S. (2009). *Ancient Glass Research along the Silk Road*. World Scientific.

Gettings, K. B., Aponte, R. A., Vallone, P. M., and Butler, J. M. (2015). STR Allele Sequence Variation: Current Knowledge and Future Issues. *Forensic Sci. Int. Genet.* 18, 118–130. doi:10.1016/j.fsigen.2015.06.005

Gettings, K. B., Kiesler, K. M., Faith, S. A., Montano, E., Baker, C. H., Young, B. A., et al. (2016). Sequence Variation of 22 Autosomal STR Loci Detected by Next Generation Sequencing. *Forensic Sci. Int. Genet.* 21, 15–21. doi:10.1016/j.fsigen.2015.11.005

Gladney, D. C. (1998). *Ethnic Identity in China: The Making of a Muslim Minority Nationality*. Harcourt Brace College Publishers.

Gladney, D. C. (2020). *Muslim Chinese: Ethnic Nationalism in the People's Republic*. Netherlands: Brill.

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

Guo, F. (2017). Population Genetics for 17 Y-STR Loci in Hui Ethnic Minority from Liaoning Province, Northeast China. *Forensic Sci. Int. Genet.* 28, e36–e37. doi:10.1016/j.fsigen.2017.02.011

Guo, F., Yu, J., Zhang, L., and Li, J. (2017). Massively Parallel Sequencing of Forensic STRs and SNPs Using the Illumina ForenSeq DNA Signature Prep Kit on the MiSeq FGx Forensic Genomics System. *Forensic Sci. Int. Genet.* 31, 135–148. doi:10.1016/j.fsigen.2017.09.003

Guo, H., Yan, J., Jiao, Z., Tang, H., Zhang, Q., Zhao, L., et al. (2008). Genetic Polymorphisms for 17 Y-Chromosomal STRs Haplotypes in Chinese Hui Population. *Leg. Med.* 10, 163–169. doi:10.1016/j.legalmed.2007.11.001

He, G., Wang, Z., Wang, M., Luo, T., Liu, J., Zhou, Y., et al. (2018). Forensic Ancestry Analysis in Two Chinese Minority Populations Using Massively Parallel Sequencing of 165 Ancestry-Informative SNPs. *Electrophoresis* 39, 2732–2742. doi:10.1002/elps.201800019

Hollard, C., Ausset, L., Chantrel, Y., Jullien, S., Clot, M., Faivre, M., et al. (2019). Automation and Developmental Validation of the ForenSeq DNA Signature Preparation Kit for High-Throughput Analysis in Forensic Laboratories. *Forensic Sci. Int. Genet.* 40, 37–45. doi:10.1016/j.fsigen.2019.01.010

Hong, W., Chen, S., Shao, H., Fu, Y., Hu, Z., and Xu, A. (2007). HLA Class I Polymorphism in Mongolian and Hui Ethnic Groups from Northern China. *Hum. Immunol.* 68, 439–448. doi:10.1016/j.humimm.2007.01.020

Hu, A. J. (2010). An Overview of the History and Culture of the Xianbei ('Monguor'/'Tu'). *Asian Ethn.* 11, 95–164. doi:10.1080/14631360903531958

Huang, E., Liu, C., Zheng, J., Han, X., Du, W., Huang, Y., et al. (2018). Genome-wide Screen for Universal Individual Identification SNPs Based on the HapMap and 1000 Genomes Databases. *Sci. Rep.* 8, 5553. doi:10.1038/s41598-018-23888-0

Hussing, C., Bytyci, R., Huber, C., Morling, N., and Børsting, C. (2019). The Danish STR Sequence Database: Duplicate Typing of 363 Danes with the ForenSeq DNA Signature Prep Kit. *Int. J. Leg. Med.* 133, 325–334. doi:10.1007/s00414-018-1854-0

Illumina (2015). *Forenseq™ Universal Analysis Software Guide, Document #15053876v01*. Available at: https://verogen.com/wp-content/uploads/2018/08/ForenSeq-Univ-Analysis-SW-Guide-VD2018007-A.pdf.

Jäger, A. C., Alvarez, M. L., Davis, C. P., Guzmán, E., Han, Y., Way, L., et al. (2017). Developmental Validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories. *Forensic Sci. Int. Genet.* 28, 52–70. doi:10.1016/j.fsigen.2017.01.011

Jin, X. Y., Guo, Y. X., Chen, C., Cui, W., Liu, Y. F., Tai, Y. C., et al. (2020). Ancestry Prediction Comparisons of Different AISNPs for Five Continental Populations and Population Structure Dissection of the Xinjiang Hui Group Via a Self-Developed Panel. *Genes (Basel)*, 11 (5), 505. doi:10.3390/genes11050505

Jordan, I. K., Rishishwar, L., and Conley, A. B. (2019). Native American Admixture Recapitulates Population-specific Migration and Settlement of the continental United States. *Plos Genet.* 15, e1008225. doi:10.1371/journal.pgen.1008225

Just, R. S., Moreno, L. I., Smerick, J. B., and Irwin, J. A. (2017). Performance and Concordance of the ForenSeq System for Autosomal and Y Chromosome Short Tandem Repeat Sequencing of Reference-type Specimens. *Forensic Sci. Int. Genet.* 28, 1–9. doi:10.1016/j.fsigen.2017.01.001

Khubrani, Y. M., Hallast, P., Jobling, M. A., and Wetton, J. H. (2019). Massively Parallel Sequencing of Autosomal STRs and Identity-Informative SNPs Highlights Consanguinity in Saudi Arabia. *Forensic Sci. Int. Genet.* 43, 102164. doi:10.1016/j.fsigen.2019.102164

Kidd, K. K., Pakstis, A. J., Speed, W. C., Grigorenko, E. L., Kajuna, S. L. B., Karoma, N. J., et al. (2006). Developing a SNP Panel for Forensic Identification of Individuals. *Forensic Sci. Int.* 164, 20–32. doi:10.1016/j.forsciint.2005.11.017

Kidd, K. K., Speed, W. C., Pakstis, A. J., Furtado, M. R., Fang, R., Madbouly, A., et al. (2014). Progress toward an Efficient Panel of SNPs for Ancestry Inference. *Forensic Sci. Int. Genet.* 10, 23–32. doi:10.1016/j.fsigen.2014.01.002

Köcher, S., Müller, P., Berger, B., Bodner, M., Parson, W., Roewer, L., et al. (2018). Inter-laboratory Validation Study of the ForenSeq DNA Signature Prep Kit. *Forensic Sci. Int. Genet.* 36, 77–85. doi:10.1016/j.fsigen.2018.05.007

Lan, Q., Chen, J., Guo, Y., Xie, T., Fang, Y., Jin, X., et al. (2018). Genetic Structure and Polymorphism Analysis of Xinjiang Hui Ethnic Minority Based on 21 STRs. *Mol. Biol. Rep.* 45, 99–108. doi:10.1007/s11033-018-4143-6

Li, H., Zhao, X., Ma, K., Cao, Y., Zhou, H., Ping, Y., et al. (2017). Applying Massively Parallel Sequencing to Paternity Testing on the Ion Torrent Personal Genome Machine. *Forensic Sci. Int. Genet.* 31, 155–159. doi:10.1016/j.fsigen.2017.09.007

Li, R., Li, H., Peng, D., Hao, B., Wang, Z., Huang, E., et al. 2018, 38.

Liu, Y., Wen, S., Guo, L., Bai, R., Shi, M., and Li, X. (2018). Haplotype Data of 27 Y-STRs Analyzed in the Hui and Tujia Ethnic Minorities from China. *Forensic Sci. Int. Genet.* 35, e7–e9. doi:10.1016/j.fsigen.2018.04.006

Liu, Y., Yu, T., Mei, S., Jin, X., Lan, Q., Zhou, Y., et al. (2020). Forensic Characteristics and Genetic Affinity Analyses of Xinjiang Mongolian Group Using a Novel Six Fluorescent Dye-Labeled Typing System Including 41 Y-STRs and 3 Y-InDels. *Mol. Genet. Genomic Med.* 8, e1097. doi:10.1002/mgg3.1097

Liu, Y., Yang, J., Li, Y., Tang, R., Yuan, D., Wang, Y., et al. (2021). 12.

Ma, X., Sun, R., and Hao, C. (2017). Polymorphism of 15 Short Tandem Repeat Loci in Hui Population of Ningxia Tongxin District. *J. forensic Leg. Med.* 52, 168–171. doi:10.1016/j.jflm.2017.08.014

Malaspinas, A.-S., Slatkin, M., and Song, Y. S. (2011). Match Probabilities in a Finite, Subdivided Population. *Theor. Popul. Biol.* 79, 55–63. doi:10.1016/j.tpb.2011.01.003

Marshall, T. C., Slate, J., Kruuk, L. E. B., and Pemberton, J. M. (1998). Statistical Confidence for Likelihood-based Paternity Inference in Natural Populations. *Mol. Ecol.* 7, 639–655. doi:10.1046/j.1365-294x.1998.00374.x

Mellars, P. (2006). Going East: New Genetic and Archaeological Perspectives on the Modern Human Colonization of Eurasia. *Science* 313, 796–800. doi:10.1126/science.1128402

Meng, H.-T., Han, J.-T., Zhang, Y.-D., Liu, W.-J., Wang, T.-J., Yan, J.-W., et al. (2014). Diversity Study of 12 X-Chromosomal STR Loci in Hui Ethnic from China. *Electrophoresis* 35, 2001–2007. doi:10.1002/elps.201400045

Nei, M. (1973). Analysis of Gene Diversity in Subdivided Populations. *Proc. Natl. Acad. Sci.* 70, 3321–3323. doi:10.1073/pnas.70.12.3321

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press.

Novroski, N. M. M., King, J. L., Churchill, J. D., Seah, L. H., and Budowle, B. (2016). Characterization of Genetic Sequence Variation of 58 STR Loci in Four Major Population Groups. *Forensic Sci. Int. Genet.* 25, 214–226. doi:10.1016/j.fsigen.2016.09.007

Pakstis, A. J., Gurkan, C., Dogan, M., Balkaya, H. E., Dogan, S., Neophytou, P. I., et al. (2019). Genetic Relationships of European, Mediterranean, and SW Asian Populations Using a Panel of 55 AISNPs. *Eur. J. Hum. Genet.* 27, 1885–1893. doi:10.1038/s41431-019-0466-6

Pakstis, A. J., Haigh, E., Cherni, L., ElGaaied, A. B. A., Barton, A., Evsanaa, B., et al. (2015). 52 Additional Reference Population Samples for the 55 AISNP Panel. *Forensic Sci. Int. Genet.* 19, 269–271. doi:10.1016/j.fsigen.2015.08.003

Pakstis, A. J., Kang, L., Liu, L., Zhang, Z., Jin, T., Grigorenko, E. L., et al. (2017). Increasing the Reference Populations for the 55 AISNP Panel: the Need and Benefits. *Int. J. Leg. Med.* 131, 913–917. doi:10.1007/s00414-016-1524-z

Pakstis, A. J., Speed, W. C., Soundararajan, U., Rajeevan, H., Kidd, J. R., Li, H., et al. (2019). Population Relationships Based on 170 Ancestry SNPs from the Combined Kidd and Seldin Panels. *Sci. Rep.* 9, 18874. doi:10.1038/s41598-019-55175-x

Parson, W., Ballard, D., Budowle, B., Butler, J. M., Gettings, K. B., Gill, P., et al. (2016). Massively Parallel Sequencing of Forensic STRs: Considerations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on Minimal Nomenclature Requirements. *Forensic Sci. Int. Genet.* 22, 54–63. doi:10.1016/j.fsigen.2016.01.009

Pereira, L., Alshamali, F., Andreassen, R., Ballard, R., Chantratita, W., Cho, N. S., et al. (2011). PopAffiliator: Online Calculator for Individual Affiliation to a Major Population Group Based on 17 Autosomal Short Tandem Repeat Genotype Profile. *Int. J. Leg. Med.* 125, 629–636. doi:10.1007/s00414-010-0472-2

Perez, A. D., and Hirschman, C. (2009). The Changing Racial and Ethnic Composition of the US Population: Emerging American Identities. *C. Hirschman* 35, 1–51. doi:10.1111/j.1728-4457.2009.00260.x

Phillips, C., Fernandez-Formoso, L., Garcia-Magariños, M., Porras, L., Tvedebrink, T., Amigo, J., et al. (2011). Analysis of Global Variability in 15 Established and 5 New European Standard Set (ESS) STRs Using the CEPH Human Genome Diversity Panel. *Forensic Sci. Int. Genet.* 5, 155–169. doi:10.1016/j.fsigen.2010.02.003

Phillips, C., Fondevila, M., and Lareau, M. V. (2012). A 34-plex Autosomal SNP Single Base Extension Assay for Ancestry Investigations. *Methods Mol. Biol. (Clifton, N.J.)* 830, 109–126. doi:10.1007/978-1-61779-461-2_8

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., Parson, W., Lundsberg, B., Santos, C., Freire-Aradas, A., Torres, M., et al. (2014). Building a Forensic Ancestry Panel from the Ground up: The EUROFORGEN Global AIM-SNP Set. *Forensic Sci. Int. Genet.* 11, 13–25. doi:10.1016/j.fsigen.2014.02.012

Romanini, C., Romero, M., Salado Puerto, M., Catelli, L., Phillips, C., Pereira, R., et al. (2015). Ancestry Informative Markers: Inference of Ancestry in Aged Bone Samples Using an Autosomal AIM-Indel Multiplex. *Forensic Sci. Int. Genet.* 16, 58–63. doi:10.1016/j.fsigen.2014.11.025

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, 1402–1422. doi:10.1086/380416

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA Sequencing with Chain-Terminating Inhibitors. *Proc. Natl. Acad. Sci.* 74, 5463–5467. doi:10.1073/pnas.74.12.5463

Schwarz, H. G. (1984). *The Minorities Of Northern China: A Survey, Center for East Asian Studies*. Bellingham, WA: Western Washington University, East Asian Studies Press, Vol. 8.

Sharma, V., Jani, K., Khosla, P., Butler, E., Siegel, D., and Wurmbach, E. (2019). Evaluation of ForenSeq™ Signature Prep Kit B on Predicting Eye and Hair Coloration as Well as Biogeographical Ancestry by Using Universal Analysis Software (UAS) and Available Web-Tools. *Electrophoresis* 40, 1353–1364. doi:10.1002/elps.201800344

Sharma, V., van der Plaat, D. A., Liu, Y., and Wurmbach, E. (2020). Analyzing Degraded DNA and Challenging Samples Using the ForenSeq DNA Signature Prep Kit. *Sci. Justice* 60, 243–252. doi:10.1016/j.scijus.2019.11.004

Sheriff, O., and Alemayehu, K. (2018). Genetic Diversity Studies Using Microsatellite Markers and Their Contribution in Supporting Sustainable Sheep Breeding Programs: A Review. *Cogent Food Agric.* 4, 1459062. doi:10.1080/23311932.2018.1459062

Shete, S., Tiwari, H., and Elston, R. C. (2000). On Estimating the Heterozygosity and Polymorphism Information Content Value. *Theor. Popul. Biol.* 57, 265–271. doi:10.1006/tpbi.2000.1452

Silvia, A. L., Shugarts, N., and Smith, J. (2017). A Preliminary Assessment of the ForenSeq FGx System: Next Generation Sequencing of an STR and SNP Multiplex. *Int. J. Leg. Med.* 131, 73–86. doi:10.1007/s00414-016-1457-6

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi:10.1093/molbev/mst197

Team, R. C. (2016). *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Tillmar, A. (2010). Populations and Statistics in Forensic Genetics [Internet]. PhD dissertation. Sweden: Linköping University Electronic Press, Linköping University Medical Dissertations. Available at: http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-54742.

Vandeputte, M. (2012). An Accurate Formula to Calculate Exclusion Power of Marker Sets in Parentage Assignment. *Genet. Sel. Evol.* 44, 36. doi:10.1186/1297-9686-44-36

Wang, C. C., Lu, Y., Kang, L., Ding, H., Yan, S., Guo, J., et al. (2019). The Massive Assimilation of Indigenous East Asian Populations in the Origin of Muslim Hui People Inferred from Paternal Y Chromosome. *Am. J. Phys. Anthropol.* 169, 341–347. doi:10.1002/ajpa.23823

Wang, J., Wen, S., Shi, M., Liu, Y., Zhang, J., Bai, R., et al. (2018). Haplotype Structure of 27 YfilerPlus Loci in Chinese Dongxiang Ethnic Group and its Genetic Relationships with Other Populations. *Forensic Sci. Int. Genet.* 33, e13–e16. doi:10.1016/j.fsigen.2017.12.014

Wendt, F. R., Churchill, J. D., Novroski, N. M. M., King, J. L., Ng, J., Oldt, R. F., et al. (2016). Genetic Analysis of the Yavapai Native Americans from West-Central Arizona Using the Illumina MiSeq FGx Forensic Genomics System. *Forensic Sci. Int. Genet.* 24, 18–23. doi:10.1016/j.fsigen.2016.05.008

Xavier, C., and Parson, W. (2017). Evaluation of the Illumina ForenSeq DNA Signature Prep Kit - MPS Forensic Application for the MiSeq FGx Benchtop Sequencer. *Forensic Sci. Int. Genet.* 28, 188–194. doi:10.1016/j.fsigen.2017.02.018

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

Xie, T., Guo, Y., Chen, L., Fang, Y., Tai, Y., Zhou, Y., et al. (2018). A Set of Autosomal Multiple InDel Markers for Forensic Application and Population Genetic Analysis in the Chinese Xinjiang Hui Group. *Forensic Sci. Int. Genet.* 35, 1–8. doi:10.1016/j.fsigen.2018.03.007

Xie, T., Shen, C., Jin, X., Lan, Q., Fang, Y., and Zhu, B. (2020). Genetic Structural Differentiation Analyses of Intercontinental Populations and Ancestry Inference of the Chinese Hui Group Based on a Novel Developed Autosomal AIM-InDel Genotyping System. *Biomed. Research International* 2020, 2124370. doi:10.1155/2020/2124370

Xie, X., and Shan, X. (2002). *Research on the Hui*.

Xu, S., Huang, W., Qian, J., and Jin, L. (2008). Analysis of Genomic Admixture in Uyghur and its Implication in Mapping Strategy. *Am. J. Hum. Genet.* 82, 883–894. doi:10.1016/j.ajhg.2008.01.017

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

Yousefi, S., Abbassi-Daloii, T., Kraaijenbrink, T., Vermaat, M., Mei, H., van 't Hof, P., et al. (2018). A SNP Panel for Identification of DNA and RNA Specimens. *BMC Genomics* 19, 90. 10.1186/s12864-018-4482-7.

Zeng, X., Chakraborty, R., King, J. L., Larue, B., Moura-Neto, R. S., and Budowle, B. (2016). Selection of Highly Informative SNP Markers for Population Affiliation of Major US Populations. *Int. J. Leg. Med.* 130, 341–352. doi:10.1007/s00414-015-1297-9

Zhou, B., Wen, S., Sun, H., Zhang, H., and Shi, R. (2020). Genetic Affinity between Ningxia Hui and Eastern Asian Populations Revealed by a Set of InDel Loci. *R. Soc. Open Sci.* 7, 190358. doi:10.1098/rsos.190358

Zhu, B.-F., Zhang, Y.-D., Liu, W.-J., Meng, H.-T., Yuan, G.-L., Lv, Z., et al. (2014). Genetic Diversity and Haplotype Structure of 24 Y-Chromosomal STR in Chinese Hui Ethnic Group and its Genetic Relationships with Other Populations. *Electrophoresis* 35, 1993–2000. doi:10.1002/elps.201300574

Zou, X., Wang, Z., He, G., Wang, M., Liu, J., Wang, S., et al. (2020). Genetic Variation and Population Structure Analysis of Chinese Wuzhong Hui Population Using 30 Indels. *Ann. Hum. Biol.* 47, 300–303. doi:10.1080/03014460.2020.1736627

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 (GH) declared a past co-authorship with one of the authors (RT) to the handling editor.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 December 2021
doi: 10.3389/fgene.2021.795570

The Genetic Structure and East-West Population Admixture in Northwest China Inferred From Genome-Wide Array Genotyping

Bin Ma[1†], Jinwen Chen[2†], Xiaomin Yang[3], Jingya Bai[1], Siwei Ouyang[1], Xiaodan Mo[1], Wangsheng Chen[1], Chuan-Chao Wang[2,3,4] and Xiangjun Hai[1]**

[1]Key Laboratory of Environmental Ecology and Population Health in Northwest Minority Areas, Northwest Minzu University, Lanzhou, 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

Northwest China is a contacting region for East and West Eurasia and an important center for investigating the migration and admixture history of human populations. However, the comprehensive genetic structure and admixture history of the Altaic speaking populations and Hui group in Northwest China were still not fully characterized due to insufficient sampling and the lack of genome-wide data. Thus, We genotyped genome-wide SNPs for 140 individuals from five Chinese Mongolic, Turkic speaking groups including Dongxiang, Bonan, Yugur, and Salar, as well as the Hui group. Analysis based on allele-sharing and haplotype-sharing were used to elucidate the population history of Northwest Chinese populations, including PCA, ADMIXTURE, pairwise Fst genetic distance, *f*-statistics, qpWave/qpAdm and ALDER, fineSTRUCTURE and GLOBETROTTER. We observed Dongxiang, Bonan, Yugur, Salar, and Hui people were admixed populations deriving ancestry from both East and West Eurasians, with the proportions of West Eurasian related contributions ranging from 9 to 15%. The genetic admixture was probably driven by male-biased migration- showing a higher frequency of West Eurasian related Y chromosomal lineages than that of mtDNA detected in Northwest China. ALDER-based admixture and haplotype-based GLOBETROTTER showed this observed West Eurasian admixture signal was introduced into East Eurasia approximately 700 ~1,000 years ago. Generally, our findings provided supporting evidence that the flourish transcontinental communication between East and West Eurasia played a vital role in the genetic formation of northwest Chinese populations.

Keywords: genetic structure, admixture history, gene flow, west Eurasia, steppe population, trans-Eurasia, gansu, northwest China

OPEN ACCESS

Edited by:
Horolma Pamjav,
Ministry of Interior, Hungary

Reviewed by:
Oscar Lao,
Center for Genomic Regulation (CRG), Spain
Jiang Huang,
Guizhou Medical University, China

****Correspondence:***
Chuan-Chao Wang
wang@xmu.edu.cn
Xiangjun Hai
yxhxj@xbmu.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: *15 October 2021*
Accepted: *06 December 2021*
Published: *21 December 2021*

Citation:
Ma B, Chen J, Yang X, Bai J, Ouyang S, Mo X, Chen W, Wang C-C and Hai X (2021) The Genetic Structure and East-West Population Admixture in Northwest China Inferred From Genome-Wide Array Genotyping. Front. Genet. 12:795570. doi: 10.3389/fgene.2021.795570

INTRODUCTION

The human history of East Asia can be traced back to the Late Paleolithic Age. The anatomically modern humans permanently made an occupation in East Asia about 50,000 years ago (Alexander et al., 2009). Numerous evidences from ancient and present-day human genomes suggested an initial settlement in East Asia about 60,000 years ago and multiple waves of population expansion in Paleolithic and Neolithic periods (Fu et al., 2013; Bai et al., 2020; Zhang et al., 2020). The Pan-Asia project suggested the main southern migration route contributed much more to the peopling of the East Asia compared to the northern migration route by analyzing genome-wide data of 1900 individuals from 73 populations (HUGO Pan-Asian SNP Consortium et al., 2009; Cao et al., 2020). However, paternal Y chromosome and maternal mitochondrial DNA indicated that the gene flows from the west and northern Eurasia into East Asia were through the northern migration route (Su et al., 1999; Wen et al., 2004). East Asia is an important earliest center of animal and plant domestication in the world (Wang et al., 2021a). Paleogenomic studies documented that the genetic diversity in prehistoric Asia was higher than in more recent periods of human history and population migration between northern and southern East Asia that started in Late Neolithic Age influenced the genetic formation of modern East Asiana (Ning et al., 2020; Yang et al., 2020; Wang et al., 2021a; Wang et al., 2021b; Xiaowei et al., 2021). These expansion events were associated with the spread of the major language families existing in East Asia. There is also a remarkable diversity of human languages spoken in East Asia, including Sino-Tibetan, Hmong-Mien, Austroasiatic, Tai-Kadai, Austronesian, Indo-European, Turkic, Mongolic, Tungusic, Japonic, Koreanic, Yukaghiric, and Chukotko-Kamchatkan (Wang et al., 2021a; Uesugi et al., 2021). The formation of East Asians is suggested to has involved genetic contributions from various ancestral human populations (Duan et al., 2018; Sun et al., 2019; Wang et al., 2021a).

The Eastern Steppe is characterized with grasslands, forest steppe, and desert steppe, connecting Russia, Mongolia, and China. The Eastern Eurasian Steppe is home to historic empires of nomadic pastoralists, including Xiongnu, Turkic Khaganate, and the Mongols. The East Steppe have also served as the important communication node between West and East Eurasia. The Central/East Steppe has witnessed intensive East and West communications and interactions in many aspects (Elfari et al., 2005; Hyten et al., 2010; Stoneking and Delfin, 2010; Liu et al., 2018; Chen et al., 2019; Lan et al., 2019; Cao et al., 2020; Tangkanchanapas et al., 2020; Rodin et al., 2021). Historical and archeological studies demonstrated that the western Eurasian cultural factors were once brought into the north region of China through the East-West communication corridors (Sanchez-Burks et al., 2003; Xu, 2008; Ning et al., 2019). In the past, the ancient Silk Road was an important connection of the West Eurasia and China, which contributed much to the intensified transcontinental culture and population communications s between the East and West Eurasia (Cheng, 1985; Robino et al., 2014). The Silk Road was at its most bustling time in Tang Dynasty, but before that time the east-west communication was established for a long time, which could be traced back to the Early Bronze Age (Haak et al., 2015; Goldberg et al., 2017; Lazaridis and Reich, 2017; Saag et al., 2017). The corresponding trans-continental population migration during the Late Neolithic Age, the Bronze Age to the Iron Age and historical period had been demonstrated in the core regions of Siberia (Abelson, 1978; Matsumoto et al., 1995; Hemphill and Mallory, 2004; Maramovich et al., 2008; Jeong et al., 2018; Juras et al., 2020; Stoof-Leichsenring et al., 2020). The archeological evidence supported the interaction between the westward spread of millet agriculture and also the eastward spread of barley and wheat agriculture with population migration (Zohary and Hopf, 1973; Medjugorac et al., 1994; Hemphill and Mallory, 2004; Saisho and Purugganan, 2007; Wang et al., 2016; De Barros Damgaard et al., 2018b; Bento et al., 2018; Jeong et al., 2018). The Trans-Eurasian cultural and genetic exchanges have significantly influenced the demographic dynamics of Eurasian populations (Peel and Talley, 1996; Khan et al., 2017; Miller et al., 2017; De Barros Damgaard et al., 2018a; De Barros Damgaard et al., 2018b; Damgaard et al., 2018; Antwerpen et al., 2019; Coulehan, 2020; Saint Onge and Brooks, 2020; Zhou et al., 2020). The EasternEurasian Forest steppe zone was genetically structured during the Pre-Bronze and Early Bronze Age, with a strong west-east admixture cline of ancestry stretching from Botai in central Kazakhstan to Lake Baikal in southern Siberia, and to the Devil's Gate Cave in the Russian Far East (Jeong et al., 2020). During the Bronze Age, the eastward migration of Western Eurasian nomadic populations related to Afanasievo and Andronovo Culture into Eastern Steppe have not only influenced the gene pool of eastern Eurasian populations (Ning et al., 2019; Wang et al., 2021a), but also drastically changed lifeways and subsistence on the Eastern Steppe. The milk consumption in Mongolia started prior to 2500 BCE by groups related to Afanasievo and Chemurchek culture (Jeong et al., 2018). Until the Iron Age, the pastoralists established the nomadic empire in Eastern Steppe. The Xiongnu empire was the first historically recorded nomadic empire in Eastern Steppe, which had a profound influence on the demographics and geopolitics of Eurasia by expanding into northern China, southern Siberia, and Central Asia, even as far as the West Eurasian (Damgaard et al., 2018). During 13th century, the Mongols group eventually controlled a vast territory and numerous trade routes stretching from China to the Mediterranean (Jeong et al., 2020). The archaeological evidence showed Mongolia Plateau is a conduit for cultural exchanges between the East and the West Eurasia (Malyarchuk et al., 2016; Wang et al., 2021a; Liu et al., 2021).

Northwest China locates in the west-east Eurasian interaction core region, populations in this region mainly belongs to Altaic language family which includes Mongolic, Turkic, and Tungusic language based on language classification. Modern populations in Northwest China were typical admixtures between populations all around the trans-Eurasia continent (Feng et al., 2017; Yao et al., 2021). Uyghur derived western related ancestry from West Eurasians and South Asians, while the eastern related components were from the East Asians, and the Siberians (Ma

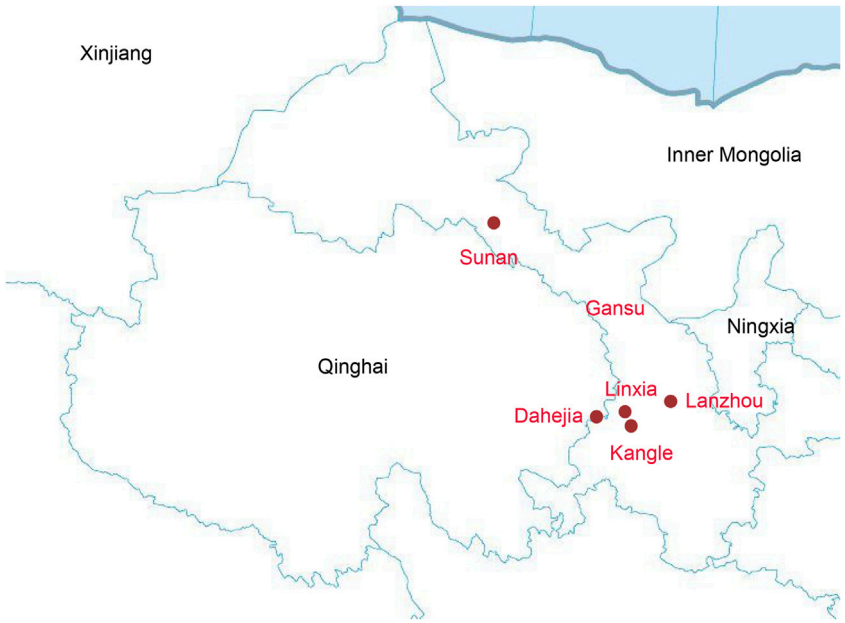


FIGURE 1 | The geographical map of our samples collection.

et al., 2014; Feng et al., 2017; Heizhati et al., 2020). Gansu province connecting the Hexi Corridor and the Tibetan-Yi Corridor in northwest China is not only takes part in the west-east Eurasian communication, but also plays an important role in the southwards population expansion which contributed to the formation of Tibeto-Burman speaking population (Feng et al., 2020; Luo et al., 2020). Human population genetic researches had been carried out based on low-density genetic markers and limited sample sizes to explore the genetic history of Gansu province (Yao et al., 2016; Yao et al., 2017; Wen et al., 2019). But a comprehensive survey of the genetic diversity and fine-scale genetic structure of Gansu province based on genome-wide data were still sparse. Therefore, to shed more light on the genetic profile of northwest China, 140 individuals from Gansu including Hui, Dongxiang, Bonan, Yugur, and Salar ethnic groups were collected and genotyping with Illumina gene arrays at approximately 700,000 genome-wide single-nucleotide polymorphisms (SNPs). We merged the genotyping data with reference data of worldwide populations, and carried out population genetics analysis to explore the genetic structure and uncovered the admixture history of Altaic speaking populations in Northwest China.

MATERIALS AND METHODS

Ethics Statement

The procedures of the sample collection and the investigations were reviewed and approved by the Medical Ethics Committee of Xiamen University and Northwest Minzu University and were in accordance with the recommendations provided by the revised Helsinki Declaration of 2000. Moreover, our study stuff had already informed these potential participants about our purposes of this project, and every participant in our study had provided the informed consent.

Sample Collection

Our study focused on Gansu province in Northwest China. We collected 140 saliva samples from unrelated individuals of Altaic speaking populations and Hui group from Sunan, Linxia, Lanzhou, Dahejia, and Kangle, including 24 samples from Hui, 30 samples from Dongxiang, 30 samples from Bonan, 30 samples from Yugur, and 26 samples from Salar (**Figure 1**). All inclued individuals were required to be indigenous self-declared, following the critera that requiring an indigenous person with at least three generations of history in the area and the offspring of a non-consanguineous marriage within populations.s.

Genotyping and Data Mergeing

We used PureLink Genomic DNA Mini Kit (Thermo Fisher Scientific) to extract DNA and measure the concentration via the Nanodrop-2000 following the manufacturer's instructions. All these qualified samples were genotyped using the Illumina WeGene Arrays covering about 700,000 Single nucleotide polymorphisms (SNPs) at the WeGene genotyping centre in Shenzhen. We first analyzed the biological relatedness of individuals using plink (Chang et al., 2015) softwere and all individuals were filtered. We also conducted the quality control process. There were 25,653 SNPs which were removed due to high percentage of missingness with "--geno 0.1 –mind 0.1" option using plink. Then we applied a HWE threshold by

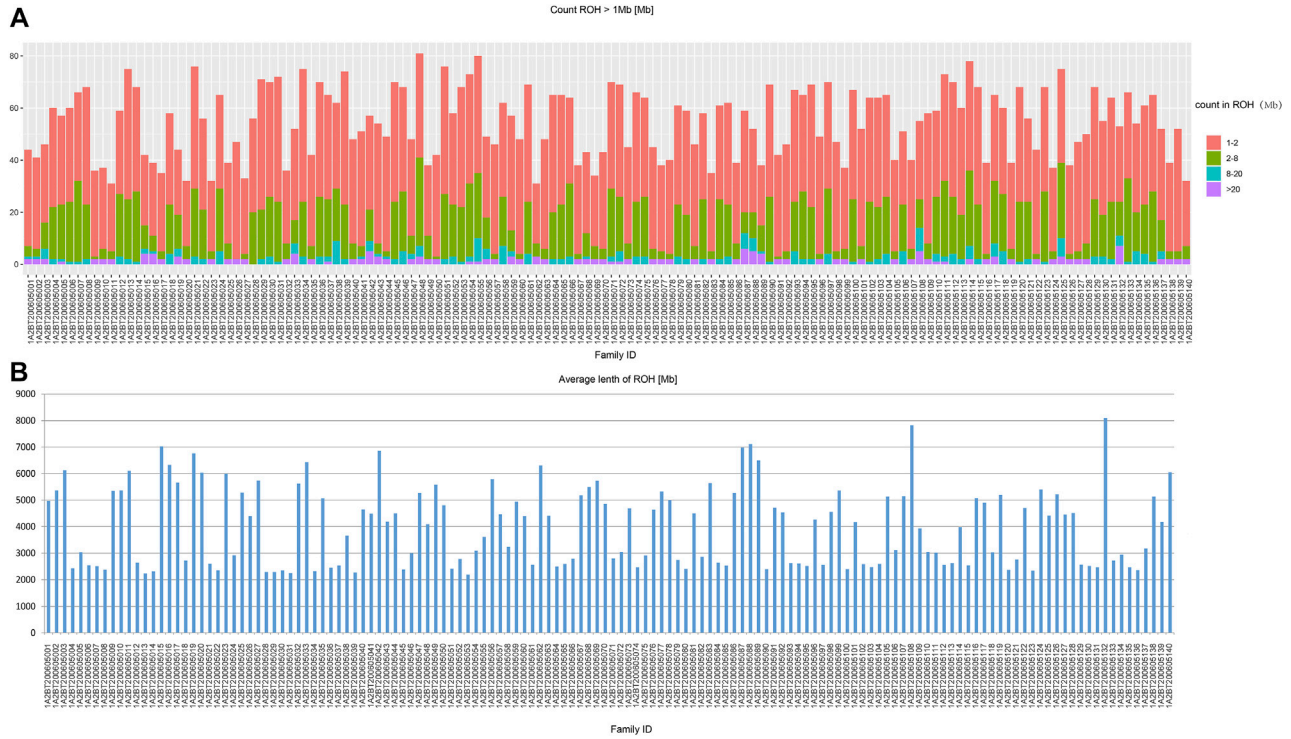


FIGURE 2 | The results of ROH calculation. **(A)** the length distribution; **(B)** the average length.

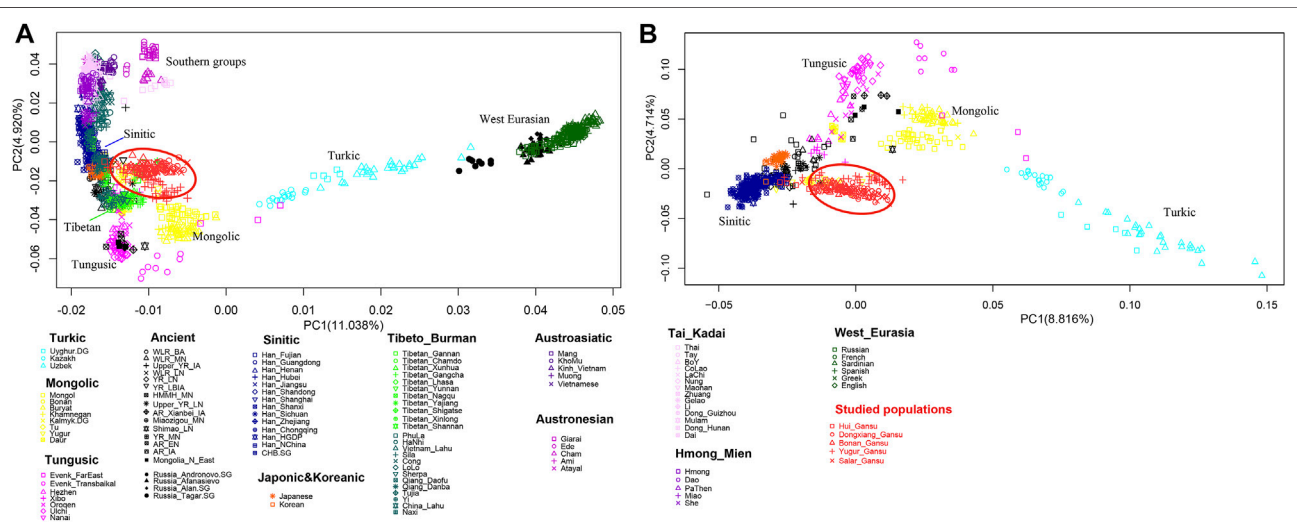


FIGURE 3 | Patterns of genetic relationship among published East Asian populations and our newly genotyped five populations inferred from the principal component analysis. **(A)** East Asians including southern populations and with the West Eurasians; **(B)** East Asians without southern populations and without the West Eurasians.

0.001, and 17,153 SNPs were removed. We pruned the Linkage Disequilibrium by "--indep-pairwise 200 25 0.4" for ADMIXTURE analysis. We obtained a dataset covring 72,541 SNPs when merged our 140 samples with the previously published data from Human Origin Datasetand a dataset covering merged 95,675 SNPs when merged with 1240 K capture dataset from David Reich Lab (https://reich.hms.harvard.edu/downloadablegenotypes-present-day-and-ancient-dna-data-compiled-published-papers) (Patterson et al., 2006; 2012).

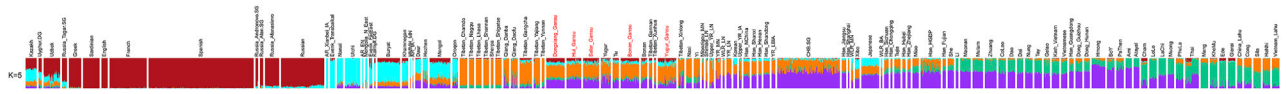


FIGURE 4 | ADMIXTURE analysis result visualization at K = 5 as the corresponding cross-validation error was the lowest. And our studied populations in Gansu were marked by red color.

Principal Component Analysis

Principal component analysis (PCA) was carried out using the software called *smartpca* built in the EIGENSOFT package (Patterson et al., 2006). The PCA analysis was performed at the individual level to describe the genetic structure of all of our samples in Gansu province and the reference populations. We used the following parameters: the numoutlieriter: 0 and lsqproject: YES options. We projected ancient individuals onto the first two components calculated by present-day samples. We visualized the PCA results by the ggplot2 package in the R software (http://www.r-project.org/).

ADMIXTURE

We carried out ADMIXTURE (Alexander et al., 2009) analysis after pruning for strong linkage disequilibrium in Plink V.1.9 (Purcell et al., 2007; Chang et al., 2015) with the parameters "-indep-pairwise 200 25 0.4". We ran ADMIXTURE with the 10-fold cross-validation (–CV = 10), varying the number of ancestral populations between K = 2 and K = 20 in 100 bootstraps with different random seeds. We chose the best run according to the highest log-likelihood with the lowest CV error value.

F-Statistics

We computed *f* statistics using ADMIXTOOLS with the default parameters, and calculated standard errors (statistical significance) using a block jackknife resampling across the genome (Patterson et al., 2006). We carried out outgroup f_3-statistics of the form f_3 (X, Y; Mbuti) to measure the shared genetic drifts between population X and population Y since their separation from an outgroup population. We here used Mbuti as an outgroup population, a group who lived in the Congo basin in the middle region of Africa. We next used admixture-f_3 statistics in the form of f_3 (X, Y; Target) for all pairs of references populations to make an evaluation of the possible admixture signals for the target populations. We conducted the heatmap visualization of the outgroup-f_3 statistics values by the pheatmap package in the R software.

Streams of Ancestry and the Inference of Admixture Proportions

We investigated the admixture source numbers, plausible admixture sources, and the corresponding admixture proportions based on *qpAdm* program as implemented in ADMIXTOOLS (Patterson et al., 2006). We used this f_4-statistics based admixture modeling to explore whether a batch of target populations were consistent with being related via N streams of source populations from a basic set of some outgroups and calculated the admixture proportions of the given source populations quantitatively.

Y-Chromosomal and mtDNA Haplogroup Assignment

We assigned the Y chromosomal haplogroups by genotyping the most derived allele upstream and the most ancestral allele downstream in the phylogenetic tree by using an in-house script following the recommendations of the International Society of Genetic Genealogy (ISOGG; http://www.isogg.org/). The mtDNA haplogroups assignment was identified with mtDNA phylogenetic tree Build 16 (http://www.phylotree.org/).

Fst Calculation

The Fst values were calculated by the *smartpca* of EIGENSOFT (Patterson et al., 2006). We ran the *smartpca* with the parameters: inbreed: YES and fstonly: YES, and then output the results by phylipoutname parameter. We found that the inbreeding corrected and uncorrected Fst were nearly identical. In the following, we performed the phylogenetic tree by the Fst values of the populations in Eurasia. We performed the phylogenetic tree by the NJ tree using MEGA software (Kumar et al., 2016).

Weighted Linkage Disequilibrium Analysis

Linkage disequilibrium decay was computed by ALDER (Loh et al., 2013) to infer the admixture time for our studied populations.

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 selected 10–20 individuals randomly in a reference group and 15 individuals in our studied group. We phased genome-wide dense SNP data using the SHAPEIT2 (Delaneau et al., 2013) and then conducted FineSTRUCTURE (Lawson et al., 2012) analysis.

ChromoPaintev2 and GLOBETROTTER Admixture Modeling

We performed a GLOBETROTTER (Hellenthal et al., 2014) analysis for our studied groups to obtain haplotype-sharing

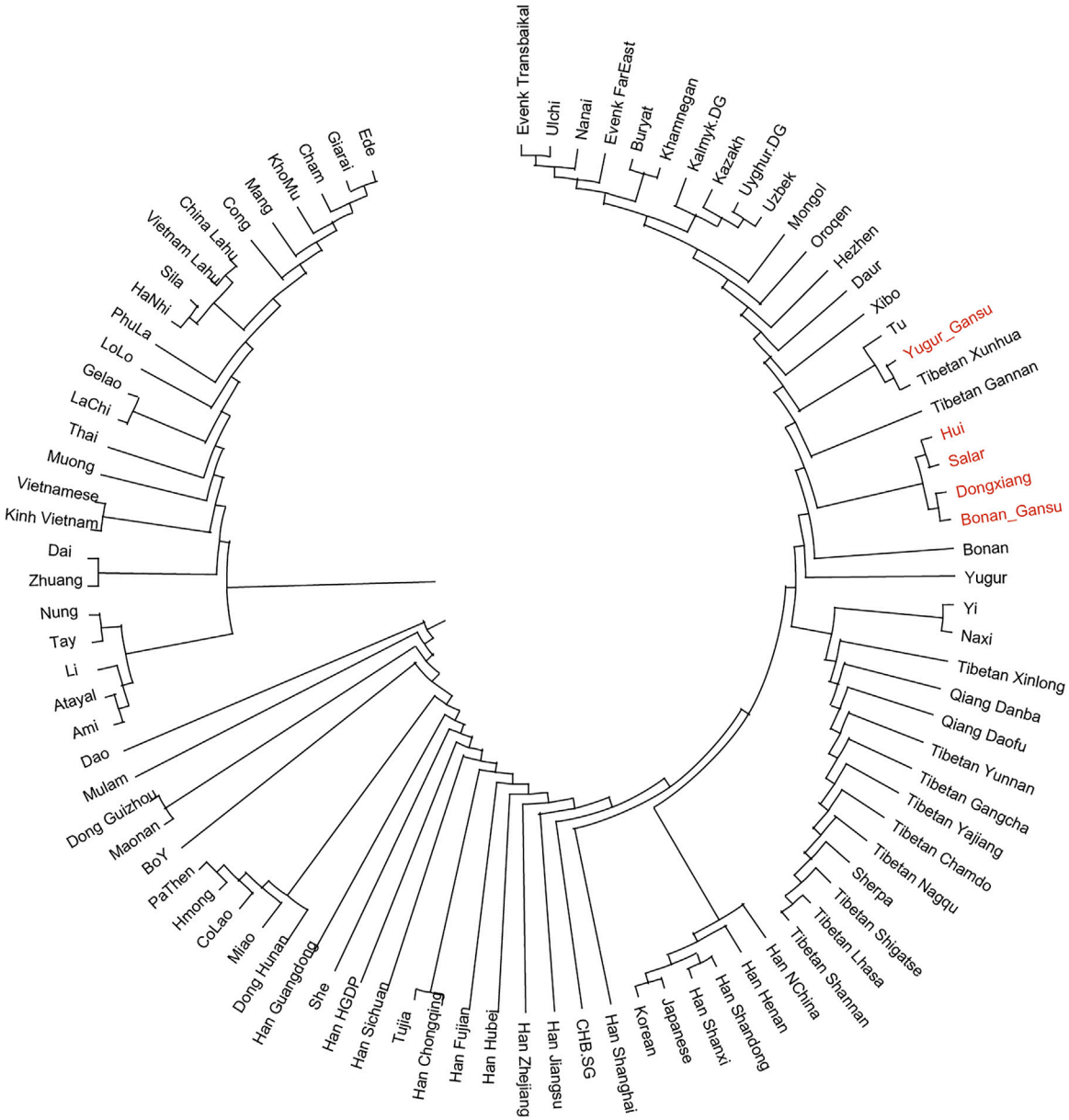


FIGURE 5 | Phylogenetic tree among our studied populations in Gansu and reference populations in Eurasia. Our samples in Gansu province were marked with red color.

based evidence of admixture. Using these haplotypes from SHAPEIT2, the "chunk length" output was obtained by running ChromoPainterv2 across all chromosomes. We ran GLOBETROTTER to estimate admixture events by 100 bootstrap replicates, assuming that there is detectable admixture using the "pro.ind:1", and "bootstrap.date.ind:1" options.

Runs of Homozygosity

We calculated the Runs of homozygosity by PLINK software. The related parameters were: "--homozyg-density 50, --homozyg-window-het 1, --homozyg-window-threshold 0.05". Then we presented the counts and lengths of ROH.

RESULTS

Population Genetic Structure of the Northwest China

In the beginning of the population genetic analysis, we presented the results of ROH computation (**Figure 2**). In our studied populations in Northwest China, the ROH segments were mainly short fractions which were between 1 and 2 Mb.

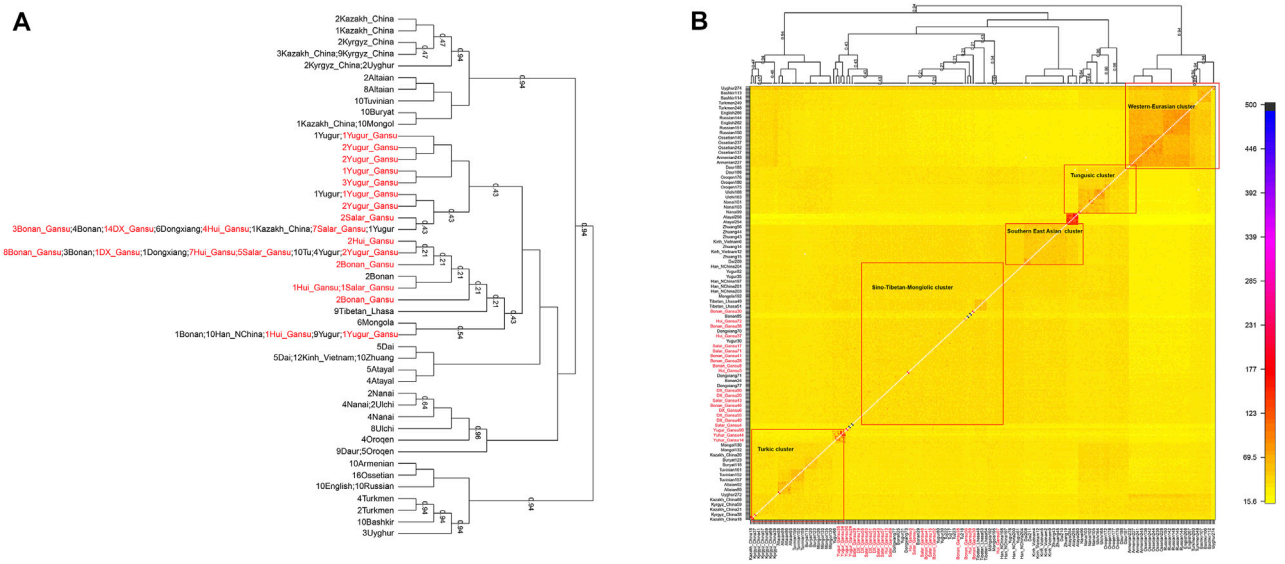


FIGURE 6 | The heat map of sharing haplotypes and clustering dendrogram by fineSTRUCTURE. DX = Dongxiang. **(A)** the dendrogram. **(B)** the heat map of sharing haplotypes.

And the long fractions which were longer than 20 Mb were rare. Therefore, our studied populations were not consanguineous communities.

We firstly conducted PCA to infer the general genetic structure of our sampled populations with other East Asians (**Figure 3**). From the PCA plot, we found the genetic clusters were consistent with the geographic, and linguistic categories in East Asia. We observed the following clear genetic clusters or clines. A genetic cline related to Turkic speaking populations, which was driven by populations with a large amount of West Eurasian related ancestry, such as Uyghur and Uzbek ethnic groups; a cluster with the Mongolic speaking populations; a cluster related to Tungusic speaking populations; a cluster of populations in West Eurasia. A cluster of Tibetan populations on the high-altitude region; a cluster with Han Chinese groups; and a huge cluster related to southern populations in East Asia speaking Hmong-Mien, Austroasiatic, Tai-Kadai, and Austronesian languages. Our newly reported samples in Gansu province clustered genetically between the Han Chinese groups and the Turkic speaking populations. We next removed the populations from southern China and Southeast Asia and the human groups in West Eurasia to show a more clearly clustering pattern among northern populations. In the zoomed PCA, our newly reported populations were close to the Han Chinese cluster, but also shifted towards the Turkic genetic cline, showing genetic affinity with both Turkic populations, and Han Chinese.

We next carried out the model-based ADMIXTURE clustering analysis. We observed the lowest CV error at K = 5. We then made the visualization of the result at K = 5 with five colors (**Figure 4**): The red component was primarily enriched in West Eurasians; the blue component was largely shown in the Mongolic and Tungusic speaking populations; the orange component was mainly detected in the Tibetan groups; the green component was largely presented in Austronesian speaking populations; the purple component was mainly enriched in some southern groups in East Asia. Our newly reported Hui, Dongxiang, Bonan, Yugur, and Salar samples harbored large orange and purple ancestral component related to East Asia and a part of red ancestral component related to the West Eurasia. The ancestry assignment was consistent with previous PCA analysis.

In the following, we calculated the pairwise Fst values for our studied populations in Gansu province together with reference populations in Eurasia and constructed a phylogenetic tree (**Figure 5**). In this phylogenetic tree, our newly reported groups in Gansu province clustered closely with the surrounding Altaic speaking populations in northern China. Notablely, The Yugur group clustered together with Tibetans from Xunhua and Gannan and Tu.

Next, we characterized the finer-scale population structure of our studied groups in Gansu by the haplotype-based fineSTRUCTURE. The inferred polygenetic tree based on the linked coancesty matrix showed all populations were clustered well according to geographical positions and language classification. Overall, our studied population clustered with published Mongolic speakers and Turkic speakers Kazakh in China, forming the major branch that also included Han, Tibetan, and Mongolia of China. Our Yugur_Gansu population formed relatively sporadic and formed serval small branches, even one individual clustered with published Yugur (**Figure 6A**). In addition, Hui people clustered with Bonan, Dongxiang, Salar, Yugur groups. Heatmap (**Figure 6B**) and the corresponding clustering

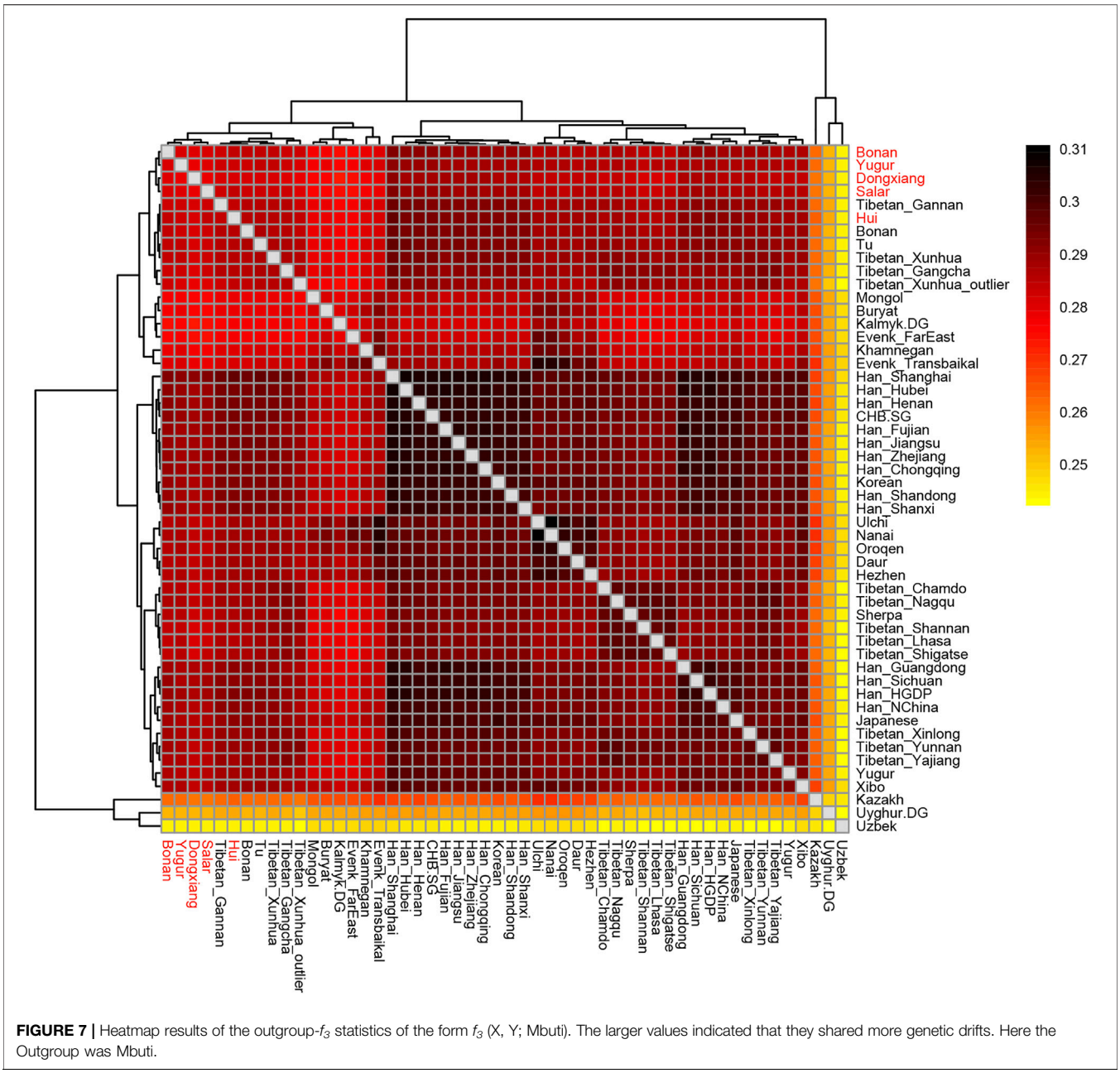


FIGURE 7 | Heatmap results of the outgroup-f_3 statistics of the form f_3 (X, Y; Mbuti). The larger values indicated that they shared more genetic drifts. Here the Outgroup was Mbuti.

patterns showed five major clusters, the Sino-Tibetan-Mongolic cluster included Chinese Mongolic populations in northwestern China, Tibetan and Han populations, our studied populations the larger amount of haplotype sharing among those populations.

Continuity and Admixture of Populations by the Allele-Shared *f*- Statistics

In the following, we calculated the outgroup-f_3 statistics in the form of f_3 (*X*, *Y*; *Mbuti*) to quantify the population differentiation across East Asia. We showed the results in a heatmap plot (**Figure 7**). The larger value of the statistics indicated that the two groups shared more genetic drifts after the separation from an African outgroup. We found the majority of Han Chinese populations shared more alleles with each other and clustered together. The Mongolic and Tungusic populations (Ulchi, Nanai, Oroqen, Daur, Hezhen) also clustered together. Our studied populations Hui, Dongxiang, Bonan, Yugur, and Salar clustered together and shared more genetic drifts with Han Chinese populations than with Tibetan groups.

In addition, we performed the admixture-f_3 statistics in the form of f_3 (Source1, Source2; Target) to explore the possible ancestral source populations for our studied populations in Gansu province. We observed the most significant negative signals when using the Neolithic Yellow River farming groups and the Bronze Age to Iron Age Steppe groups from West Eurasia

TABLE 1 | Admixture *f3* statistics of the form (Source1, Source2; Target) with the lowest *f3* values.

Source 1	Source 2	Target	f_3	Std. err	Z	SNPs
Kazakhstan_Andronovo.SG	Upper_YR_LN	Hui	−0.010305	0.001639	−6.288	55271
Kazakhstan_Andronovo.SG	Upper_YR_IA	Hui	−0.009701	0.001942	−4.995	53232
Kazakhstan_Kangju.SG	Shimao_LN	Hui	−0.009247	0.001029	−8.99	161279
Russia_Alan.SG	WLR_LN	Hui	−0.009221	0.001209	−7.63	136872
Russia_Alan.SG	YR_LBIA	Hui	−0.009003	−0.000714	−12.609	165864
Russia_Alan.SG	WLR_LN	Dongxiang	−0.014067	−0.001139	−12.348	138809
CHB.SG	Anatolia_N	Dongxiang	−0.013755	−0.000316	−43.553	172663
CHB.SG	Russia_MLBA_Sintashta	Dongxiang	−0.013573	0.000324	−41.928	171061
Anatolia_N	YR_LBIA	Dongxiang	−0.013525	−0.000628	−21.527	170054
Kazakhstan_Kangju.SG	Shimao_LN	Dongxiang	−0.013429	−0.00097	−13.847	163617
Russia_MLBA_Sintashta	Miaozigou_MN	Bonan	−0.010536	0.001382	−7.621	49714
Kazakhstan_Andronovo.SG	Upper_YR_LN	Bonan	−0.010432	0.001621	−6.436	55987
Russia_Alan.SG	WLR_LN	Bonan	−0.010298	0.001191	−8.645	138290
Kazakhstan_Kangju.SG	Shimao_LN	Bonan	−0.010185	−0.000996	−10.226	163067
CHB.SG	Anatolia_N	Bonan	−0.010179	−0.000311	−32.782	172482
Kazakhstan_Andronovo.SG	Upper_YR_LN	Yugur	−0.008896	0.001651	−5.388	55670
Russia_Alan.SG	Wuzhuangguoliang	Yugur	−0.008814	0.002125	−4.148	28513
Russia_MLBA_Sintashta	Miaozigou_MN	Yugur	−0.008676	0.001367	−6.346	49569
Kazakhstan_Kangju.SG	Shimao_LN	Yugur	−0.00843	0.001004	−8.393	162332
Kazakhstan_Andronovo.SG	Upper_YR_IA	Yugur	−0.008279	0.001914	−4.326	53651
Kazakhstan_Andronovo.SG	Upper_YR_LN	Salar	−0.011922	0.001611	−7.4	55593
Kazakhstan_Andronovo.SG	Upper_YR_IA	Salar	−0.011297	0.001916	−5.897	53574
Kazakhstan_Kangju.SG	Shimao_LN	Salar	−0.011024	−0.001022	−10.781	162165
Russia_Alan.SG	WLR_LN	Salar	−0.010966	0.001182	−9.28	137507
Russia_Alan.SG	Shimao_LN	Salar	−0.010849	−0.000996	−10.893	165726

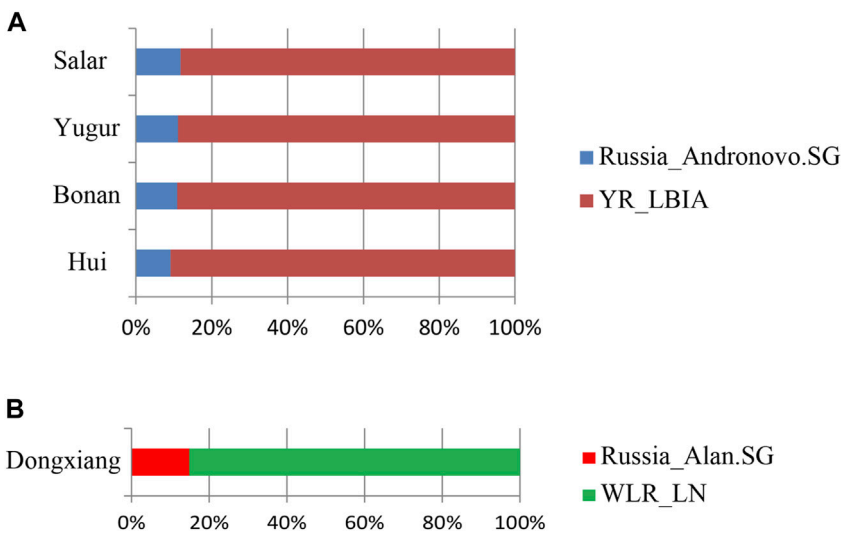


FIGURE 8 | qpAdm based admixture models for the populations in our study in Gansu province. The 2-way admixture models for our Gansu samples were presented when the *p* values >0.05 at the rank = 1. **(A)** Hui, Bonan, Yugur, Salar ethnic groups. **(B)** Dongxiang ethnic group.

and Central Asia as sources (**Table 1**), suggesting the gene flow from West Eurasia into northwest China.

The Ancestry Inference of the Populations in Northwest China

We next carried out *qpAdm* analysis to infer the admixture proportions in our studied Gansu populations (**Figure 8**; **Table 2**). The eastern ancestral source populations we selected were the Yellow River farming groups from the Bronze Age to Iron Age, and the western ancestral source populations we selected were ancient populations of Andronovo and Alan cultures, since theyprovided the most significant negative admixture-f_3 values. We used the following set of populations as outgroups: Mbuti, Russia_EBA_Yamnaya_Samara, Anatolia_N, Russia_MA1, Russia_Afanasievo, Mongolia_N_East,

TABLE 2 | Two-way *qpAdm* models of studied populations in Gansu.

Studied population	Proportion	Std. err	Proportion	Std. err	*p* value
	Russia_Andronovo.SG		YR_LBIA		
Hui	0.091	0.007	0.909	0.007	0.206
Bonan	0.109	0.007	0.891	0.007	0.0536
Yugur	0.111	0.007	0.889	0.007	0.546
Salar	0.118	0.007	0.882	0.007	0.098
	Russia_Alan.SG			WLR_LN	
Dongxiang	0.149	0.011	0.851	0.011	0.304

TABLE 3 | The Y-chromosome haplogroups distribution of our studied populations.

	Y Haplogroup	Frequency
Hui	D1a1a1a1a2a~	0.100
	H1a1a1a	0.100
	J2a1a	0.100
	J2a1h2b	0.100
	J2a2	0.100
	N1a2b3	0.100
	O2a2a1a2a1a	0.100
	O2a2b1a1a6b	0.100
	R1a1a1b2	0.200
Dongxiang	D1a1a1a2	0.133
	E1b1a1a1a2a1a3b1a10b~	0.067
	J2a1h2	0.200
	J2a2	0.067
	L1a2a1b2~	0.067
	N1a1a1a1a3a2a~	0.067
	N1a3~	0.067
	O2a2b1a1a6	0.133
	R1a1a1b2	0.067
	R2a2	0.133
Bonan	C2b1a2a2a~	0.0625
	D1a1a1a1a2a~	0.125
	D1a2a1~	0.0625
	J2a1h2	0.0625
	N1a2b3a~	0.0625
	O1b1a1a1a1b1b	0.0625
	O1b1a1a1a2	0.0625
	O2a2b1a1a	0.125
	O2a2b1a2a1d	0.0625
	Q1b1a3b1a1~	0.0625
	Q2a1c1b1~	0.0625
	R1a1a1b2	0.1875
Yugur	C2b1a1	0.133
	C2b1a3b~	0.067
	D1a1a1a1a2a~	0.133
	D1a1a1a2	0.067
	O2a2b1a1a6	0.067
	O2a2b1a2a1a2	0.133
	O2a2b1a2b2	0.133
	O2a2b2a2a1	0.067
	Q1b1a3a~	0.067
	Q1b2b1b2b2~	0.133
Salar	I2a2a1b2a1b1b2a2~	0.091
	J2a1	0.091
	N1b2a2~	0.091
	O1b1a1a1b2	0.182
	O2a1a1b1a2	0.091
	O2a1c1a1a1a1a1b1a~	0.091
	O2a2b1a2a1a1a1	0.091
	R1a1a1b2	0.273

Ust_Ishim, Russia_Kostenki14, Iran_C_SehGabi. Our studied populations could be modeled by two-way admixture with the p-value > 0.05 at rank = 1. We estimated the genetic proportions of Russia_Andronovo related ancestry were 9.1 ~ 11.8%, while the genetic proportions of YR_LBIA farming group related ancestry were 88.2 ~ 90.9% in Hui, Bonan, Yugur, and Salar groups. Given the pair groups consisting of Late Neolithic farmers in West Liao River (WLR_LN) and Iron Age Alan people in Russia (Russia_Alan) in admixture f_3 showed the most significant admixture signal, we found that the Dongxiang group derived 14.9% western Eurasian ancestry from Russia_Alan related groups and the left from WLR_LN related groups. In general, the qpAdm model indicated the west-east admixture in our five studied populations, showing East Asian related ancestry dominantly made contribution to the genetic formation of Northwest Chinses Altaic speaking groups with different proportions of West Eurasian related ancestry.

Y Chromosomal and MtDNA Haplogroup Assignment

We assigned the haplogroups of Y chromosome and mtDNA for our newly genotyped samples (**Table 3**). The haplogroup R1a1a1b2 was the most frequent patrilineal lineage in the Hui, Bonan, and Salar groups. We also detected haplogroup D1a1a1a1a2a~, H1a1a1a, J2a1a, J2a1h2b, J2a2, N1a2b3, O2a2a1a2a1a, and O2a2b1a1a6b in our Hui samples. Haplogroup D1a1a1a1a2a~ and O2a2b1a1a were also found in Bonan group. The haplogroup O1b1a1a1b2 was also presented in Salar group. Haplogroup J2a1h2, which was mostly found in the Middle East, was the most prevailing lineage in Dongxiang people. We also found D1a1a1a2, O2a2b1a1a6, and R2a2 in the Dongxiang group. Haplogroups C2b1a1, D1a1a1a1a2a~, O2a2b1a2a1a2, O2a2b1a2b2, and Q1b2b1b2b2~ were the prevalent lineages in the studied Yugur group. The distribution of Y haplotype indicated the influence of westward expansion of several ancestral sources in genetic formation of Northwest Chinese Altaic populations, including West Eurasian, Sino-Tibetan, common ancestor of Altaic related ancestry.

We next assigned the matrilineal mtDNA haplogroups for our studied populations. In the Hui group, we observed diverse mtDNA haplogroups, including D4, D5a2a1, F1, G3a1′2, M7, M8, Z3, and Z4. The maternal profile of Dongxiang group was

TABLE 4 | The mtDNA haplogroups distribution for our studied populations.

Hui		Dongxiang		Bonan		Yugur		Salar	
Haplogroup	**Frequency**	**Haplogroup**	**Frequency**	**Haplogroup**	**Frequency**	**Haplogroup**	**Frequency**	**Haplogroup**	**Frequency**
A16	0.041666667	A	0.06666667	A	0.03333333	A1	0.1	A	0.115385
B4c1b2c	0.041666667	A1	0.1	B4	0.06666667	A6b	0.033333333	A18	0.038462
B5b2	0.041666667	A6b	0.03333333	B5	0.06666667	B4a3	0.033333333	A5b1b	0.038462
C4d	0.041666667	B4	0.1	B6a	0.03333333	C4	0.1	A8a	0.038462
D4	0.083333333	C5d2	0.03333333	C4	0.1	D4	0.4	B4b1a2a	0.038462
D5	0.041666667	D4	0.1	C5b1b	0.03333333	D5a2a1	0.033333333	C4d	0.038462
D5a2a1a1	0.04166667	D5	0.06666667	D4	0.2	F1g	0.1	D4	0.076923
F1	0.125	F1	0.06666667	F1g	0.03333333	M9a1a1c1b1a	0.066666667	F1	0.115385
F4a2	0.041666667	F2	0.1	G2	0.1	M9a1b1	0.033333333	F3a1	0.038462
G3a1′2	0.083333333	F4b	0.03333333	H	0.06666667	R9b1a3	0.033333333	G1a1	0.038462
M7	0.125	H15	0.03333333	M10a1a1b	0.03333333	U4b1a1a1	0.033333333	G2a	0.076923
M8	0.083333333	H5	0.06666667	M7b1a	0.03333333	U7a	0.033333333	H7b1	0.038462
N9a2	0.04166667	M7b1a1a3	0.03333333	M8	0.06666667			M11a2	0.038462
Z3	0.083333333	M8	0.06666667	M9a1a1c1a	0.03333333			M21b	0.038462
Z4	0.083333333	T2a1a	0.03333333	X2b4	0.03333333			M9a1b1	0.115385
		X2	0.03333333	Z3a	0.06666667			Z3	0.115385
		Y1b1a	0.03333333						

TABLE 5 | The admixture time estimation by ALDER for our studied populations.

Population	1-Ref weighted LD with weights Sardinian (generation)	Z-score	1-Ref weighted LD with weights Han_HGDP (generation)	Z-score	2-Ref weighted LD with weights Sardinian and Han_HGDP (generation)	Z-score
Hui	34.98 ± 4.20	8.32	97.35 ± 35.92	2.71	31.36 ± 3.27	9.58
Dongxiang	28.71 ± 2.60	11.03	40.77 ± 7.40	5.51	26.73 ± 2.61	10.24
Bonan	33.21 ± 2.42	13.72	-	-	26.08 ± 2.50	10.42
Yugur	33.65 ± 4.53	7.42	-	-	25.32 ± 3.81	6.65
Salar	25.70 ± 3.64	7.07	33.74 ± 11.91	2.83	24.77 ± 3.83	6.47

similar to that in the Hui group, but the haplogroup A, B4, and F2 were more prevalent in Dongxiang. We found D4 was the most dominant lineage in Bonan group and we also detected B and G2a in Bonan group. Haplogroup D4 was also the most dominant haplogroup in Yugur group, following by A1, C4, F1g, and M9a1 haplogroups. Haplogroup A was the most prevailing haplogroup in the Salar group, following by F1, M9a1b1, and Z3. The main mtDNA haplogroups in our samples were also prevalent in East Asia, suggesting the local East Asians largely contributed to the maternal gene pool of Gansu Altaic speaking populations. The genetic influence from the West Eurasian human populations were more significant in the patrilineal lineages than in the matrilineal lineages. The details of the distribution of mtDNA haplogroups were listed in **Table 4**.

The Admixture Time Estimation for the Populations in Northwest China

We estimated the admixture time between the East and West Euraisan related ancestry in Northwest Chinese populaton using the weighted linkage disequilibrium-based admixture inference implemented in ALDER (Loh et al., 2013). We used Han_HGDP and Sardinian as two ancestral surrogates to calculated the east-west admixture time and listed the results in **Table 5**. The average admixture time calculated by the 2-ref weighted LD for our five studied populations ranged from 25 to 31 generations, which was approximately 750–930 years before present assuming 30 years one generation (**Table 5**). The east-west interactions were suggested to have occurred in about the Song and Yuan Dynasty of China.

We further performed haplotype-based GLOBETROTTER to obtain the admixture landscaped of our studied northwestern Chinese populations (**Table 6**). The east-west admixture could be traced back to ~21 to ~25 generations ago (approximately ~630–750 years ago assuming 30 years one generations), with inferring western Eurasian related ancestry represented by English ranging from 16 to 24%, coinciding with the results from ALDER. In addition, we observed the minor southern population admixture in Hui, Yugur and Salar (0.2, 0.06, and 0.04, respectively).

DISCUSSION

The East Asia is a region with diverse culture communications, multiple language interactions, and complex population history.

TABLE 6 | The admixture events of our studied populations by GLOBETROTTER.

Recipient. Population	Model	Gen.1date	Proportion. source1	Bestmatch. event1. source1	Bestmatch. event1. source2	Proportion. event2. source1	Bestmatch. event2. source1	Bestmatch. event2. source2	MaxR2fit. 1date	Fit.quality. 1event	Fit.quality. 2events
Hui_Gansu	1-DATE	24.88801703	0.18	English	Han_NChina	0.2	Kinh_Vietnam	Salar_Gansu	0.920102825	0.9996966	0.999996517
Dongxiang_Gansu	1-DATE	21.43093694	0.24	English	Han_NChina	0.41	Mongol	Dongxiang	0.943671053	0.999996881	0.999999612
Bonan_Gansu	multiple-dates	24.95247758	0.19	English	Han_NChina	0.49	Yugur	Bonan	0.918212959	0.999969001	0.9999979
Yugur_Gansu	1-DATE	23.47510082	0.16	English	Yugur	0.06	Atayal	Tibetan_Lhasa	0.884253347	0.999991289	0.999998974
Salar_Gansu	1-DATE	20.80295557	0.2	English	Han_NChina	0.04	Atayal	Hui_Gansu	0.943365942	0.99999999	1

Gen.2dates. date1	Gen.2dates. date2	Proportion. date1.source1	Bestmatch. date1.source1	Bestmatch. date1.source2	Proportion. date2.source1	Bestmatch. date2.source1	Bestmatch. date2.source2	MaxScore. 2events
1.000004327	24.44514919	0.42	Atayal	Salar_Gansu	0.18	English	Han_NChina	0.111809238
1.000023459	26.76153143	0.13	Turkmen	Tu	0.24	English	Han_NChina	0.154655956
8.270277733	30.57366368	0.34	Uyghur.DG	Bonan	0.18	English	Han_NChina	0.478201465
11.71119508	44.60320405	0.07	English	Tu	0.13	English	Yugur	0.073485128
10.74070225	43.5990788	0.07	English	Hui_Gansu	0.16	English	Han_NChina	0.176636782

Many previous studies provided that the genetic substructure of populations in East Asia was consistent with the language affinities. The Hexi Corridor and its surrounding regions were known for the famous Majiayao civilization in the middle and late Neolithic Age and subsequently controlled by the Rong-Di tribes before the Han Dynasty. Moreover, the Northwest China witnessed the intersection of the eastward expansion of the barley and wheat agriculture and the westward expansion of the millet agriculture in the Neolithic to Bronze Age. Gansu province isone of the key regions in Northwest China where also connects the Hexi Corridor and Tibetan-Yi Corridor. The genetic diversity, fine-scale genetic substructure, and the western Eurasian admixture in the populations of Gansu are still needed to be fully explored. We collected 140 modern individuals from Hui, Dongxiang, Bonan, Yugur, and Salar groups from the Gansu province and genotyped with genome-wide SNPs. We reconstructed the population admixture history of the Altaic speaking populations in northwest China.

Our studied populations of Northeast China showed similar genetic profile among those populations, suggesting the relatively genetic homogeneity in Northwest China, even though harboring subtle different proportions of East, and West Eurasian related ancestry. The close genetic affinity among Chinese Turkic speakers, Tungusic, and Mongolic populations indicated the probability of common ancestor of Altaic speakers. Our results showed that both West and East Eurasian contributed the genetic formation of Altaic populations in Northwest China, which coinciding with previous studies suggested the east-west admixture in Alatic populations and Hui population (Xu and Jin, 2008; Bai et al., 2018; Jeong et al., 2019; Zhao et al., 2020; Ma et al., 2021). The closer genetic relationship between our studied population and Sino-Tibetan populations and the results of qpAdm and GLOBETROTTER suggested the majority contributing East Eurasian ancestry might derived from millet farmers in Yellow River Basin related population. The eastward expansion of Bronze Age West Steppe nomadic groups limitedly impacted the gene pool of the East Eurasian. The five studied Altaic speaking groups were suggested to harbored the lower proportion of Middle and Late Bronze West Steppe pastoralists represented by Andronovo culture. This was also supported by the high frequencies of Y chromosomal haplogroup R1a1a1b2 which prevailed Middle and Late Bronze Age Steppe populations in Hui, Bonan, and Salar groups (Narasimhan et al., 2019). The genetic admixture from West Eurasians was probably driven by male dominant migration which showing the higher frequencies of West Eurasian related paternal Y chromosome lineages and the absence of maternal mtDNA lineage related to West Eurasian. The paleogenomic studies exhibited the most complex pattern of male-biased admixture in the demographic dynamics of the East Steppe (Jeong et al., 2020).

Considering that the West Eurasian related ancestry proportions were limited in our studied populations (<15%), we noted that it was hard to determine the exact genetic source for the admixture. The sequencing of more ancient genomes from Northwest China may shed more light on determining the West Eurasian sources. We estimated the admixture event to have occurred in historic period based on

ALDER and GLOBETROTTER (approximately dating to ~750–930 years ago, ~630–750 years ago, respectively). The ancient admixture we identified was roughly corresponding to the Song to Yuan Dynasty. But we noted if the admixture did not happen immediately after arrival or multiple times over an extended period, however, the true start of admixture would have been more ancient. Furthermore, the intensive and continuous contact between West and East Eurasian population started as early as the Bronze Age due to the advantage of horses, and the interaction became more frequent with the opening of Silk Road in the Han Dynasty. The establishment of Mongols empire and the Mongolian Conquests in the 13th and 14th centuries facilitated the west-east contacts. The true admixture history in Northwest China could be more complex than the simplified models as we presented in this study, the populations studied here, however, harbored prominent local East Eurasian related ancestry and limited West Eurasian related ancestry.

Running through the ancient Silk Road, the human groups were all presented a west-east admixture structure. The Uyghur in Xinjiang was a typical one. Besides, the Altaic speaking populations in Central Asia all have the west-east interactions in genetic structure and culture. The east endpoint of the ancient Silk Road was near Chang'an City, and the Gansu pathway was the only route to it. The Altaic populations in this region lack of large-scale sampling and genome-wide genetic analysis. Our research answered this issue at a certain degree, but the more elaborate admixture history needed to be explored from the whole genome sequencing next.

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/, https://doi.org/10.5281/zenodo.5542715.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Medical Ethics Committee of Xiamen University and Northwest Minzu University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

C-CW and XH designed the study. JC and C-CW wrote the article. BM, JB, SO, XM, WC, and XH collected the samples. BM, JB, SO, XM, WC, and XH conducted the experiment. JC, XY, and C-CW analyzed the data. All authors reviewed the article.

FUNDING

This work was funded by the Major Project of National Social Science Foundation of China (the origin and evolution of Sino-Tibetan language family from a multidisciplinary perspective) granted to C-CW (21 and ZD285), 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), National Natural Science Foundation of China (NSFC 31801040), 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).

ACKNOWLEDGMENTS

S. Fang and Z. Xu from Information and Network Center of Xiamen University are acknowledged for the help with the high-performance computing.

REFERENCES

Abelson, A. (1978). Population Structure in the Western Pyrenees: Social Class, Migration and the Frequency of Consanguineous Marriage, 1850 to 1910. *Ann. Hum. Biol.* 5, 165–178. doi:10.1080/03014467800002761

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

Antwerpen, M., Beyer, W., Bassy, O., Ortega-Garcia, M. V., Cabria-Ramos, J. C., Grass, G., et al. (2019). Phylogenetic Placement of Isolates within the Trans-eurasian Clade A.Br.008/009 of Bacillus Anthracis. *Microorganisms* 7, 689. doi:10.3390/microorganisms7120689

Bai, F., Zhang, X., Ji, X., Cao, P., Feng, X., Yang, R., et al. (2020). Paleolithic Genetic Link between Southern China and Mainland Southeast Asia Revealed by Ancient Mitochondrial Genomes. *J. Hum. Genet.* 65, 1125–1128. doi:10.1038/s10038-020-0796-9

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

Bento, C. B., Filoso, S., Pitombo, L. M., Cantarella, H., Rossetto, R., Martinelli, L. A., et al. (2018). Impacts of Sugarcane Agriculture Expansion over Low-Intensity Cattle Ranch Pasture in Brazil on Greenhouse Gases. *J. Environ. Manage.* 206, 980–988. doi:10.1016/j.jenvman.2017.11.085

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

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, P., Zou, X., Wang, M., Gao, B., Su, Y., and He, G. (2019). Forensic Features and Genetic Structure of the Hotan Uyghur Inferred from 27 Forensic Markers. *Ann. Hum. Biol.* 46, 589–600. doi:10.1080/03014460.2019.1687751

Cheng, T. O. (1985). Medicine and Health Care along the Silk Road. *Arch. Intern. Med.* 145, 137–138. doi:10.1001/archinte.1985.00360010175029

Coulehan, J. (2020). Cultural Exchange. *Ann. Intern. Med.* 172, 158. doi:10.7326/m19-0932

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, 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. (2018b). The First Horse Herders and the Impact of Early Bronze Age Steppe Expansions into Asia. *Science* 360, eaar7711. doi:10.1126/science.aar7711

De Barros Damgaard, P., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., Korneliussen, T., et al. (2018a). Author Correction: 137 Ancient Human Genomes from across the Eurasian Steppes. *Nature* 563, E16. doi:10.1038/s41586-018-0488-1

Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved Whole-Chromosome Phasing for Disease and Population Genetic Studies. *Nat. Methods* 10, 5–6. doi:10.1038/nmeth.2307

Duan, S.-F., Han, P.-J., Wang, Q.-M., Liu, W.-Q., Shi, J.-Y., Li, K., et al. (2018). The Origin and Adaptive Evolution of Domesticated Populations of Yeast from Far East Asia. *Nat. Commun.* 9, 2690. doi:10.1038/s41467-018-05106-7

Elfari, M., Schnur, L. F., Strelkova, M. V., Eisenberger, C. L., Jacobson, R. L., Greenblatt, C. L., et al. (2005). Genetic and Biological Diversity Among Populations of Leishmania Major from Central Asia, the Middle East and Africa. *Microbes Infect.* 7, 93–103. doi:10.1016/j.micinf.2004.09.010

Feng, Q., Lu, Y., Ni, X., Yuan, K., Yang, Y., Yang, X., et al. (2017). Genetic History of Xinjiang's Uyghurs Suggests Bronze Age Multiple-Way Contacts in Eurasia. *Mol. Biol. Evol.* 34, 2572–2582. doi:10.1093/molbev/msx177

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. Leg. Med* 134, 981–983. doi:10.1007/s00414-019-02115-z

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

Goldberg, A., Günther, T., Rosenberg, N. A., and Jakobsson, M. (2017). Reply to Lazaridis and Reich: Robust Model-Based Inference of Male-Biased Admixture during Bronze Age Migration from the Pontic-Caspian Steppe. *Proc. Natl. Acad. Sci. USA* 114, E3875–E3877. doi:10.1073/pnas.1704442114

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. doi:10.1038/nature14317

Heizhati, M., Wang, L., Yao, X., Li, M., Hong, J., Luo, Q., et al. (2020). Prevalence, Awareness, Treatment and Control of Hypertension in Various Ethnic Groups (Hui, Kazakh, Kyrgyz, Mongolian, Tajik) in Xinjiang, Northwest China. *Blood Press.* 29, 276–284. doi:10.1080/08037051.2020.1745055

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

Hemphill, B. E., and Mallory, J. P. (2004). Horse-mounted Invaders from the Russo-Kazakh Steppe or Agricultural Colonists from Western Central Asia? A Craniometric Investigation of the Bronze Age Settlement of Xinjiang. *Am. J. Phys. Anthropol.* 124, 199–222. doi:10.1002/ajpa.10354

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

Hyten, D. L., Cannon, S. B., Song, Q., Weeks, N., Fickus, E. W., Shoemaker, R. C., et al. (2010). High-throughput SNP Discovery through Deep Resequencing of a Reduced Representation Library to Anchor and orient Scaffolds in the Soybean Whole Genome Sequence. *BMC Genomics* 11, 38. doi:10.1186/1471-2164-11-38

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. 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, 890–904. 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, E11248–e11255. doi:10.1073/pnas.1813608115

Juras, A., Makarowicz, P., Chyleński, M., Ehler, E., Malmström, H., Krzewińska, M., et al. (2020). Mitochondrial Genomes from Bronze Age Poland Reveal Genetic Continuity from the Late Neolithic and Additional Genetic Affinities with the Steppe Populations. *Am. J. Phys. Anthropol.* 172, 176–188. doi:10.1002/ajpa.24057

Khan, M. N., Kalsoom, S., and Khan, A. A. (2017). Food Exchange List and Dietary Management of Non-communicable Diseases in Cultural Perspective. *Pak J. Med. Sci.* 33, 1273–1278. doi:10.12669/pjms.335.13330

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

Lan, T., Lin, H., Zhu, W., Tellier, L. C. A. M., Yang, M., Liu, X., et al. (2019). Correction to: Deep Whole-Genome Sequencing of 90 Han Chinese Genomes. *Gigascience* 8, giz001. doi:10.1093/gigascience/giz001

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

Lazaridis, I., and Reich, D. (2017). Failure to Replicate a Genetic Signal for Sex Bias in the Steppe Migration into central Europe. *Proc. Natl. Acad. Sci. USA* 114, E3873–E3874. doi:10.1073/pnas.1704308114

Liu, J., Wu, D., Wang, T., Ji, M., and Wang, X. (2021). Interannual Variability of Dust Height and the Dynamics of its Formation over East Asia. *Sci. Total Environ.* 751, 142288. doi:10.1016/j.scitotenv.2020.142288

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

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

Luo, L., Gao, H., Yao, L., Long, F., Zhang, H., Zhang, L., et al. (2020). Genetic Diversity, Forensic Feature, and Phylogenetic Analysis of Guizhou Tujia Population via 19 X-STRs. *Mol. Genet. Genomic Med.* 8, e1473. doi:10.1002/mgg3.1473

Ma, J., Sun, Q., Zhang, X., and Du, H. (2014). Correlation between the Single Nucleotide Polymorphisms of the Human Phosphodiesterase 4D Gene and the Risk of Cerebral Infarction in the Uygur and Han Ethnic Groups of Xinjiang, China. *Exp. Ther. Med.* 7, 155–160. doi:10.3892/etm.2013.1370

Ma, X., Yang, W., Gao, Y., Pan, Y., Lu, Y., Chen, H., et al. (2021). Genetic Origins and Sex-Biased Admixture of the Huis. *Mol. Biol. Evol.* 38, 3804–3819. doi:10.1093/molbev/msab158

Malyarchuk, B. A., Derenko, M., Denisova, G., Woźniak, M., Rogalla, U., Dambueva, I., et al. (2016). Y Chromosome Haplotype Diversity in Mongolic-Speaking Populations and Gene Conversion at the Duplicated STR DYS385a,b in Haplogroup C3-M407. *J. Hum. Genet.* 61, 491–496. doi:10.1038/jhg.2016.14

Maramovich, A. S., Kosilko, S. A., Innokent'eva, T. I., Voronova, G. A., Bazanova, L. P., Nikitin, A. I., et al. (2008). Plague in China. Threat of Transmission to Regions of Siberia and Far East. *Zh Mikrobiol Epidemiol. Immunobiol*, 95–99.

Matsumoto, G. I., Friedmann, E. I., and Gilichinsky, D. A. (1995). Geochemical Characteristics of Organic Compounds in a Permafrost Sediment Core Sample from Northeast Siberia, Russia. *Proc. NIPR Symp. Antarct Meteorites* 8, 258–267.

Medjugorac, I., Kustermann, W., Lazar, P., Russ, I., and Pirchner, F. (1994). Marker-derived Phylogeny of European Cattle Supports Demic Expansion of Agriculture. *Anim. Genet.* 25 (Suppl. 1), 19–27. doi:10.1111/j.1365-2052.1994.tb00399.x

Miller, J. G., Akiyama, H., and Kapadia, S. (2017). Cultural Variation in Communal versus Exchange Norms: Implications for Social Support. *J. Personal. Soc. Psychol.* 113, 81–94. doi:10.1037/pspi0000091

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. 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, 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

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

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

Peel, K. A., and Talley, C. B. (1996). Scientific and Cultural Exchange Trip to China. *S C Nurse (1994)* 3, 28–29. doi:10.7748/en.3.4.29.s17

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

Robino, A., Mezzavilla, M., Pirastu, N., Dognini, M., Tepper, B. J., and Gasparini, P. (2014). A Population-Based Approach to Study the Impact of PROP Perception on Food Liking in Populations along the Silk Road. *PLoS One* 9, e91716. doi:10.1371/journal.pone.0091716

Rodin, R. E., Dou, Y., Kwon, M., Sherman, M. A., D'gama, A. M., Doan, R. N., et al. (2021). Author Correction: The Landscape of Somatic Mutation in Cerebral Cortex of Autistic and Neurotypical Individuals Revealed by Ultra-deep Whole-Genome Sequencing. *Nat. Neurosci.* 24, 611. doi:10.1038/s41593-021-00830-8

Saag, L., Varul, L., Scheib, C. L., Stenderup, J., Allentoft, M. E., Saag, L., et al. (2017). Extensive Farming in Estonia Started through a Sex-Biased Migration from the Steppe. *Curr. Biol.* 27, 2185–2193. doi:10.1016/j.cub.2017.06.022

Saint Onge, J. M., and Brooks, J. V. (2020). The Exchange and Use of Cultural and Social Capital Among Community Health Workers in the United States. *Sociol. Health Illn* 43, 299–315. doi:10.1111/1467-9566.13219

Saisho, D., and Purugganan, M. D. (2007). Molecular Phylogeography of Domesticated Barley Traces Expansion of Agriculture in the Old World. *Genetics* 177, 1765–1776. doi:10.1534/genetics.107.079491

Sanchez-Burks, J., Lee, F., Choi, I., Nisbett, R., Zhao, S., and Koo, J. (2003). Conversing across Cultures: East-West Communication Styles in Work and Nonwork Contexts. *J. Personal. Soc. Psychol.* 85, 363–372. doi:10.1037/0022-3514.85.2.363

Stoneking, M., and Delfin, F. (2010). The Human Genetic History of East Asia: Weaving a Complex Tapestry. *Curr. Biol.* 20, R188–R193. doi:10.1016/j.cub.2009.11.052

Stoof-Leichsenring, K. R., Liu, S., Jia, W., Li, K., Pestryakova, L. A., Mischke, S., et al. (2020). Plant Diversity in Sedimentary DNA Obtained from High-Latitude (Siberia) and High-Elevation Lakes (China). *Biodivers Data J.* 8, e57089. doi:10.3897/BDJ.8.e57089

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

Sun, N., Ma, P.-C., Yan, S., Wen, S.-Q., Sun, C., Du, P.-X., et al. (2019). Phylogeography of Y-Chromosome Haplogroup Q1a1a-M120, a Paternal Lineage Connecting Populations in Siberia and East Asia. *Ann. Hum. Biol.* 46, 261–266. doi:10.1080/03014460.2019.1632930

Tangkanchanapas, P., Haegeman, A., Ruttink, T., Höfte, M., and De Jonghe, K. (2020). Whole-Genome Deep Sequencing Reveals Host-Driven In-Planta Evolution of Columnea Latent Viroid (CLVd) Quasi-Species Populations. *Int. J. Mol. Sci.* 21, 3262. doi:10.3390/ijms21093262

Uesugi, R., Jouraku, A., Sukonthabhirom Na Pattalung, S., Hinomoto, N., Kuwazaki, S., Kanamori, H., et al. (2021). Origin, Selection, and Spread of Diamide Insecticide Resistance Allele in Field Populations of Diamondback Moth in East and southeastAsia. *Pest Manag. Sci.* 77, 313–324. doi:10.1002/ps.6020

Wang, C.-C., Huang, Y., Yu, X. e., Chen, C., Jin, L., and Li, H. (2016). Agriculture Driving Male Expansion in Neolithic Time. *Sci. China Life Sci.* 59, 643–646. doi:10.1007/s11427-016-5057-y

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

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, 815–820. doi:10.1038/s10038-019-0618-0

Xiaowei, M., Hucai, Z., Shiyu, Q., Yichen, L., Fengqin, C., Ping, X., 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

Xu, G. (2008). Building a Platform for East-West Communication in Stroke Research: Report of the Third International Stroke Summit, Wuhan, China, November 1-3, 2007. *Cerebrovasc. Dis.* 25, 279–280. doi:10.1159/000119636

Xu, S., and Jin, L. (2008). A Genome-wide Analysis of Admixture in Uyghurs and a High-Density Admixture Map for Disease-Gene Discovery. *Am. J. Hum. Genet.* 83, 322–336. doi:10.1016/j.ajhg.2008.08.001

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. B., Tang, S., Yao, X., Yeh, H. Y., Zhang, W., Xie, Z., et al. (2017). The Genetic Admixture in Tibetan-Yi Corridor. *Am. J. Phys. Anthropol.* 164, 522–532. doi:10.1002/ajpa.23291

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

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, 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

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, 620–628. doi:10.1080/03014460.2020.1837952

Zhou, X., Yu, J., Spengler, R. N., Shen, H., Zhao, K., Ge, J., et al. (2020). 5,200-year-old Cereal Grains from the Eastern Altai Mountains Redate the Trans-eurasian Crop Exchange. *Nat. Plants* 6, 78–87. doi:10.1038/s41477-019-0581-y

Zohary, D., and Hopf, M. (1973). Domestication of Pulses in the Old World. *Science* 182, 887–894. doi:10.1126/science.182.4115.887

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: 06 January 2022
doi: 10.3389/fgene.2021.819337

Mitochondrial DNA Footprints from Western Eurasia in Modern Mongolia

Irene Cardinali[1]†, Martin Bodner[2†], Marco Rosario Capodiferro[3†], Christina Amory[2], Nicola Rambaldi Migliore[3], Edgar J. Gomez[4,5], Erdene Myagmar[6], Tumen Dashzeveg[6], Francesco Carano[7], Scott R. Woodward[4], Walther Parson[2,8], Ugo A. Perego[3,4,9], Hovirag Lancioni[1] and Alessandro Achilli[3]**

[1]Department of Chemistry, Biology and Biotechnology, University of Perugia, Perugia, Italy, [2]Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria, [3]Department of Biology and Biotechnology "L. Spallanzani", University of Pavia, Pavia, Italy, [4]Sorenson Molecular Genealogy Foundation, Salt Lake City, UT, United States, [5]FamilySearch Int., Salt Lake City, UT, United States, [6]Department of Anthropology and Archaeology, National University of Mongolia, Ulaanbaatar, Mongolia, [7]Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy, [8]Forensic Science Program, The Pennsylvania State University, State College, PA, United States, [9]Department of Math and Science, Southeastern Community College, Burlington, IA, United States

OPEN ACCESS

Edited by:
Horolma Pamjav,
Hungarian Institute for Forensic Sciences, Hungary

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

***Correspondence:**
Irene Cardinali
cardinali_irene@libero.it
Alessandro Achilli
alessandro.achilli@unipv.it

†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: 21 November 2021
Accepted: 14 December 2021
Published: 06 January 2022

Citation:
Cardinali I, Bodner M, Capodiferro MR, Amory C, Rambaldi Migliore N, Gomez EJ, Myagmar E, Dashzeveg T, Carano F, Woodward SR, Parson W, Perego UA, Lancioni H and Achilli A (2022) Mitochondrial DNA Footprints from Western Eurasia in Modern Mongolia. Front. Genet. 12:819337. doi: 10.3389/fgene.2021.819337

Mongolia is located in a strategic position at the eastern edge of the Eurasian Steppe. Nomadic populations moved across this wide area for millennia before developing more sedentary communities, extended empires, and complex trading networks, which connected western Eurasia and eastern Asia until the late Medieval period. We provided a fine-grained portrait of the mitochondrial DNA (mtDNA) variation observed in present-day Mongolians and capable of revealing gene flows and other demographic processes that took place in Inner Asia, as well as in western Eurasia. The analyses of a novel dataset (N = 2,420) of mtDNAs highlighted a clear matrilineal differentiation within the country due to a mixture of haplotypes with eastern Asian (EAs) and western Eurasian (WEu) origins, which were differentially lost and preserved. In a wider genetic context, the prevalent EAs contribution, larger in eastern and central Mongolian regions, revealed continuous connections with neighboring Asian populations until recent times, as attested by the geographically restricted haplotype-sharing likely facilitated by the Genghis Khan's so-called *Pax Mongolica*. The genetic history beyond the WEu haplogroups, notably detectable on both sides of Mongolia, was more difficult to explain. For this reason, we moved to the analysis of entire mitogenomes (N = 147). Although it was not completely possible to identify specific lineages that evolved *in situ*, two major changes in the effective (female) population size were reconstructed. The more recent one, which began during the late Pleistocene glacial period and became steeper in the early Holocene, was probably the outcome of demographic events connected to western Eurasia. The Neolithic growth could be easily explained by the diffusion of dairy pastoralism, as already proposed, while the late glacial increase indicates, for the first time, a genetic connection with western Eurasian refuges, as supported by the unusual high frequency and internal sub-structure in Mongolia of haplogroup H1, a well-known post-glacial marker in Europe. Bronze Age events, without a significant demographic impact, might explain the age of some mtDNA haplogroups. Finally, a diachronic comparison with available ancient mtDNAs made it

possible to link six mitochondrial lineages of present-day Mongolians to the timeframe and geographic path of the Silk Route.

Keywords: Eurasian Steppe, Inner Asia, Mongolia genetic history, modern mitogenomes, mitochondrial DNA phylogeny, mtDNA haplogroups

INTRODUCTION

The Eurasian Steppe stretches from Europe to Inner Asia and represents an important crossroad in human history, characterized by migrations and admixtures of culturally and genetically distinct populations (Palstra et al., 2015). Mongolia covers most of the Eastern Steppes. Nowadays, it is a presidential republic divided into 21 provinces (*aimags*) and one provincial municipality (Ulaanbaatar); most of the population (71%) lives in urban centers, while the remaining 29%, often tied to nomadic lifestyles, lives in rural areas.

Archaeological evidence (Kovalev and Erdenebaatar, 2009; Wilkin et al., 2021) and genetic studies provided the first information on the complex Mongolian past (Cavalli-Sforza et al., 1994; Comas et al., 1998; Yao et al., 2004; Yang et al., 2008; Yunusbayev et al., 2015; Pugach et al., 2016; Bai et al., 2018). As in another population context (Achilli et al., 2018), archaeogenomics unveiled further details on the emerging scenario of admixture between Eastern and Western Eurasians for the origin of the Central Asian populations, with distinct west-east genetic gradients between different western and eastern Eurasian groups (Jeong et al., 2019; Narasimhan et al., 2019; Ning et al., 2021). In particular, recent analyses of ancient genomes spanning from 6000 before the common era (BCE) to present days revealed at least four ancestral sources that arose in Mongolia through the Neolithic. Two were identified in pre-Bronze Age individuals from northeastern and northern Mongolia and are associated to hunter-gatherer populations from northeast Asia and northern Eurasia, respectively; the third was connected with the Afanasievo culture, an eastward extension of the Yamnaya culture from the Pontic-Caspian steppe (ca. 3300–2200 BCE), which probably introduced dairying practices to the region (ca. 3000 BCE) and was later followed or replaced by the Chemurchek culture (2750–1900 BCE). A genetic mixture of Yamnaya pastoralists and European farmers, the fourth source, appeared ca. 1400 BCE (Jeong et al., 2020; Wang C. C. et al., 2021). During the Middle and Late Bronze Age (ca. 1900–900 BCE), ruminant dairying characterized by intensive nomadic herding without farming was widespread, leading to the development of large-scale polities since the late first millennium BCE. The Xiongnu was the first of different historically documented dynasties and empires, founded by pastoralists in the Early and Late Medieval periods such as Xianbei, Türk, Uyghur, Khitan, and Mongol. During those centuries Mongolia and the Eurasian Steppe represented an important crossroad through the notorious *Silk Road* (founded by the Chinese Han dynasty in 130 BCE) that played a major role in the economic, demographic and cultural processes shaping the history of several Eurasian populations (Comas et al., 1998). The Mongol empire arose in the late 12th century CE when the chieftain Temüjin took the title of *Genghis Khan* ("*Universal Ruler*"). At its peak (1206–1368 CE), the empire stretched from present-day Poland in the west to Korea in the east, and from Siberia in the north to the Gulf of Oman and Vietnam in the south, covering approximately 22% of Earth's total land and with a population of over 100 million people. At the beginning, Genghis Khan used to destroy most infrastructures along the Silk Route, but eventually he decided to adopt a politics of supporting and facilitating commercial and cultural exchanges between regions under his dominion (Liu, 2010). The empire allowed the establishment of the *Pax Mongolica* (1280–1360 CE), indicating a pacific and flourishing period characterized by commercial, cultural, religious, and scientific exchanges between western and eastern populations, including trades between nomadic groups and urban centers (Köstenbauer, 2017). The historical stability under Genghis Khan's rule is also supported when comparing genetic profiles of ancient Mongols with contemporary Mongolians (Zerjal et al., 2003; Bai et al., 2018).

Mitochondrial DNA (mtDNA) provided several pieces of information. Evidence deriving from the mtDNA haplogroups shared between Afanasievo and Yamnaya people supports an eastward migration from the Pontic-Caspian steppes (Allentoft et al., 2015; Narasimhan et al., 2019). The presence of a U5a1 mitochondrial haplotype in an Eneolithic grave, dated at ca. 3000 BCE and associated with the Afanasievo archaeological culture in the Khangai Mountains, attested the presence of people with "western" origin in the east of the Altai Mountains before the Bronze Age (Rogers et al., 2020), in contrast to what was previously proposed (Ricaut et al., 2004a; Ricaut et al., 2004b; Lalueza-Fox et al., 2004; Chikisheva et al., 2007; Keyser et al., 2009; González-Ruiz et al., 2012; Wang C. C. et al., 2021). To further investigate the impact and legacy of mitochondrial lineages with eastern and western origins on the gene pool of modern Mongolian populations, we analyzed the mtDNA profiles of 2,420 individuals with a last known terminal maternal ancestor (TMA) from one of the 20 different Mongolian provinces.

MATERIALS AND METHODS

Sample Collection and DNA Extraction

A total of 2,420 biological samples belonging to unrelated subjects with a Mongolian TMA were collected in different areas of Mongolia, using 10 ml of commercially available mint-flavored mouthwash. Pedigree charts and informed consents were obtained from all participants. The samples were collected in 20 (out of 21) Mongolian provinces: Arkhangai (n = 4), Bayankhongor (n = 2), Bayan-Ölgii (n = 216), Bulgan (n = 5),

Darkhan-Uul (n = 1), Dornod (n = 370), Dornogovi (n = 26), Dundgovi (n = 1), Govi-Altai (n = 8), Khentii (n = 132), Khovd (n = 429), Khövsgöl (n = 307), Ömnögovi (n = 2), Övörkhangai (n = 8), Selenge (n = 4), Sükhbaatar (n = 246), Töv (n = 10), Ulaanbaatar (n = 2), Uvs (n = 132), Zavkhan (n = 167); the remaining samples (n = 348) belonged to individuals who did not provide province information and therefore were associated to an "unspecified" group. Provinces with less than 30 individuals were grouped into three geographic macro-areas by considering their geographic position, biome, and orography: "Gobi Desert" (Bayankhongor, Dornogovi, Dundgovi, Govi-Altai and Ömnögovi), "Khangai Mountains" (Arkhangai and Övörkhangai), "Near Ulaanbaatar" (Bulgan, Darkhan-Uul, Selenge, Töv and Ulaanbaatar) (**Supplementary Table S1**).

Mitochondrial DNA Control Region Analysis

DNA extraction and mtDNA control-region sequencing were performed as in Perego et al. (2012).

All resulting sequences have been deposited in GenBank under accession numbers OL632312-OL634731 and are also available in the EMPOP mtDNA population database (https://empop.online/) under accession number EMP00853. A total of 2,133 haplotypes encompassed the entire mitochondrial control region (CR, ~1122 bps from np 16024 to np 576), while 2,335 haplotypes encompassed at least the HVS1 segment (nps 16024–16365). The sequences were aligned to the revised Cambridge Reference Sequence (rCRS; NC_012920.1) (Andrews et al., 1999) using Sequencher 5.10 (Gene Codes Corporation), in order to visualize electropherograms and identify and register any mutational differences. All samples were classified into haplogroups according to their respective mutational motifs by referring to PhyloTree build 17 (van Oven and Kayser, 2009). Considering the large number of maternal lineages identified in our study we grouped each of them into 18 macro-haplogroups (**Supplementary Table S2**).

Several mtDNA sequence variation parameters were estimated by using DnaSP 5.1 software (Librado and Rozas, 2009). Nucleotide diversity (π or Pi) and haplotype diversity (Hd) were plotted with Tableau 2021tbl 2021.3 onto Mongolian geographic map. In order to graphically display and summarize the mitogenetic relationships among the analyzed individuals, Principal Component Analyses (PCA) were performed using prcomp () from the stats R package (R Core Team, 2021) or the Excel software implemented by XLSTAT. Macro-haplogroup frequencies were used as input data. In intra-Mongolia analyses, the Khangai Mountains and the cosmopolitan "Near Ulaanbaatar" macro-area were excluded due to the low number (<30) of individuals in each of them.

The Mongolian maternal gene pool was further compared with a Eurasian dataset encompassing 546 bps of the control region (nps 16024–16569) obtained from 30,400 individuals from 69 countries/geographic areas (**Supplementary Table S3**). Four population groups outside Mongolia with less than 30 individuals were excluded from these analyses. The final dataset (for a total of 32,486 sequences including 2,133 Mongolian CR sequences from this study) was aligned with MEGAX (Kumar et al., 2018). The genetic distance between groups was also calculated with MEGAX using the p-distance (proportion of nucleotides at which two sequences being compared are different). The obtained distance matrix was used to construct a multidimensional scaling (MDS) using the R function cmdscale () (R Core Team, 2021). The heteroplasmic bases were converted to Ns with DNAsp 6 (Rozas et al., 2017), before calculating the haplotype sharing with Arlequin (Excoffier and Lischer, 2010). The ratio of haplotype sharing was calculated for each population pair by dividing the number of shared haplotypes by the total haplotypes in each paired group. The pairwise haplotype sharing ratio was also calculated separately for haplotypes belonging to eastern Asia (EAs) and western Eurasian (WEu) haplogroups.

Complete Mitogenome Analysis

The entire mitogenome sequences of 147 individuals, representative of different macro-haplogroups, were obtained (**Supplementary Table S4**) by using the Ion Torrent Personal Genome Machine (PGM) and following manufacturers' protocols. The entire mtDNA molecule was amplified with the HID-Ion Ampliseq Mitochondrial Library Preparation; then the template-positive Ion PGM Hi-Q Ion Sphere Particles were prepared with Ion OneTouch 200 Template Kit v2 and sequenced with the Ion PGM Hi-Q Sequencing Kit chemistry on an Ion 318 v2 chip using a multiplexing approach (Parson et al., 2013; Strobl et al., 2018).

The IGV (Integrative Genomics Viewer) software (Thorvaldsdóttir et al., 2013) was used to visualize the BAM files (aligned to the rCRS and produced by the sequencing machine aligning software) and to verify or to search for specific mutational differences throughout the entire mitochondrial genome. Ambiguous positions were analyzed by Sanger-type sequencing until clarity was reached. The quality of mitogenome sequences was checked through SAM2 on EMPOP (Parson and Dür, 2007; Huber et al., 2018), and all samples were classified into haplogroups according to PhyloTree build 17 (http://www.phylotree.org/) (van Oven and Kayser, 2009) using both Haplogrep 2.0 (Weissensteiner et al., 2016) and SAM2 on EMPOP. A few discrepancies using the two approaches were expected, as explained in Dür et al. (2021). However, these differences do not affect the outcomes of the present work. All control-region haplotypes were confirmed by the mitogenome sequences.

All complete mitogenomes (N = 147) are available in GenBank under accession numbers OL619795-OL619941 and in the EMPOP mtDNA population database (https://empop.online/) under accession number EMP00853.

The evolutionary relationships among our modern haplotypes were visualized through the construction of a most parsimonious (MP) tree, built with an updated version of mtPhyl v.5.003 and checked with MEGAX software. One published L3 sequence (accession number DQ341081) was used as outgroup to reconstruct time estimates and demographic trends in BEAST v2.6.6 (Bouckaert et al., 2019), as previously reported (Modi et al., 2020; Capodiferro et al., 2021). 95% of High Posterior Densities (HPD) were plotted for haplogroups younger than 20 thousand years ago (kya). A total of 693 published ancient mitogenomes from

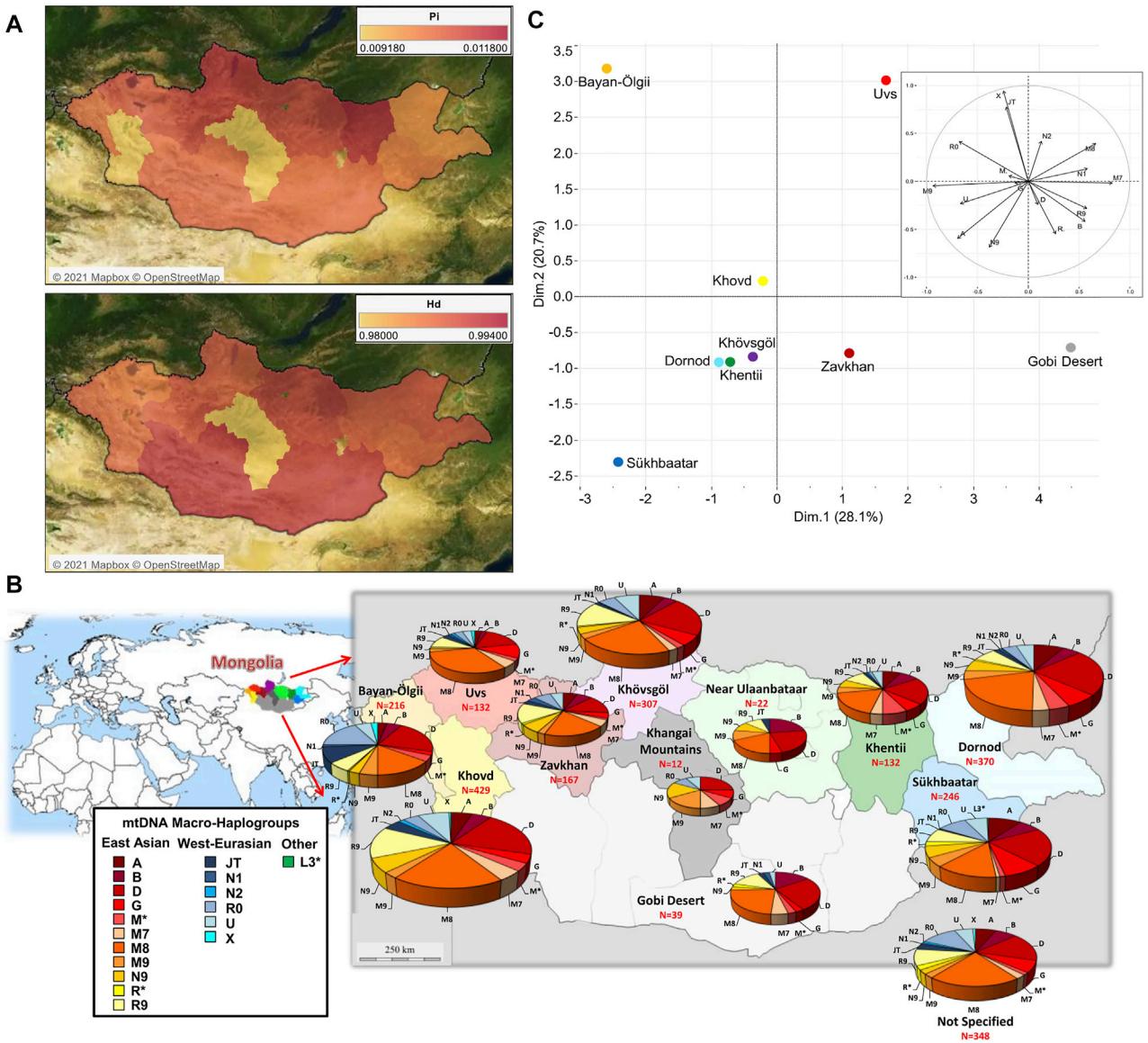


FIGURE 1 | The mtDNA variation within Mongolia based on 2,420 modern control-region sequences. **(A)** Map of genetic variability in each macro-area expressed as nucleotide diversity (Pi) and haplotype diversity (Hd); this analysis was restricted to 2,335 complete HVS1 sequences. **(B)** Pie charts of the distribution of mtDNA macro-haplogroups. **(C)** PCA plot representing the genetic landscape of Mongolia based on macro-haplogroup frequencies; the following groups were excluded: "Khangai Mountains", "Near Ulaanbaatar" and "Not Specified".

different Eurasian regions (including 25 excavated in Mongolia) were also analyzed taking into account haplogroup classification, age and location of the remains (**Supplementary Table S5**).

RESULTS

The Mitochondrial DNA Variation Within Mongolia

The concomitant analysis of 2,335 HVS1 sequences from modern Mongolians identified 160 polymorphic sites (excluding gaps and ambiguous sites) with a Nei's nucleotide diversity (π or Pi) of 0.00809, and a very high haplotype diversity (Hd = 0.986). This genetic diversity is heterogeneously distributed within Mongolia (**Figure 1A**). The "Near Ulaanbaatar" macro-area is characterized by high diversity values, probably due to very recent migrations towards regions around the capital. On the other hand, isolation could better explain the very low diversity in the Khangai Mountains, even if a sample bias due to the low number of individuals (N = 12) cannot be excluded. Interestingly, the Khovd region shows an average number of different haplotypes but a very low nucleotide diversity, which might indicate that people of this western area stretched between the Altai Mountains and the Gobi Desert brought different haplotypes in this area only

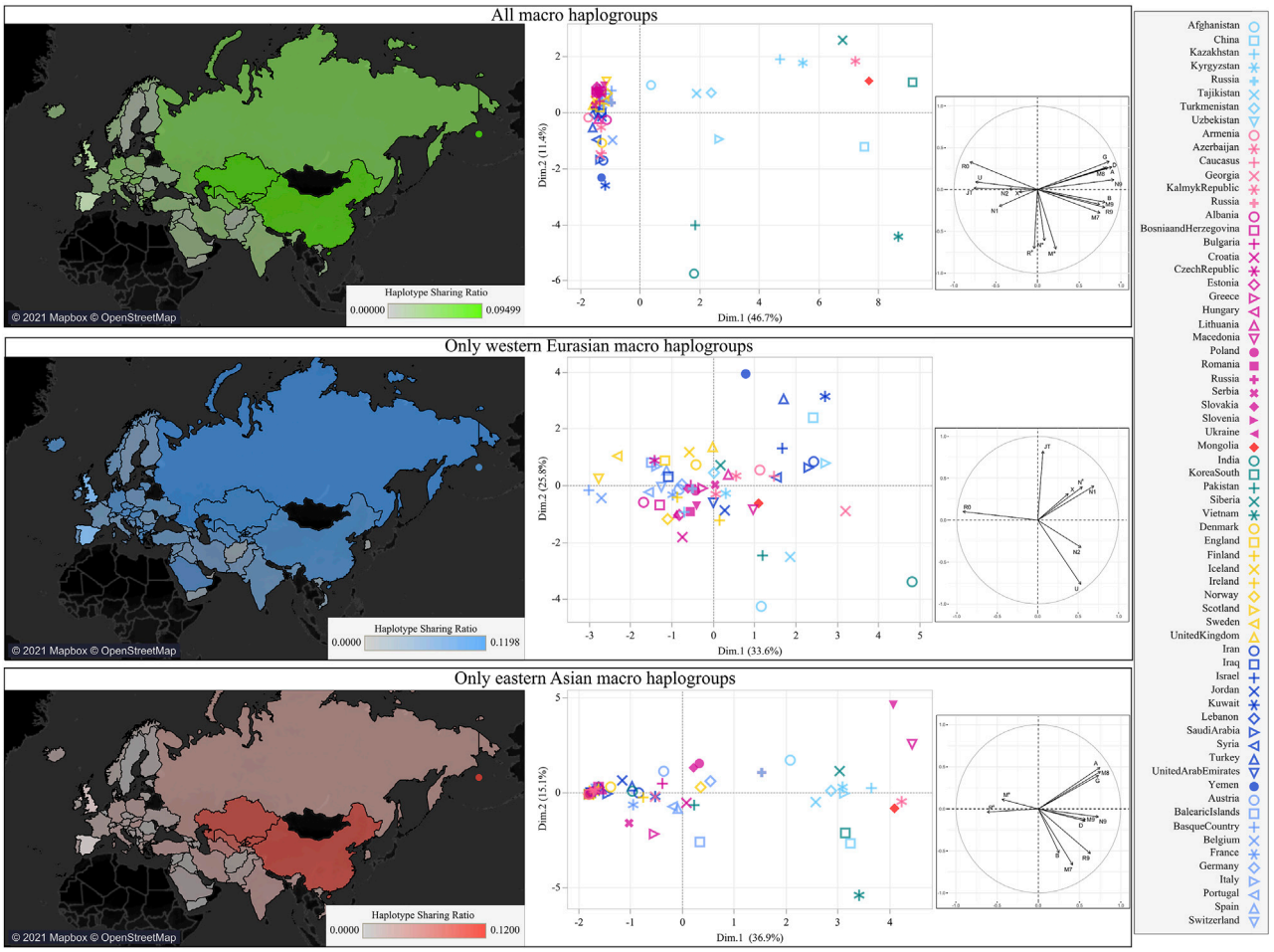


FIGURE 2 | The mtDNA variation of Mongolia in the Eurasian context. Heatmaps based on haplotype sharing between Mongolia and other Eurasian populations are shown on the left. PCA plots based on macro-haplogroup frequencies are shown on the right. All lineages were included in the top panel, the differential contributions of WEu and EAs lineages were explored in the middle and lower panel, respectively.

recently, and their mitogenomes did not yet differentiate from each other.

The overall 2,420 Mongolian mtDNAs were classified into different 413 lineages and sub-lineages, ultimately grouped into 18 macro-haplogroups (A, B, D, G, JT, L3*, M*, M7, M8, M9, N1, N2, N9, R*, R0, R9, U, and X; **Supplementary Table S2**). The macro-haplogroup distribution across the country clearly shows a differential contribution of haplogroups with two distinct geographic origins (**Figure 1B**; **Supplementary Table S1**). As expected, most mtDNAs (1,987 out of 2,420: 82.1%) belong to eastern Asia (EAs) haplogroups with a notable incidence of C (19.6%) and D4 (19.8%) and higher frequencies in the eastern part of Mongolia. In the west, the presence of western Eurasian (WEu) haplogroups is significant (21.7%) and tends to decrease eastwards, but with the lowest occurrences in central regions. The most represented WEu haplogroup is H (6.5%), thus confirming the results obtained in previous studies on Inner Asia (Comas et al., 1998; Wells et al., 2001; Quintana-Murci et al., 2004; Derenko et al., 2014; Lan et al., 2019; Chen et al., 2020; Wang W. et al., 2021; Keyser et al., 2021). A geographic differentiation is clear in the PCA that represents the mtDNA genetic landscape of Mongolia (**Figure 1C**). The PC2 separates the westernmost regions, Bayan-Ölgii, Khovd and Uvs, due to the high contribution of the typical WEu lineages JT, N2, R0 and X. The northern regions (Dornod, Khentii and Khövsgöl) cluster together in the middle of the plot, while PC1 pushes the southern regions of Gobi Desert and Sükhbaatar apart from each other due to the differential distribution of typical EAs lineages.

The Mitochondrial DNA Variation of Mongolia in the Eurasian Context

Considering the various mtDNA contributions to different regions within the country, we evaluated the Mongolian mitochondrial gene pool in the Eurasian context through a MDS plot based on genetic distances, detecting an outlier behavior (**Supplementary Figure S1**). This peculiarity was further investigated through the analysis of haplotype sharing and by building a Eurasian PCA based on macro-haplogroup frequencies (**Figure 2**). The map based on haplotype sharing shows a greater proximity to

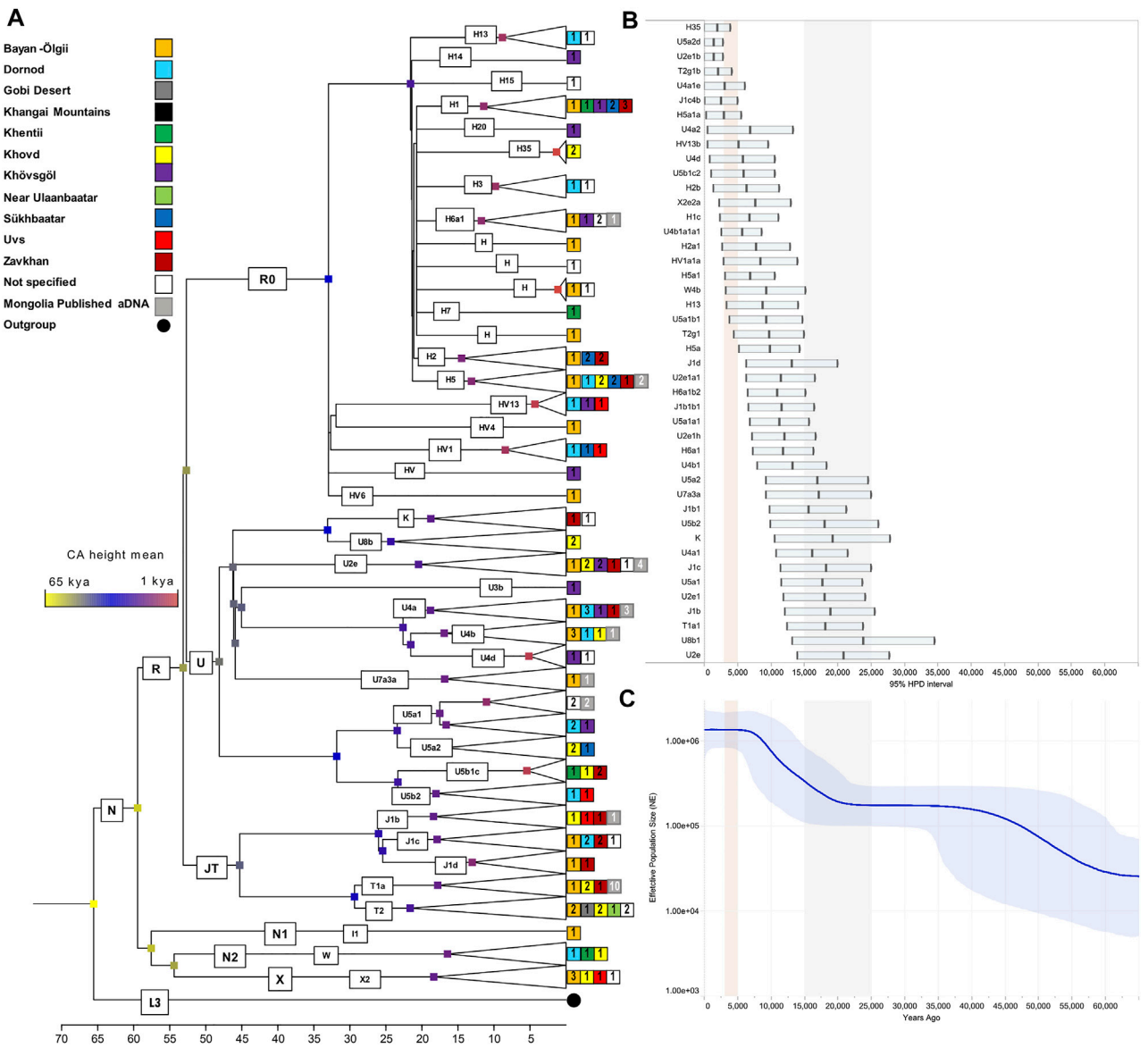


FIGURE 3 | The mtDNA variation in Mongolia based on 121 (new) modern and 25 (published) ancient complete mitogenomes belonging to WEu haplogroups. **(A)** Bayesian tree with internal nodes colored according to common ancestor (CA) average age estimates. **(B)** Age estimates and 95% high posterior densities. LGM timeframe is highlighted in grey, while the red shade indicates Bronze Age. **(C)** Bayesian Skyline Plot (BSP) displaying changes in the effective population size (Ne) through time and considering a generation time of 25 years.

surrounding Asian populations. When we tried to differentiate eastern and western contributions, the legacy of WEu haplogroups was less marked and more widespread. Therefore, it is apparent that the genetic relationships with neighboring populations were mostly driven by EAs lineages and the higher haplotype sharing suggests continuous interactions.

The Mitochondrial DNA Variation in Mongolia Based on Complete Mitogenomes

To deepen the understanding of mtDNA peculiarities of Mongolians, we extended the analysis to the maximum level of resolution. A total of 147 complete mitogenomes were obtained, including 26 mtDNAs representative of EAs lineages (B, C, D, F, M*, R1, R2, R11) and 121 belonging to WEu haplogroups (H, HV, I, J, K, T, U, W, X) (**Supplementary Figure S2**; **Supplementary Table S4**). Through the phylogenetic analysis of our complete mitogenomes, we identified two novel mtDNA sub-branches of haplogroups HV and U5. Three different HV haplotypes from the provinces of Dornod, Khövsgöl and Uvs showed common mutational motifs both in coding (at nps 1654, 9377 and 11152) and control regions (at nps 16184 and 16291) in addition to HV13b motif, thus allowing us to classify these

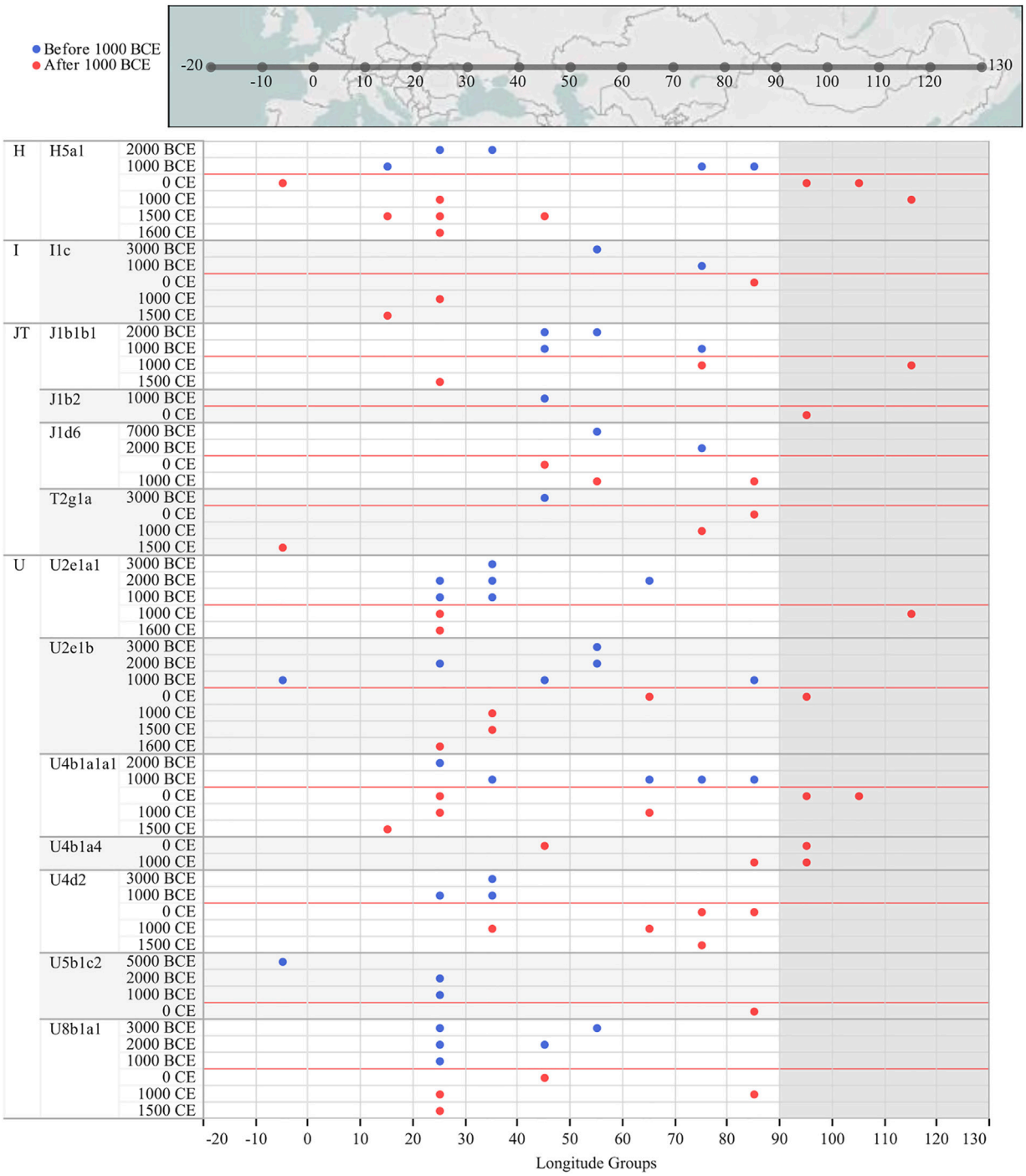


FIGURE 4 | Ancient mitogenomes typical of western Eurasia that were identified among contemporary Mongolians and in ancient remains excavated in Mongolia as well as in other Eurasian regions to the west. A longitude axis is indicated at the bottom and in the geographic map on the top. Mongolia longitudes are shaded. A timeline of 1000 years BCE is reported in red and only those lineages identified in ancient individuals from Mongolia (or nearby regions) and dated after this timeline are reported; see **Supplementary Figure S3** and **Supplementary Table S5** for the entire dataset.

sequences into a postulated haplogroup HV13b1. Four Mongolian individuals from Khentii, Khovd and Zavkhan presenting three different haplotypes were classified as U5b1c2, due to a common transition at np 9110. The entire haplogroup H1 was represented in our control-region dataset with a frequency three times higher (~3%) than the value reported in literature for Inner Asia (Ottoni et al., 2010) and through complete mitogenome analysis, we identified different

sub-clades: H1b and H1c, previously found in Asia as well as in Europe, and H1j, that is very uncommon in the Asian continent.

In order to provide a timeframe to the mtDNA inputs from the west, we focused on the WEu mitogenomes providing Bayesian coalescence ages of internal nodes. Most of these lineages, with a western Eurasian origin but now also identified in Mongolia, coalesced during and soon after the last glacial maximum (LGM, ~25–15 kya; purple nodes) (**Figures 3A,B**). The BSP describes a demographic trend with two major increases of the effective population size (Ne; **Figure 3C**). The first one between 60 and 45 kya likely reflects the initial increase of Ne due to the initial settlement of Eurasia by modern humans during the Pleistocene after the Out-of-Africa exit. Another major increment of Ne seems to be characterized by two steps: the first one started in the late glacial period ~18–15 kya, while the second took place in early Holocene and was probably facilitated by ecological changes associated with the Holocene Climatic Optimum (HCO; ~10–6 kya) (An et al., 2008; Orkhonselenge et al., 2018). When considering the 95% HPD, we noticed that some lineages started to coalesce in the early Bronze Age (~5 kya). They were probably carried by nomadic steppe populations but did not contribute significantly to the Ne. Finally, a very few haplogroups originated in more recent times (<3 kya) and could be linked to historical events. The latter timeframe cannot be accurately described through coalescent estimates taking into account that mitogenomes accumulate one mutation in more than two thousand years on average (Soares et al., 2009; Posth et al., 2016). Therefore, we tried to temporally and geographically reconstruct possible routes marked by the WEu lineages of modern Mongolians, by building a map with the ancient mitogenomes belonging to these lineages that were identified in ancient remains excavated in Mongolia (and nearby regions) as well as in other Eurasian regions to the west. Considering a post-Bronze Age timeline of one thousand years BCE, it is clear that the majority of these WEu lineages were already present in Mongolia in prehistoric times (**Supplementary Figure S3**). However, other WEu lineages (H5a1, I1c, J1b1b1, J1b2, J1d6, T2g1a, U2e1a1, U2e1b, U4b1a1a1, U4b1a4, U4d2, U5b1c2, U8b1a1) reached Mongolia (or nearby regions) after 1000 BCE (**Figure 4**) after appearing in areas 20–40 degrees of longitude to the west.

DISCUSSION

Mongolia is one of the most sparsely populated countries in the world, but complex population interactions occurred across this Eastern Steppe region over the past several millennia (Schurr and Pipes, 2011). Here we characterized the mtDNA of 2,420 modern individuals with a TMA from Mongolia. Our analyses showed a high mitochondrial variation that is heterogeneously distributed across the country. The higher diversity values are present in the northern regions with a dramatic increase in the cosmopolitan area near the capital Ulaanbaatar. If the first finding agrees with reported paleogenomic data (Jeong et al., 2019), the peculiarity of the area around the capital could be better explained by more recent migrations that probably wiped out the original mtDNA gene pool. Other populations probably remained more isolated due to geographic barriers (mountains and desert), which reduced the number of different mitogenomes since ancient times (e.g., in the Khangai Mountains) or only recently (e.g., in the Khovd region).

The majority of mtDNAs belong to haplogroups typical of eastern Asian populations whose frequency decreases from eastern to western regions. An opposite pattern could be observed for those lineages characteristic of western Eurasia. Overall, both EAs and WEu haplogroups contributed to create the mtDNA differentiation currently detectable in Mongolia, as highlighted by the PCA (**Figure 1**). The WEu lineages determine the genetic distinction of the three westernmost provinces (Bayan-Ölgii, Khovd and Uvs), mostly due to macro-haplogroups JT, R0 and X. In particular, macro-haplogroup R0 (mostly made of H mtDNAs, 36.1%) characterizes the outlier position of people living in the Bayan-Ölgii province, which encompasses the Altai Mountains. The Altai Mountains have initially been considered a genetic barrier to gene flows from the west until the recent discovery of ancient people with a WEu mtDNA living on the Mongol Steppe east of the Altai Mountains before the Bronze Age (Rogers et al., 2020). Different EAs lineages distinguish the southern regions, while the northeastern provinces (Dornod, Khentii, Khövsgöl and Sükhbaatar) cluster together, separately from the others, and are characterized by a high number of different mitogenomes that arrived mostly from the surrounding eastern Asian countries. Actually, genetic closeness and continuous interactions of Mongolia with neighboring populations are witnessed by the shared haplotypes of typical EAs lineages. During Early and Late Medieval time, these interactions across the east Asian Steppe were probably facilitated by a series of organized and highly influential nomadic empires, which had a major impact on the demography and geopolitics of Eurasia until the fall of the Mongol Empire (Jeong et al., 2020). A different pattern, more homogeneous and widespread, has been observed when considering shared haplotypes belonging to WEu lineages without pointing to any apparent connection. Therefore, we pushed the analysis to the highest level of resolution, by considering the information hidden in complete mitogenomes and reconstructing and dating the phylogenetic tree of western Eurasian haplogroups found in Mongolia. The ages of some lineages fall in late and post-glacial times and the demographic analysis highlighted a significant increase of mtDNA lineages and population size that started right after the LGM and became steeper during the HCO, probably marking two different demographic events. The first reflects post-glacial re-populations from glacial refuges in western Eurasia, as testified by the haplotype sharing with contemporary populations from Europe and the Balkans peninsula (a well-known refuge area during

LGM), and the high frequency of haplogroup H1, which was indicated as a genetic marker of the post-glacial expansions from western European refuge areas (Achilli et al., 2004). A post-glacial expansion in eastern Asia was already proved for another mtDNA post-glacial marker, haplogroup U5b (Achilli et al., 2005). A later expansion can be probably connected to the climatic amelioration of the early Holocene that was accompanied by the development of farming and pastoralism and more sedentary communities. A mixed ancestry between Yamnaya and European farmers was recently identified by analyzing ancient Bronze Age Mongolians (Jeong et al., 2020; Wang C. C. et al., 2021). We could not identify sub-branches of WEu lineages specific to Mongolia. Therefore, most of the WEu lineages detected in modern Mongolians actually evolved in western Eurasia, and the increments of the population size depicted by our BSP might mirror demographic events that took place in regions to the west of Mongolia. The lack of Mongolia-specific sub-branches might also suggest that the WEu lineages arrived in the Eastern Steppe in more recent times. Certainly, the ages of some WEu lineages between 5 and 3 kya could be linked to Bronze Age migrations across the Eurasian steppes that probably involved also the Afanasievo first (ca. 3300–2500 BCE) and later the Sintashta culture (ca. 2100–1800 BCE). Finally, by searching the available database of ancient mitogenomes for WEu lineages identified in our modern Mongolians, we identified 13 different sub-lineages among remains excavated in Mongolia and dated after the Bronze Age. They might testify for small population movements from the west less than 3,000 ya that can be probably related to commercial routes. Actually, the migration path from western Eurasia to Mongolia marked by some of these mitochondrial sub-lineages (H5a1, J1b2, T2g, U2e1b, U4b1a1a1, and U4b1a4) occurred about 2,500 ya, thus temporally and geographically overlapping with the Silk Route, while other sub-haplogroups, such as J1b1b1 and U2e1a1, seem to have arrived in Mongolia later.

CONCLUSION

The gene pool of present-day Mongolians reflects gene flows and demographic processes that occurred over the past several millennia across the Eurasian Steppe, thus representing an important key to reconstruct the genetic history of Inner Asia as well as western Eurasia. The analyses of a large set (N = 2,420) of mtDNAs allowed us to identify peculiarities of the mitochondrial gene pools of different Mongolian regions. A clear matrilineal differentiation was identified across the country due to the differential contribution of mitochondrial lineages of eastern Asian and western Eurasian origins. The EAs contribution was probably linked to continuous interactions with neighboring regions until present days, presumably including those related to the Mongol Empire expansions (1206–1368 CE). The inputs from the west were more difficult to pinpoint. Therefore, we moved to the analysis of entire mitogenomes (N = 147), which allowed us to date the WEu lineages and to reconstruct demographic trends across time. After the first migration of Paleolithic hunter-gatherers, the major increases of the population size could be linked to post-glacial late Pleistocene expansions and to changes towards a more sedentary lifestyle during the Holocene. However, the lack of Mongolia-specific lineages did not allow to directly study mitogenomes that evolved *in situ*. A few and more recent events have been also reconstructed through the analysis of modern and ancient mitogenomes, some during the Bronze Age, others in the last three thousand years. As if haplogroup H1 (and its sub-clades) might suggest a direct link between Europe and Mongolia, six sub-lineages identified in ancient mitogenomes perfectly match the timeframe and path of the Silk Route and can be still identified in present-day Mongolians. Finally, rather than finding long-distance traces of the Mongol Empire expansion to the west, we identified continuous and recent (female-mediated) connections with neighboring Eastern Asian populations. The geographically restricted sharing of haplotypes from typical EAs mtDNA lineages might represent an outcome of Genghis Khan's so-called *Pax Mongolica* still detectable in present-day Mongolians.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: OL632312–OL634731 for mtDNA control regions and OL619795–OL619941 for complete mitogenomes.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Western Institutional Review Board (WIRB), Olympia, Washington (United States). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

UAP and AA conceived the study. EG, TD, EM, SW, and UAP provided samples and genealogical data. IC, MB, MRC, CA, UAP, HL, and AA performed the laboratory experiments. IC, MB, MRC, CA, NRM, UAP, HL, and AA conducted the data analyses. IC, MB, MRC, CA, NRM, and WP performed the data quality control and validation. EG, EM, FC, and WP contributed to data interpretation. IC, MB, MRC, HL, and AA wrote the original draft with inputs from all co-authors. All authors reviewed and approved the final manuscript.

FUNDING

This research received support from: the National Geographic Society (NGS) grant number HJ-115ER-17 (to

IC); the Italian Ministry of Education, University and Research (MIUR) for Progetti PRIN 2017 20174BTC4R (to AA) and Dipartimenti di Eccellenza Program (2018–2022) - Department of Biology and Biotechnology "L. Spallanzani," University of Pavia (to MC, NRM, and AA); the University of Pavia–INROAd program (to AA). TD and EM are supported by the Ministry of Education, Culture, Science and Sport of Mongolia (grant #2018/25) and the Russian Foundation for Basic Research (grant #18–59–94,020); EM is also supported by the National University of Mongolia (grant number P2020-3955).

ACKNOWLEDGMENTS

We are grateful to all the volunteers who generously participated in this study and made this research possible.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.819337/full#supplementary-material

REFERENCES

Achilli, A., Olivieri, A., Semino, O., and Torroni, A. (2018). Ancient Human Genomes-Keys to Understanding Our Past. *Science* 360 (6392), 964–965. doi:10.1126/science.aat7257

Achilli, A., Rengo, C., Battaglia, V., Pala, M., Olivieri, A., Fornarino, S., et al. (2005). Saami and Berbers-An Unexpected Mitochondrial DNA Link. *Am. J. Hum. Genet.* 76 (5), 883–886. doi:10.1086/430073

Achilli, A., Rengo, C., Magri, C., Battaglia, V., Olivieri, A., Scozzari, R., et al. (2004). The Molecular Dissection of mtDNA Haplogroup H Confirms that the Franco-Cantabrian Glacial Refuge Was a Major Source for the European Gene Pool. *Am. J. Hum. Genet.* 75 (5), 910–918. doi:10.1086/425590

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

An, C.-B., Chen, F.-H., and Barton, L. (2008). Holocene Environmental Changes in Mongolia: A Review. *Glob. Planet. Change* 63 (4), 283–289. doi:10.1016/j.gloplacha.2008.03.007

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

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

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

Capodiferro, M. R., Aram, B., Raveane, A., Rambaldi Migliore, N., Colombo, G., Ongaro, L., et al. (2021). Archaeogenomic Distinctiveness of the Isthmo-Colombian Area. *Cell* 184 (7), 1706–1723. doi:10.1016/j.cell.2021.02.040

Cavalli-Sforza, L. L., Menozzi, P., and Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton NJ: Princeton University Press.

Chen, C., Li, Y., Tao, R., Jin, X., Guo, Y., Cui, W., et al. (2020). The Genetic Structure of Chinese Hui Ethnic Group Revealed by Complete Mitochondrial Genome Analyses Using Massively Parallel Sequencing. *Genes* 11 (11), 1352. doi:10.3390/genes11111352

Chikisheva, T. A., Gubina, M. A., Kulikov, I. V., Karafet, T. M., Voevoda, M. I., and Romaschenko, A. G. (2007). A Paleogenetic Study of the Prehistoric Populations of the Altai. *Archeol. Ethnol. Anthropol. Eurasia* 32, 130–142. doi:10.1134/S156301100704012

Comas, D., Calafell, F., Mateu, E., Pérez-Lezaun, A., Bosch, E., Martínez-Arias, R., et al. (1998). Trading Genes along the Silk Road: mtDNA Sequences and the Origin of central Asian Populations. *Am. J. Hum. Genet.* 63 (6), 1824–1838. doi:10.1086/302133

Derenko, M., Malyarchuk, B., Denisova, G., Perkova, M., Litvinov, A., Grzybowski, T., et al. (2014). Western Eurasian Ancestry in Modern Siberians Based on Mitogenomic Data. *BMC Evol. Biol.* 14, 217. doi:10.1186/s12862-014-0217-9

Dür, A., Huber, N., and Parson, W. (2021). Fine-tuning Phylogenetic Alignment and Haplogrouping of mtDNA Sequences. *Int. J. Mol. Sci.* 22 (11), 5747. doi:10.3390/ijms22115747

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

González-Ruiz, M., Santos, C., Jordana, X., Simón, M., Lalueza-Fox, C., Gigli, E., et al. (2012). Tracing the Origin of the East-West Population Admixture in the Altai Region (Central Asia). *PloS one* 7 (11), e48904. doi:10.1371/journal.pone.0048904

Huber, N., Parson, W., and Dür, A. (2018). Next Generation Database Search Algorithm for Forensic Mitogenome Analyses. *Forensic Sci. Int. Genet.* 37, 204–214. doi:10.1016/j.fsigen.2018.09.001

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. doi:10.1016/j.cell.2020.10.015

Keyser, C., Bouakaze, C., Crubézy, E., Nikolaev, V. G., Montagnon, D., Reis, T., et al. (2009). Ancient DNA Provides New Insights into the History of South Siberian Kurgan People. *Hum. Genet.* 126 (3), 395–410. doi:10.1007/s00439-009-0683-0

Keyser, C., Zvénigorosky, V., Gonzalez, A., Fausser, J.-L., Jagorel, F., Gérard, P., et al. (2021). Genetic Evidence Suggests a Sense of Family, Parity and Conquest in the Xiongnu Iron Age Nomads of Mongolia. *Hum. Genet.* 140 (2), 349–359. doi:10.1007/s00439-020-02209-4

Köstenbauer, J. (2017). Surgical Wisdom and Genghis Khan's Pax Mongolica. *ANZ J. Surg.* 87 (3), 116–120. doi:10.1111/ans.13813

Kovalev, A., and Erdenebaatar, D. (2009). "Discovery of New Cultures of the Bronze Age in Mongolia According to the Data Obtained by the International Central Asian Archaeological Expedition," in *Current Archaeological Research in Mongolia*. Editors J. Bemmann, H. Parzinger, E. Pohl, and D. Tseveendorj (Bonn, DE: University of Bonn).

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 (6), 1547–1549. doi:10.1093/molbev/msy096

Lalueza-Fox, C., Sampietro, M. L., Gilbert, M. T. P., Castri, L., Facchini, F., Pettener, D., et al. (2004). Unravelling Migrations in the Steppe: Mitochondrial DNA Sequences from Ancient central Asians. *Proc. R. Soc. Lond. B* 271 (1542), 941–947. doi:10.1098/rspb.2004.2698

Lan, Q., Xie, T., Jin, X., Fang, Y., Mei, S., Yang, G., et al. (2019). MtDNA Polymorphism Analyses in the Chinese Mongolian Group: Efficiency Evaluation and Further Matrilineal Genetic Structure Exploration. *Mol. Genet. Genomic Med.* 7 (10), e00934. doi:10.1002/mgg3.934

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, X. (2010). *The Silk Road in World History*. New York: Oxford University Press.

Modi, A., Lancioni, H., Cardinali, I., Capodiferro, M. R., Rambaldi Migliore, N., Hussein, A., et al. (2020). The Mitogenome Portrait of Umbria in Central Italy as Depicted by Contemporary Inhabitants and Pre-roman Remains. *Sci. Rep.* 10 (1), 10700. doi:10.1038/s41598-020-67445-0

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., Zheng, H.-X., Zhang, F., Wu, S., Li, C., Zhao, Y., et al. (2021). Ancient Mitochondrial Genomes Reveal Extensive Genetic Influence of the Steppe Pastoralists in Western Xinjiang. *Front. Genet.* 12, 740167. doi:10.3389/fgene.2021.740167

Orkhonselenge, A., Komatsu, G., and Uuganzaya, M. (2018). Middle to Late Holocene Sedimentation Dynamics and Paleoclimatic Conditions in the Lake Ulaan basin, Southern Mongolia. *Geomorphologie* 24 (4), 351–363. doi:10.4000/geomorphologie.12219

Ottoni, C., Primativo, G., Hooshiar Kashani, B., Achilli, A., Martínez-Labarga, C., Biondi, G., et al. (2010). Mitochondrial Haplogroup H1 in north Africa: an Early Holocene Arrival from Iberia. *PloS one* 5 (10), e13378. doi:10.1371/journal.pone.0013378

Palstra, F. P., Heyer, E., and Austerlitz, F. (2015). Statistical Inference on Genetic Data Reveals the Complex Demographic History of Human Populations in central Asia. *Mol. Biol. Evol.* 32 (6), 1411–1424. doi:10.1093/molbev/msv030

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., Strobl, C., Huber, G., Zimmermann, B., Gomes, S. M., Souto, L., et al. (2013). Evaluation of Next Generation mtGenome Sequencing Using the Ion Torrent Personal Genome Machine (PGM). *Forensic Sci. Int. Genet.* 7 (5), 543–549. doi:10.1016/j.fsigen.2013.06.003

Perego, U. A., Lancioni, H., Tribaldos, M., Angerhofer, N., Ekins, J. E., Olivieri, A., et al. (2012). Decrypting the Mitochondrial Gene Pool of Modern Panamanians. *Plos One* 7 (6), e38337. doi:10.1371/journal.pone.0038337

Posth, C., Renaud, G., Mittnik, A., Drucker, D. G., Rougier, H., Cupillard, C., et al. (2016). Pleistocene Mitochondrial Genomes Suggest a Single Major Dispersal of Non-africans and a Late Glacial Population Turnover in Europe. *Curr. Biol.* 26 (6), 827–833. doi:10.1016/j.cub.2016.01.037

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

Quintana-Murci, L., Chaix, R., Wells, R. S., Behar, D. M., Sayar, H., Scozzari, R., et al. (2004). Where West Meets East: the Complex mtDNA Landscape of the Southwest and Central Asian Corridor. *Am. J. Hum. Genet.* 74 (5), 827–845. doi:10.1086/383236

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ricaut, F.-X., Keyser-Tracqui, C., Bourgeois, J., Crubézy, E., and Ludes, B. (2004a). Genetic Analysis of a Scytho-Siberian Skeleton and its Implications for Ancient Central Asian Migrations. *Hum. Biol.* 76 (1), 109–125. doi:10.1353/hub.2004.0025

Ricaut, F.-X., Keyser-Tracqui, C., Cammaert, L., Crubézy, E., and Ludes, B. (2004b). Genetic Analysis and Ethnic Affinities from Two Scytho-Siberian Skeletons. *Am. J. Phys. Anthropol.* 123 (4), 351–360. doi:10.1002/ajpa.10323

Rogers, L. L., Honeychurch, W., Amartuvshin, C., and Kaestle, F. A. (2019). U5a1 Mitochondrial DNA Haplotype Identified in Eneolithic Skeleton from Shatar Chuluu, Mongolia. *Hum. Biol.* 91 (4), 213–223. doi:10.13110/humanbiology.91.4.01

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol. Biol. Evol.* 34 (12), 3299–3302. doi:10.1093/molbev/msx248

Schurr, T., and Pipes, L. (2011). "Prehistory of Mongolian Populations as Revealed by Studies of Osteological, Dental, and Genetic Variation," in *Mapping Mongolia: Situating Mongolia in the World from Geologic Time to the Present*. Editor P. L. W. Sabloff (Philadelphia: University of Pennsylvania Museum Press), 125–165.

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

Strobl, C., Eduardoff, M., Bus, M. M., Allen, M., and Parson, W. (2018). Evaluation of the Precision ID Whole MtDNA Genome Panel for Forensic Analyses. *Forensic Sci. Int. Genet.* 35, 21–25. doi:10.1016/j.fsigen.2018.03.013

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration. *Brief Bioinform.* 14 (2), 178–192. doi:10.1093/bib/bbs017

van Oven, M., and Kayser, M. (2009). Updated Comprehensive Phylogenetic Tree of Global Human Mitochondrial DNA Variation. *Hum. Mutat.* 30 (2), E386–E394. doi:10.1002/humu.20921

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

Wang, W., Ding, M., Gardner, J. D., Wang, Y., Miao, B., Guo, W., et al. (2021). Ancient Xinjiang Mitogenomes Reveal Intense Admixture with High Genetic Diversity. *Sci. Adv.* 7 (14), eabd6690. doi:10.1126/sciadv.abd6690

Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.-J., et al. (2016). HaploGrep 2: Mitochondrial Haplogroup Classification in the Era of High-Throughput Sequencing. *Nucleic Acids Res.* 44 (W1), W58–W63. doi:10.1093/nar/gkw233

Wells, R. S., Yuldasheva, N., Ruzibakiev, R., Underhill, P. A., Evseeva, I., Blue-Smith, J., et al. (2001). The Eurasian Heartland: a continental Perspective on Y-Chromosome Diversity. *Proc. Natl. Acad. Sci.* 98 (18), 10244–10249. doi:10.1073/pnas.171305098

Wilkin, S., Ventresca Miller, A., Fernandes, R., Spengler, R., Taylor, W. T.-T., Brown, D. R., et al. (2021). Dairying Enabled Early Bronze Age Yamnaya Steppe Expansions. *Nature* 598, 629–633. doi:10.1038/s41586-021-03798-4

Yang, L., Tan, S., Yu, H., Zheng, B., Qiao, E., Dong, Y., et al. (2008). Gene Admixture in Ethnic Populations in Upper Part of Silk Road Revealed by mtDNA Polymorphism. *Sci. China Ser. C* 51 (5), 435–444. doi:10.1007/s11427-008-0056-2

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

Zerjal, T., Xue, Y., Bertorelle, G., Wells, R. S., Bao, W., Zhu, S., et al. (2003). The Genetic Legacy of the Mongols. *Am. J. Hum. Genet.* 72 (3), 717–721. doi:10.1086/367774

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: 07 January 2022
doi: 10.3389/fgene.2021.801295

Genetic Relationship Among the Kazakh People Based on Y-STR Markers Reveals Evidence of Genetic Variation Among Tribes and Zhuz

Elmira Khussainova[1], Ilya Kisselev[1,2], Olzhas Iksan[1,3], Bakhytzhan Bekmanov[1,3], Liliya Skvortsova[1], Alexander Garshin[1,3], Elena Kuzovleva[1], Zhassulan Zhaniyazov[1], Gulnur Zhunussova[1], Lyazzat Musralina[1,3], Nurzhibek Kahbatkyzy[1], Almira Amirgaliyeva[1], Mamura Begmanova[1], Akerke Seisenbayeva[1], Kira Bespalova[1], Anastasia Perfilyeva[1], Gulnar Abylkassymova[1], Aldiyar Farkhatuly[4], Sara V. Good[2]* and Leyla Djansugurova[1]*

[1]Institute of Genetics and Physiology, Almaty, Kazakhstan, [2]The University of Winnipeg, Winnipeg, MB, Canada, [3]Al-Farabi Kazakh National University, Almaty, Kazakhstan, [4]Oskemen Bilim-Innovation Lyceum, Ust'-Kamenogorsk, Kazakhstan

Ethnogenesis of Kazakhs took place in Central Asia, a region of high genetic and cultural diversity. Even though archaeological and historical studies have shed some light on the formation of modern Kazakhs, the process of establishment of hierarchical socioeconomic structure in the Steppe remains contentious. In this study, we analyzed haplotype variation at 15 Y-chromosomal short-tandem-repeats obtained from 1171 individuals from 24 tribes representing the three socio-territorial subdivisions (Senior, Middle and Junior zhuz) in Kazakhstan to comprehensively characterize the patrilineal genetic architecture of the Kazakh Steppe. In total, 577 distinct haplotypes were identified belonging to one of 20 haplogroups; 16 predominant haplogroups were confirmed by SNP-genotyping. The haplogroup distribution was skewed towards C2-M217, present in all tribes at a global frequency of 51.9%. Despite signatures of spatial differences in haplotype frequencies, a Mantel test failed to detect a statistically significant correlation between genetic and geographic distance between individuals. An analysis of molecular variance found that ~8.9% of the genetic variance among individuals was attributable to differences among zhuzes and ~20% to differences among tribes within zhuzes. The STRUCTURE analysis of the 1164 individuals indicated the presence of 20 ancestral groups and a complex three-subclade organization of the C2-M217 haplogroup in Kazakhs, a result supported by the multidimensional scaling analysis. Additionally, while the majority of the haplotypes and tribes overlapped, a distinct cluster of the O2 haplogroup, mostly of the Naiman tribe, was observed. Thus, firstly, our analysis indicated that the majority of Kazakh tribes share deep heterogeneous patrilineal ancestries, while a smaller fraction of them are descendants of a founder paternal ancestor. Secondly, we observed a high frequency of the C2-M217 haplogroups along the southern border of Kazakhstan, broadly corresponding to both the path of the Mongolian invasion and the ancient Silk Road. Interestingly, we detected three subclades of the C2-M217 haplogroup that broadly exhibits zhuz-specific clustering. Further study of Kazakh haplotypes variation within a Central Asian context is required to untwist this complex process of ethnogenesis.

OPEN ACCESS

Edited by:
Maxat Zhabagin, National Center for Biotechnology, Kazakhstan

Reviewed by:
Riga Wu, Sun Yat-sen University, China
Miriam Baeta, University of Basque Country UPV/EHU, Spain
Atif Adnan, Naif Arab University for Security Sciences (NAUSS), Saudi Arabia

***Correspondence:**
Elmira Khussainova
khussainova@mail.ru
Sara V. Good
s.good@uwinnipeg.ca

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

Received: *25 October 2021*
Accepted: *10 December 2021*
Published: *07 January 2022*

Citation:
Khussainova E, Kisselev I, Iksan O, Bekmanov B, Skvortsova L, Garshin A, Kuzovleva E, Zhaniyazov Z, Zhunussova G, Musralina L, Kahbatkyzy N, Amirgaliyeva A, Begmanova M, Seisenbayeva A, Bespalova K, Perfilyeva A, Abylkassymova G, Farkhatuly A, Good SV and Djansugurova L (2022) Genetic Relationship Among the Kazakh People Based on Y-STR Markers Reveals Evidence of Genetic Variation Among Tribes and Zhuz. Front. Genet. 12:801295. doi: 10.3389/fgene.2021.801295

Keywords: Y-chromosome, Y-STR, haplotypes, haplogroups, Kazakhstan, MDS plot, Kazakh tribes

INTRODUCTION

Central Asia is a region populated by a wide range of ethnicities and characterized by heterogeneous economic and linguistic landscapes. Being located along the Silk Road, Central Asian populations have been genetically and culturally influenced by a millennia-long interplay between East and West that underpins its highly diverse genetic landscape. Genetic studies of Bronze Age (3100–1300 BC) remains from Central Asia show substantial temporal changes in the genetic composition of populations, indicating extensive migrations and west-to-east expansions of sedentary herders from the western steppe that formed a homogeneous gene pool by the end of the second millennium BCE (Allentoft et al., 2015; Narasimhan et al., 2019; Lalueza-Fox et al., 2004). In the Iron Age (1300–900 BC), nomadic pastoralists spread through the Eurasian steppe, dispersing the Scythian culture. Analysis of ancient DNA from Sakas and Sarmatians burials, belonging to the Scythian culture, demonstrates an increase of Iranian and eastern Eurasian genetic influx in southern and eastern samples, respectively (Gnecchi-Ruscone et al., 2021; Unterländer et al., 2017). During the first millennium CE, multiple confederations and empires were formed on the territory of modern Kazakhstan that were associated with substantial gene flow. For instance, the male-biased westward expansion of the Xiongnu nomads from the eastern steppe led to significant admixture of east Eurasian lineages into Central Sakas and displacement of the Indo-European Kangju and Wusun people (Damgaard et al., 2018). Subsequently, diverse Turkic nomadic states formed and blended into each other, resulting in gene flows between heterogeneous populations of the former Hunnic empire (Damgaard et al., 2018; Gnecchi-Ruscone et al., 2021). Following the Mongol invasion of the territory in 1211, the Golden Horde was established in the 13th century, that underwent series of fragmentation in the following centuries, resulting in the establishment of the Kazakh Khanate (1465–1847). During this time, nomadic tribes of different origins lived throughout the territory of present-day Kazakhstan, and eventually they were organized into three socio-territorial groups (zhuzes) based largely on geographical origin: Senior zhuz, Middle zhuz, and Junior zhuz (**Figure 1**) (Akishev et al., 1996).

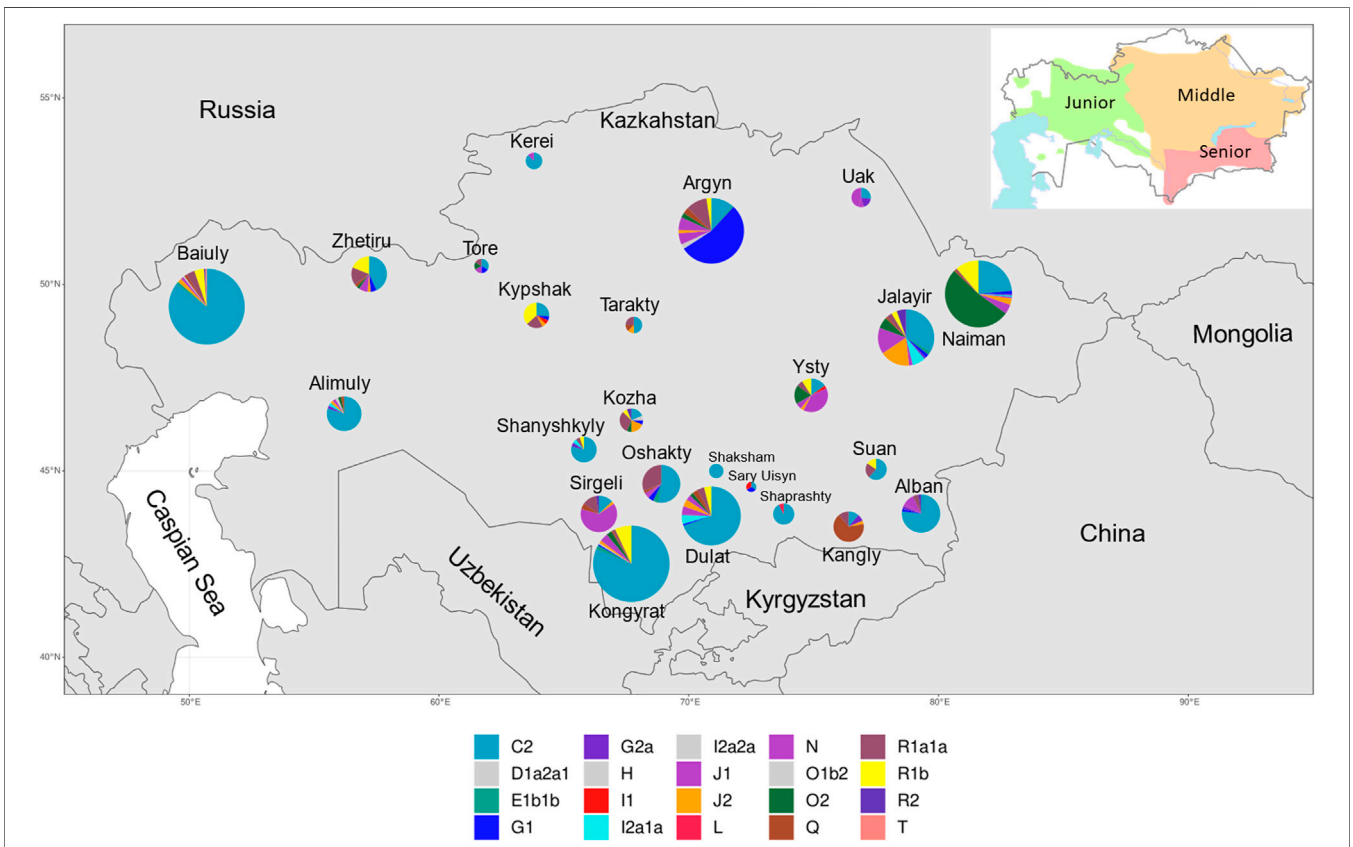


FIGURE 1 | Frequency of Y-STR haplogroups by approximate geographic center of 22 Kazakh tribes, plus the Kozha and Tore in Kazakhstan. Size of pie chart is proportional to the total number of individuals sampled. The relative proportion of each haplogroup (*see* legend for colouration) is depicted in each pie chart. The smaller map in the top right corner shows approximate territories of Kazakh zhuzes in early 20th century. Reproduced from Wassily 2014.

The nomadic society of the Kazakh Steppe was organized based on a hierarchical patrilineal clan system of genealogical lineages. Individuals of the same genealogical lineage claim to share a common ancestor, and multiple genealogical lineages combine into clans that, collectively, form tribes. The 12 tribes of the Senior zhuz primarily occupied Southern and South-Eastern Kazakhstan, the seven tribes of the Middle zhuz reside in Eastern, Northern and Central Kazakhstan, while the three tribes in the Junior zhuz traditionally lived in Western Kazakhstan (**Figure 1**). Some of the steppe clans were not affiliated to the zhuzes, notably the clergy (Kozha and Sunak) and aristocracy (Tore). Representatives of the Kozha and Sunak clans link their ancestry to Islamic missionaries who originated from paternal-line relatives of the Prophet Muhammad. Tore people claim to be direct descendants of Genghis Khan. In contrast to sedentary farmer populations of Central Asia, Kazakhs have practiced exogamous marriages: a partner must be chosen from a different clan, and women integrate into the clan of their husband.

Despite the globalization of the last centuries and the move to a sedentary lifestyle, the tribal-clan structure of the Kazakh people has persisted, and many modern Kazakhs know the tribal affiliation and history of their clan. Being a patrilineal custom, the transgenerational transmission of tribal-clan affiliation resembles inheritance of the non-recombining part of Y chromosome, even though the former is a social entity. The analysis of genetic markers of the Y chromosome has been successfully employed in many studies of human populations to reconstruct migration routes; the combination of extended Y-haplotypes in patrilocal communities with genealogical data can enhance our understanding of the fine-scale demographic dynamics of a population (Wells et al., 2001).

To date, a multitude of studies has employed genetic markers to investigate the genetic diversity and differentiation of the Kazakh population in global (Wells et al., 2001; Underhill et al., 2010; Underhill et al., 2015; Unterländer et al., 2017), regional (Karafet et al., 2002; Lalueza-Fox et al., 2004; M.; Zhabagin et al., 2017) and local (Gokcumen et al., 2008; Balmukhanov et al., 2013; Tarlykov et al., 2013; Wen et al., 2020; M.; Zhabagin et al., 2021) contexts. The accumulated data provide preliminary insights into the demographic history of Kazakhs. For instance, Central Asian populations possess high levels of mtDNA and Y chromosomal haplotype diversity (Underhill et al., 1997; Comas et al., 2004), although paternal genetic markers are less polymorphic than maternal ones (Gokcumen et al., 2008; Tarlykov et al., 2013; Shan et al., 2014a). An earlier study also indicated potential discrepancies between the present-day geographic distribution of the Kazakh tribes and their ancestral relationships to neighboring populations (Tarlykov et al., 2013). Overall, however, prior studies have been hampered by one or more weaknesses such as small sample size, disregard for tribal affiliation, genealogical information, and/or insufficient geographical coverage. Here, we performed one of the largest study to date, N = 1,171 of Y-chromosomal haplotype diversity among all extant Kazakh tribes including the Tore and Kozha, with the primary goal of assessing the relationship of tribes and zhuz among the Kazakh people of modern-day Kazakhstan.

MATERIALS AND METHODS

Samples and DNA extraction: A total of 1171 Kazakh males were included in this study. Blood or saliva was sampled from unrelated males from all-known Kazakh clans throughout five geographic regions in Kazakhstan (**Supplementary Tables S1, S2**). Individual and ethnological information, such as ethnicity and tribal-clan affiliation were self-declared and collected using an approved interview form from all individuals for which blood/saliva samples were obtained. Individuals with admixed ethnicity in their paternal lineage were removed from the study. Written informed consent was obtained from all participants to perform genetic analyses. The study was approved by the local ethics committee for biological research at the National Center for Biotechnology. Genomic DNA was extracted from all samples using the QIAamp DNA Mini Kit (Qiagen, Germany) according to the manufacturer's protocol and quantified spectrophotometrically (BioPhotometer Plus, Eppendorf, Germany) and fluorometrically (Qubit 2.0, Thermo Fisher Scientific, United States).

Y-STR genotyping: Samples were genotyped with the AmpFLSTR Y-filer PCR Amplification Kit (Thermo Fisher Scientific, United States) generating STR profiles at 17-loci (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385a, DYS385b, DYS438, DYS439, DYS437, DYS448, DYS458, DYS456, DYS635, and Y-GATA-H4) (**Supplementary Table S2**). Fragments were separated and visualized on the ABI PRISM 310 Genetic Analyzer (Applied Biosystems, United States) and alleles called using GeneMapper IDX V1.4 (Applied Biosystems, United States). Due to difficulties in correctly assigning alleles to the duplicated DYS385a/b loci, these loci were removed and the haplotype of all individuals at 15 loci were used in further analyses. The raw data were submitted to the Y-Chromosome Haplotype Reference Database (YHRD) and is under the accession number YA004686.

Haplogroup prediction and Y-SNP genotyping: Y-haplogroups were predicted using the online Y-DNA Haplogroup Predictors NevGen (http://www.nevgen.org) as well as Whit-Athey's (http://www.hprg.com). Putative Y-haplogroups for 16 of the 20 predicted haplogroups were then definitively determined through the analysis of 16 Y-SNPs, genotyped by RFLP or allele specific PCR using nine primers designed from previous studies and seven primer pairs designed for this study (**Supplementary Table S3**). In Y-STR haplotypes confirmed by Y-SNP analyses were confirmed for 1164 men (**Supplementary Table S2**). The nomenclature of Karafet et al., 2008, Underhill et al., 2010, and Myres et al., 2011 was used for SNP-analyses as they incorporate the latest information from the International Society of Genetic Genealogy (ISOGG) regarding Y-haplogroup confirmation (https://isogg.org/tree/index.html).

DATA ANALYSES

Haplogroups: To visualize the clustering of Y-haplogroups in the region, the Y-haplogroups of all individuals were plotted with respect to their birthplace and sampling location (**Supplementary Figures S1, S2**, respectively). Haplogroup frequencies were calculated by direct counting and the frequency of Y-haplogroups by tribe calculated and plotted with respect to the approximate geographic center of the tribes.

Haplotypes: The frequency of alleles (repeat lengths) for each Y-STR marker was calculated by direct counting, and the genetic diversity (GD) of single-markers was calculated using Nei's formula $GD = n(1-\Sigma p_i^2)/(n-1)$, where P_i is the relative frequency of the i^{th} allele and n is the sample size (Nei, 1987). The number of unique haplotypes and their frequencies were obtained with the help of the R package Pegas (Paradis, 2010). Haplotype diversity was calculated in an analogous way to GD except by replacing the allele frequencies (Pi) by the relative frequencies of each haplotype. Haplotype discrimination capacity (DC) was calculated as the ratio of unique haplotypes in the sample (i.e. (n haplotypes)/N) ∗100).

Genetic distance and differentiation among tribes: The pairwise genetic distances between individuals in each of the 24 tribes was estimated using Weir and Cockerham pairwise FST for haploid data (Weir and Cockerham 1984). An analysis of molecular variance was performed to assess the components of molecular variance as explained by both tribe (model: ~tribe), and tribe nested in zhuz (~zhuz/tribe). The significance of the covariance components was assessed using a permutation test following Excoffier (Excoffier et al., 1992). A Mantel test was performed to test for a linear relationship between the Euclidean geographic distance between all pairs of individuals' birthplace and the corresponding genetic distance of their Y-haplotypes estimated by Edward's distance, and the significance of the regression was tested by a randomization procedure as implemented in the R-package Ade4 (v 1.7–15).

Structure and Principal Coordinate Analysis: An analysis of the inferred number of ancestral Y-STR groups, K was performed using the program Structure 2.3.4 (Pritchard et al., 2000). Given the nearly complete linkage and haploid nature of the Y-STR, we assessed the haplotype structure within haplogroups and tribes using a "no admixture" model, which assumes that each individual originate from one of the K ancestral groups. The structure analyses was performed on 1163 individuals (all of those that were SNP-genotyped and excluding haplogroup T present in a single individual) using a person's tribal affiliation and haplogroup as priors. To assess the best estimate of K, the number of ancestral Y-STR groups in the total population, simulations were carried out for K = 1 to 25 with three replicates for each value of K. To select the best estimate of K, the second-order rate of change of the likelihood function with respect to K was estimated, using the program Structure Harvester (Earl and vonHoldt 2012), and the value of K with the highest likelihood selected. We also examined the inferred posterior probability that individual i is from the *k*th group assuming a prior probability of 1/K. To further explore the relationship among tribes, a principal coordinate analysis (PcoA aka, multidimensional scaling, MDS) was performed by using the pairwise genetic distance between individuals based on Meirmans PT, as above, performing the PcoA and then extracting the first two dimensions for visualization with the help of the R package Ape (Paradis et al., 2004). The individual coordinates of all 1171 individuals were plotted with respect to *1*) Y-haplogroups and *2*) the tribal affiliations of individuals.

RESULTS

Y-chromosome haplogroups in Kazakh population: 1171 Kazakh males were included in the study: 433 total individuals representing the 12 tribes in the Senior zhuz, 475 from seven tribes in the Middle zhuz, 241 from the Junior zhuz, and 22 samples from the Kozha (n = 16) and Tore (n = 6) tribes (**Supplementary Table S1**). Although the analysis performed here is based on the highest hierarchical affiliation of a person to their tribe/zhuz, all but 142 individuals reported their family lineage/clan and this data could be used in future analyses (given in **Supplementary Table S2**). All individuals included in the study were living in Kazakhstan at the time of the study, 61 were born outside of Kazakhstan (**Supplementary Table S2**, **Supplementary Figure S1** for map of birth locations and **Supplementary Figure S2** for sampling locations). Gene diversity of the single locus markers for each of the 15 Y-STR's ranged from 0.3305 (DYS438) to 0.7699 (DYS635) (**Supplementary Table S4**). The locus with the highest diversity, DYS635, harbored eight allelic classes, while the least diverse loci DYS393, DYS391 and DYS437 each had five alleles, with 129 alleles scored at all 15 loci (**Supplementary Table S4**). Haplotypes were submitted to two online Y-DNA haplogroup predictors (NevGen and Whit-Athey's) to assign a tentative haplogroup to each individual (given in **Supplementary Table S2**). Sixteen of the 20 predicted haplogroups were confirmed by SNP genotyping using primers obtained from the literature or developed for this study (**Supplementary Table S3**). Of the 1171 males for which a Y-haplotype was obtained, 1164 were assigned to one of 16 Y-haplogroups based on the combined Y-STR - SNP data, while seven individuals were assigned to one of four additional Y-haplogroups (D1a2a1, H, I2a2a, O1b2) based on the NevGen prediction alone (**Supplementary Table S2**). These seven individuals were included in some analyses (those based purely on haplotypes for which only the allele sizes/locus are used) but not all analyses; these four haplogroups require Y-SNP genotyping to be confirmed.

The most frequent haplogroup in the Kazakh population was C2-M217 (51.9% - 608 men). Haplogroup C2-M217 was present in all examined tribes and its frequency ranged from 11% in Kangly and Argyn to 100% in Shaksham (n = 6) (**Figure 1** and **Supplementary Table S5**). Another important component of the Kazakh gene pool is represented by the haplogroup R (12.8%), which has three subclades: R1a1a-M17 (6.5%, 76 individuals), R1b-M343 (5.6%, 65 individuals) and R2-M479 (0.8%, nine individuals). The R1a1a-M17 haplogroup was observed in 18 of the 24 tribes, and was most frequent in the Kozha clan (31.3%, five individuals) and Oshakty tribe (31%, 13 individuals).

Subclade R1b-M343 was found in 12 Kazakh tribes and had the highest frequency (36.8%, seven individuals) in the Kypshak tribe. Lastly, the subclade R2-M479 was observed in five tribes at low frequencies, and was most prevalent in the Kozha clan (6.3%, one person) and Jalayir tribe (5.4%, five individuals) (**Figure 1** and **Supplementary Table S5**).

Haplogroups O (represented by O2-M122 and O1b2, the latter predicted by NevGen), G (G1-M285 and G2a-P15), N-M231, J (J1-M267 and J2-M172), and Q-M242 were observed at frequencies <10% across tribes, but were found at higher frequencies in one or a few tribes. For example, haplogroup O2-M122, had a frequency of 8.03% in the sampled population, but was found in 52.3% of the individuals from the Naiman tribe. Haplogroup G had a global frequency of 7.9%, with the majority (7.1%) belonging to subclade G1-M285 and 0.8% to subclade G2a-P15, but the frequency varied between tribes, and ~54% of the males from the Argyn tribe (n = 126) harbored haplogroup G1-M285. The highest frequency of the haplogroup G2a-P15 was observed in the Uak tribe (18.2%, 11 individuals). The N-M231 haplogroup had a global frequency of ~6.9%, but was prevalent in the Sirgeli (64.1% of 39 individuals) and Uak (45.5% of 11 individuals) tribes, Tore tribe (16.7% from six individuals) and Jalayir tribe (15.1% from 93 individuals). Haplogroup J, represented by J1-M267 and J2-M172, had a global frequency of ~6.2%, being frequent in the Ysty tribe (J1-M267 (39.4%, 13 individuals) and J2-M172 (3%, one person)). Additionally, J2-M172 was also observed in the Kozha clan (18.8%, three individuals). Lastly, haplogroup Q had a low overall frequency of ~3.1%, but was highly represented in the Kangly tribe (66.7%, 27 individuals), while its frequency was <5.5% in all other tribes (**Supplementary Table S5**). The other haplogroups show frequencies in the sampled population lower than 2% (**Supplementary Figure S3**).

The assigned Y-haplogroup of each individual was plotted with respect to the location of their birth and sampling location (**Supplementary Figures S1, S2**) and the frequency of the haplogroups by tribe was plotted with respect to the approximate geographic center of the territory occupied by a tribe in the past (**Figure 1**). Overlaying the Y-haplogroup assignments on the map of Kazakhstan with the approximate route of the Mongolian invaders who rampaged through Central Asia in the 13th century (Zerjal et al., 2003), shows the historical mark of this invasion since haplogroup C2 is more frequent in the southern and western portions of the country where the Mongols passed (**Supplementary Figure S1**). For example, tribes in the southern and western regions of Kazakhstan, such as the Alban, Kongyrat, Dulat, Baiuly and Alimuly, have frequencies of C2-M217 ≥ 70%, while tribes located in the center and northeast of the country, such as the Argyn, Uak, Naiman, Kangly, Ysty and Kypshak, have lower frequencies (<30% in most cases) (*see* **Figure 1** and **Supplementary Table S5**).

Haplotype Diversity: From the 1171 individuals, 577 distinct haplotypes were found, of which 429 were observed once and the remaining 148 were observed between 2 and 51 times and 15 haplotypes observed ≥10 times (**Supplementary Table S6**). Overall, this resulted in a Y chromosome haplotype diversity of 0.9938 ± 0.0006, reflecting the deep paternal lineages of different origins in the sample. However, the discriminatory capacity of the samples based on the haplotype frequency distribution was 55.17%, reflecting the high frequency of a few haplotypes. Seven of the top nine most frequent haplotypes belonged to the C2 haplogroup (**Supplementary Table S6**). The two most common haplotypes (Ht1, Ht2, **Supplementary Table S6**) belonged to the C2 haplogroup and were both observed in 51 individuals while a third C2 haplotype (Ht3) was observed in 37 individuals. Ht1 was present in 38 individuals from the Baiuly tribe and six individuals from the Alimuly tribe (**Supplementary Table S7**) both of which belong to the Junior zhuz, located in western Kazakhstan (**Figure 1**). Ht2 was identified in 32 individuals in the Dulat tribe and 10 individuals in Alban, both belonging to the Senior zhuz in the southeastern region (**Figure 1** and **Supplementary Table S7**). Finally, Ht3 was observed in 37 individuals from the Kongyrat, a member of the Middle zhuz (**Figure 1** and **Supplementary Table S7**). The most frequent non-C2 haplotype (Ht4) belonged to haplogroup G1 and was found in 34 individuals, 26 of which were in the Argyn tribe belonging to the Middle zhuz, located in the center-north of the country (**Figure 1** and **Supplementary Table S7**).

Population Structure: The analysis of molecular variance revealed that between ~73% (Model: ~tribe) and ~71% (Model: ~zhuz/tribe) of the variation in Y-STR diversity is attributable to variation within tribes. Variation in Y-STR diversity among zhuz, explained ~9% of the variation in YSTR-diversity, a value that was lower than expected if the diversity was distributed randomly among the highest hierarchical level using a Monte Carlo permutation test, while the amount of variation within tribes was greater than expected (**Table 1** and **Supplementary Figure S4**). Estimates of the fixation index ΦST, were significant and between 0.273 (model ~tribe) and 0.2923 (model ~ zhus/tribe), such that between 27.3 and 29.3% of the total variation is due to inter-tribal differentiation (**Table 1**). Pairwise Weir and Cockerham's F_{ST} values between tribes varied between ~0 (−0.0003, Baiuly vs Alimuly) and 0.19 (Sirgeli vs Shaksham) (**Supplementary Table S8**).

Mantel test: Despite the apparent higher frequency of the C2-M217 haplogroup along the southern and western borders of Kazakhstan, tracking the Mongolian invasion (**Supplementary Figure S1**), there was not a significant correlation between the genetic and geographic distance among individuals based on a linear regression between all individuals pairwise. The regression line explained only 0.12% of the variation in the data and the permutation test failed to reject the hypothesis of no spatial structure (**Supplementary Figure S5**).

Structure analyses: Further analysis of the haplotype structure using the program Structure implementing a "no admixture" model and using both the tribe and haplogroup as priors, found that the best estimate of K was 20 (**Supplementary Table S9**). Visualization of the structure of the Y-STR's by tribe (**Figure 2A**) and haplogroup (**Figure 2B**), confirms the presence of at least three subgroups within C2-M217: lime green (Senior zhuz), teal blue (Middle zhuz) and pink (Junior zhuz). Similarly, it identified the presence of some homogenous haplogroups (e.g. G1, O2 and

TABLE 1 | Analysis of Molecular Variance (AMOVA) among Y-STR haplotypes of 1171 men in Kazakhstan using two models *1*) tribal affiliations (model: ~tribe), and *2*) tribal affiliation nested in Zhus (model: ~zhus/tribe).

Structure design	Source of variation	d.f.	Sum of squares	Variance components (sigma)	Percentage of variation	Permutation test (α = 0.01)
Model: ~Tribe	Among tribes	23	2792.23	2.47	26.77	Less
	Within tribes	1147	7751.01	6.75	73.23	Greater
	Total	1170	10543.24	9.01	ΦST: 0.268	
Model: ~Zhuz/Tribe	Among zhuzs	2	1103.96	0.844	8.91	Less
	Among tribes Within zhuzs	21	1688.27	1.86	19.68	Less
	Within tribes	1147	7751.01	6.75	71.40	Greater
	Total	1170	12353.99	11.125	Φ_{SC}: 0.216	
					Φ_{ST}: 0.286	
					Φ_{CT}: 0.089	

The significance of the covariance components was tested using a Monte Carlo permutation test using an alpha = 0.01: variance components that were less than or greater than expected under the null (permuted) distribution are marked as < or > respectively.

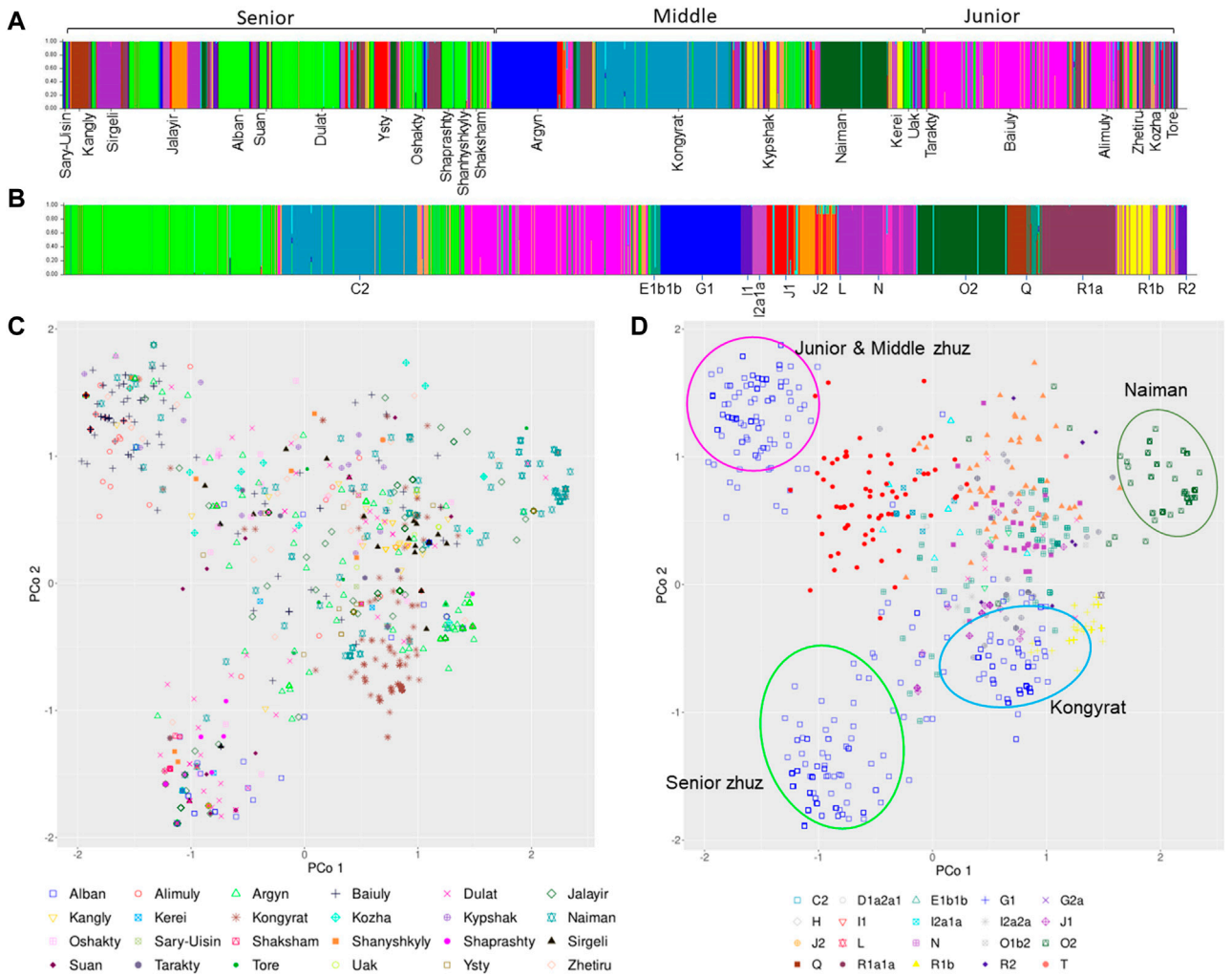


FIGURE 2 | Posterior Probability that an individual is assigned to a specific tribe **(A)** or haplogroup **(B)** based on a no admixture model in the program Structure. Principal Coordinate Analysis (also known as Multidimensional Scaling Plot) showing the relationship among 1171 Kazakh men genotyped at 17 Y-STR's by tribe **(C)** or and haplogroup **(D)**.

R1a1a) that are strongly associated with some tribes (e.g. G1-Argyn, O2 – Naiman).

MultiDimensional Scaling: The principal coordinate analysis (PCoA) provided a closer examination of the relationship among A) tribes and B) haplogroups. This revealed that while many of the haplogroups fall in the middle of the coordinate with somewhat distinct clustering (**Figure 2D**), the haplogroups are diffusely distributed among tribes (**Figure 2C**). On the other hand, there are three distinct clusters of haplogroup C2-M217 that fall in the upper and lower left or middle bottom (**Figure 2B**), which broadly correspond to the three subgroups found in the Structure analyses. Haplogroup C2-M217 cluster 1 is found within numerous tribes in the Junior and Middle zhuz, notably the Baiuly, Argyn, Alimuly and Kypshak and corresponds to the pink haplotypes in the Structure analyses (**Figures 2B,D**). On the other hand, the second C2-M217 haplogroup is found among diverse members of the Senior zhuz, including the Alban, Shaprashty, Oshakty, Dulat and Suan, all found in south-east Kazakhstan and corresponds to the lime-green Structure haplotypes (**Figures 2B,D**). Lastly, the third cluster of C2 haplogroups is found predominantly among members of the Kongyrat tribe corresponding to the teal blue Structure haplotype (**Figures 2B,D**). Other distinct clusters in the MDS analyses pertain to haplotype O2 (upper right, **Figure 2D**), which is most frequent in the Naiman tribe, but is also found in the Jalayir tribe (**Figure 2C**) and for haplotype G1 (brown crosses middle right, **Figure 2D**), which is found in many tribes, particularly the Argyn (**Figure 2C**). The remaining haplotypes have overlapping ranges on the pCoA plot (**Figures 2A,B**). This indicates that there are further subclade differences among individuals within the C2-M217 haplogroup that require further subtyping. Thus, by combining the results of the pCoA and Structure analyses indicate that are broad differences among the haplogroups among the three zhuzes, in particular differences in the C2-M217, haplogroup among zhuz, as well as large differences in the frequency of certain Y-haplogroups among tribes.

DISCUSSION

In this study, we present the most comprehensive study of Y-STR diversity in Kazakhstan, with 1171 samples representing all of the extant tribes living within the territory of Kazakhstan. Haplotype diversity of Y-STR in Kazakhs reached a value of 0.9929, reflecting the deep paternal lineages of different origins in the sample. Our results agree with another recent study from Kazakhstan (0.9936) (Zhabagin et al., 2019), while the haplotype diversity of Kazakhs from Xinjiang (Shan et al., 2014a; Shan et al., 2014b) was found to be lower, possibly due to a founder effect of the Kazakhs that migrated to China. Most of the gene diversity (GD) estimates derived from the Y-profiler Y-STRs were consistent between Kazakhs from Kazakhstan and China (Shan et al., 2014a; Shan et al., 2014b), however, GD of DYS448 was two-fold higher in Xinjiang Kazakhs. Interestingly, the most frequent haplotype in Kazakhs from China is also one the most common haplotypes among Kazakhs from Kazakhstan. The Ht8 haplotype is associated with O2 haplogroup that accounted for 52.2% of individuals from the Naiman tribe, historically situated in Eastern Kazakhstan. Moreover, Kazakh populations in the Altai Region, Russia, are also characterized by a significant fraction of O2 individuals (31–40%) (Dulik et al., 2011; Kharkov 2012). Surprisingly, despite the high frequency of O2 haplotype in a Kazakh population studied by Shan et al. (Shan et al., 2014a; Shan et al., 2014b), no individuals with the O2 haplogroup were found in Kazakhs from Northwest China (Shou et al., 2010).

In contrast to high haplotype diversity, the overall discrimination capacity of the 15 Y-STR loci was only 0.5517. This suggests that despite including only unrelated males in the study, individuals from the same or different tribes may have identical haplotypes presumably reflecting the deep patrilineal descent among Kazakhs. For example, the discrimination capacity in Chinese Han from Shanxi Province, Northern China was 0.9865, which indicates a high potential for differentiating between male individuals in this population (Bai et al., 2013). But in Kazakh populations from Xinjiang, Northwest China the discrimination capacity was 0.5950 (Shan et al., 2014a). The discrimination capacity in our Kazakh sample is also lower compared to some data from European and Asian populations for the same set of 15 Y-STR loci (Turrina et al., 2006; Roewer et al., 2007; Robino et al., 2008; Lacau et al., 2011; Liu et al., 2020). It should be borne in mind that diversity indices vary depending on the Y-STR genotyping systems used. As the number of marker sets increases, diversity indices also increase (Purps et al., 2014; Khubrani et al., 2018; Zhabagin et al., 2019; Liu et al., 2020).

The results of the AMOVA and Mantel tests in our study confirmed that there is significantly less genetic variation among zhuzes than expected under a hierarchical model of genetic structure. This suggests that the zhuz structure is not the primary influence on genetic relationships among Kazakh tribes, and that the division into zhuzes was conditional rather than socio-territorial as suggested by other authors (Ashirbekov et al., 2018; Zhabagin et al., 2018). Nevertheless, approximately, 10% of the genetic variation among individuals was accounted for by variation among zhuzes. Furthermore, the MDS analyses indicated some differences in haplogroup structure among tribes/zhuzes, particularly for the C2-M217 haplogroup (*see* below). In the AMOVA analyses, partitioning the genetic variance within and among tribes (model: ~tribe), revealed that ~27% of the genetic variance is found between tribes and ~73% within tribes. This is similar to a recent study by Ashirbekov et al. (2017) who surveyed Y-STR polymorphism, including more detailed analyses of haplogroup subclades based on SNP polymorphism, for 1269 Kazakh men sampled from 10 tribes in Southern Kazakhstan (Ashirbekov et al., 2017). Overall, they find ~22% of the genetic variance between tribes and 78% within tribes. We did not find evidence of a relationship between genetic and geographic (birthplace) distances of individuals, despite the apparent higher frequency of the C-M217 haplogroup along the southern and western borders of Kazakhstan. We suggest that a more sophisticated spatial analysis is required to show that the C haplogroup exhibits a higher frequency in the southern part of the country.

The MDS and Structure analysis showed that there is considerable diversity in some haplogroups. Particularly interesting is the haplogroup C-M217. It is the most common haplogroup in modern Kazakhs, but the analysis shown here reveal that there at least three distinct sub-clusters of this haplogroup. One of the C2-M217 subgroups is dominant in tribes of the Junior zhuz (mostly in tribes Baiuly, Alimuly), one in the middle (Kongyrat tribe) and one in the Senior zhuz (in almost all tribes). We assume that these clusters represent different daughter branches of the C-M217 haplogroup. A Y-STR study by Ashirbekov et al. sampled 564 individuals (Ashirbekov et al., 2018) from ten tribes in the Senior zhuz, five tribes in the Middle zhuz and three tribes in Junior zhuz, and identified three daughter branches of the haplogroup C-M217: C-M401, C-M86 and C-M407. The authors note the predominance of the C-M401 subgroup in the tribes of the Senior zhuz, the C-M86 subgroup in the Junior zhuz tribes, and the C-M407 subgroup in the tribe of the Middle zhuz - the Kongyrat. This result is consistent with our observation. Further in-depth analysis of a number of single nucleotide markers of the C-M217 haplogroup will make it possible to determine which subgroups of this haplogroup are precisely present in the population of modern Kazakhs.

In conclusion, although several papers have described genetic polymorphism at Y-STR's among Kazakh tribes, this is the largest study, published in English, and represents individuals from tribes in all three zhuzes as well as individuals from the Kozha and Tore tribes. Overall, we find evidence of genetic differentiation between zhuz (~10%) and between tribes within zhuz (~20%) suggesting that there are differences in haplogroup structure among Kazakh tribes. Although we did not find evidence of a linear relationship between genetic and geographic distance among paired individuals (i.e. non-significant Mantel test), we observed the imprint of higher frequencies of C2 haplogroups along the southern and western border of Kazakhstan, which corresponds to both the path of the Mongolian invasion and the approximate route of the ancient Silk Road. A broader spatial and temporal analysis of the Y-STR diversity among Kazakh tribes within the context of other groups in Central Asia is needed to further elucidate this dynamic history Abilev et al., 2012, Akerov, 2016, Artykbaev, 2020, Balaganskaya et al., 2011, Balanovsky et al., 2015, Barinova, 2016, Beisenov et al., 2015, Cai et al., 2011, Damba et al., 2018, Derenko et al., 2007, Ding et al., 2020, Herrera and Garcia-Bertrand, 2018, Hollard et al., 2014, Ilumä e et al., 2016, Ismagulov, 1970, Ismagulov, 1982, Jeong et al., 2019, Keyser et al., 2009, Kozhanuly, 2018, Malyarchuk et al., 2010, Meirmans, 2006, Roewer et al., 2013, Sabitov, 2013, Shi et al., 2005, Shi et al., 2013, Wei et al., 2018, Huang et al., 2018, Zegura et al., 2004, Zhabagin et al., 2016, Zhabagin et al., 2014, Zhabagin et al., 2020, Zhong et al., 2010.

DATA AVAILABILITY STATEMENT

The raw Y-STR data were submitted to the Y-Chromosome Haplotype Reference Database (YHRD) under the accession number YA004686. Other datasets for this study can be found in the **Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Local ethics committee for biological research at the National Center for Biotechnology, Nur-Sultan, Kazakhstan. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: EKh, SG, and LD Performed the experiments: EKh, IK, OI, BB, LS, AG, EKh, GZ, LM, NK, AA, MB, KB, AP, GA, AF, and AS Analyzed the data: EK, ZZ, IK, and SG Contributed reagents/materials/analysis tools: BB and AA Wrote the paper: EKh, SG, and IK.

FUNDING

This research was funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Programs No. OR11465435 and BR05233709).

ACKNOWLEDGMENTS

We thank all the donors for their contributions to this work and all those who helped with sample collection. We are grateful to Oraz Sapargali, Saltanat Abdikerim and other our colleagues for laboratory assistance. We are grateful to Nursaule Rsalieva for verification the correct spelling of the names of the Kazakh tribes. We also thank Talgat Yechshzhanov and Sergey Yegorov for assistance in scientific collaboration.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.801295/full#supplementary-material

REFERENCES

Abilev, S., Malyarchuk, B., Derenko, M., Wozniak, M., Grzybowski, T., and Zakharov, I. (2012). The Y-Chromosome C3* Star-Cluster Attributed to Genghis Khan's Descendants Is Present at High Frequency in the Kerey Clan from Kazakhstan. *Hum. Biol.* 84 (1), 79–89. doi:10.3378/027.084.0106

Akerov, T. A. (2016). On the Origin of the Naiman. *J. Sib. Fed. Univ. Humanit. Soc. Sci.* 9 (9), 2071–2081. doi:10.17516/1997-1370-2016-9-9-2071-2081

Akishev, K., Baipakov, K., and Kumekov, B. (1996). *History of Kazakhstan in 4 Volumes*, 1. Almaty: Atamura. Available at: https://www.twirpx.com/file/988541/.

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

Artykbaev, Z. (2020). O Kazahskom Plemeni Sirgeli: Problemy Proishozhdenija [Kazakh Tribe Sirgeli: Problems of Origin]. *North-Eastern humanitarian Bull.* 2, 34–39. doi:10.25693/SVGV.2020.31.2.004

Ashirbekov, Y., Botbaev, D., Belkozhaev, A., Abaildayev, A., and Aitkhozhina, N. (2017). *Raspredelenie Gaplogrupp Y-Hromosomy Kazahov Juzhno-Kazahstanskoj, Zhambylskoj I Almatinskoj Oblasti [Distribution of Y-Chromosome Haplogroups of Kazakhs in South Kazakhstan, Zhambyl and Almaty Regions]*. Almaty: Reports of National Academy of Sciences of the Republic of Kazakhstan, 6, 25–30.

Ashirbekov, Y. Y., Khrunin, A. V., Botbayev, D. M., Belkozhaev, A. M., Abaildayev, A. O., Rakhimgozhin, M. B., et al. (2018). Molecular Genetic Analysis of Population Structure of the Great Zhuz Kazakh Tribal Union Based on Y-Chromosome Polymorphism. *Mol. Genet. Microbiol. Virol.* 33 (2), 91–96. doi:10.3103/S0891416818020040

Bai, R., Zhang, Z., Liang, Q., Lu, D., Yuan, L., Yang, X., et al. (2013). Haplotype Diversity of 17 Y-STR Loci in a Chinese Han Population Sample from Shanxi Province, Northern China. *Forensic Sci. Int. Genet.* 7 (1), 214–216. doi:10.1016/j.fsigen.2012.10.004

Balaganskaya, O., Lavryashina, M., Kuznetchova, M. A., Romanov, A. G., Dibirova, Kh. D., Frolova, S. A., et al. (2011). *Gene Pool of the Altay Ethnic Groups (From Russia, Kazakhstan, and Mongolia) Analyzed by the Y Chromosomal Markers*. Moscow: Moscow University Anthropology Bulletin (Vestnik Moskovskogo Universiteta. Seria XXIII. Antropologia), 2, 25–36.

Balanovsky, O., Zhabagin, M., Agdzhoyan, A., Chukhryaeva, M., Zaporozhchenko, V., Utevska, O., et al. (2015). Deep Phylogenetic Analysis of Haplogroup G1 Provides Estimates of SNP and STR Mutation Rates on the Human Y-Chromosome and Reveals Migrations of Iranic Speakers. *PLoS ONE* 10 (4), e0122968. doi:10.1371/journal.pone.0122968

Balmukhanov, T., Bekseitov, E., Akhmetollaev, I., Khanseitova, A., Botbaev, D., Belkozhaev, A., et al. (2013). Investigation of Y-Chromosome Microsatellite STR Loci in Kazakh Population. *Proc. Natl. Acad. Sci. Republic Kazakhstan* 4, 91–95.

Barinova, E. B. (2016). Some Aspects of Forming the Population of East and Central Asia in Ancient Times. *RUDN J. World Hist.* 3 (December), 40–51.

Beisenov, A. Z., Ismagulova, A. O., Kitov, E. P., and Kitova, A. O. (2015). *Naselenie Central'nogo Kazahstana V I Tys. Do n.Je. Almaty [The Population of Central Kazakhstan in the 1st Millennium BC]*.

Cai, X., Qin, Z., Wen, B., Xu, S., Wang, Y., Lu, Y., et al. (2011). Human Migration Through Bottlenecks from Southeast Asia into East Asia During Last Glacial Maximum Revealed by Y Chromosomes. *PLoS ONE* 6 (8), e24282. doi:10.1371/journal.pone.0024282

Comas, D., Plaza, S., Wells, R. S., Yuldaseva, N., Lao, O., Calafell, F., et al. (2004). Admixture, Migrations, and Dispersals in Central Asia: Evidence from Maternal DNA Lineages. *Eur. J. Hum. Genet.* 12 (6), 495–504. doi:10.1038/sj.ejhg.5201160

Damba, L. D., Balanovskaya, E. V., Zhabagin, M. K., Yusupov, Y. M., Bogunov, Y. V., Sabitov, Z. M., et al. (2018). Estimating the Impact of the Mongol Expansion Upon the Gene Pool of Tuvans. *Vestn. Vogis* 22 (August), 611–619. doi:10.18699/VJ18.402

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

Derenko, M. V., Malyarchuk, B. A., Wozniak, M., Denisova, G. A., Dambueva, I. K., Dorzhu, C. M., et al. (2007). Distribution of the Male Lineages of Genghis Khan's Descendants in Northern Eurasian Populations. *Russ. J. Genet.* 43 (3), 334–337. doi:10.1134/S1022795407030179

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.* 0 (0), 1–8. doi:10.1080/20961790.2020.1810882

Dulik, M. C., Osipova, L. P., and Schurr, T. G. (2011). Y-chromosome Variation in Altaian Kazakhs Reveals a Common Paternal Gene Pool for Kazakhs and the Influence of Mongolian Expansions. *PLoS ONE* 6 (3), e17548. doi:10.1371/journal.pone.0017548

Earl, D. A., and vonHoldt, B. M. (2012). STRUCTURE HARVESTER: A Website and Program for Visualizing STRUCTURE Output and Implementing the Evanno Method. *Conservation Genet. Resour.* 4 (2), 359–361. doi:10.1007/s12686-011-9548-7

Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of Molecular Variance Inferred from Metric Distances Among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. *Genetics* 131 (2), 479–491. doi:10.1093/genetics/131.2.479

Gnecchi-Ruscone, G. A., Khussainova, E., Kahbatkyzy, N., Musralina, L., Spyrou, M. A., Bianco, R. A., et al. (2021). Ancient Genomic Time Transect from the Central Asian Steppe Unravels the History of the Scythians. *Sci. Adv.* 7 (13), eabe4414. doi:10.1126/sciadv.abe4414

Gokcumen, O., Dulik, M. C., Pai, A. A., Zhadanov, S. I., Rubinstein, S., Osipova, L. P., et al. (2008). Genetic Variation in the Enigmatic Altaian Kazakhs of South-Central Russia: Insights into Turkic Population History. *Am. J. Phys. Anthropol.* 136 (3), 278–293. doi:10.1002/ajpa.20802

Herrera, R. J., and Garcia-Bertrand, R. (2018). *Ancestral DNA, Human Origins, and Migrations*. Elsevier Science.

Hollard, C., Keyser, C., Giscard, P.-H., Tsagaan, T., Bayarkhuu, N., Bemmann, J., et al. (2014). Strong Genetic Admixture in the Altai at the Middle Bronze Age Revealed by Uniparental and Ancestry Informative Markers. *Forensic Sci. Int. Genet.* 12 (September), 199–207. doi:10.1016/j.fsigen.2014.05.012

Huang, Y.-Z., Pamjav, H., Flegontov, P., Stenzl, V., Wen, S.-Q., Tong, X.-Z., et al. (2018). Dispersals of the Siberian Y-Chromosome Haplogroup Q in Eurasia. *Mol. Genet. Genomics* 293 (1), 107–117. doi:10.1007/s00438-017-1363-8

Ilumäe, A.-M., Reidla, M., Chukhryaeva, M., Järve, M., Post, H., Karmin, M., et al. (2016). Human Y Chromosome Haplogroup N: A Non-trivial Time-Resolved Phylogeography that Cuts Across Language Families. *Am. J. Hum. Genet.* 99 (1), 163–173. doi:10.1016/j.ajhg.2016.05.025

Ismagulov, O. (1982). *Jetnicheskaja Antropologija Kazahstana: Somatologicheskoe Issledovanie [Ethnic Anthropology of Kazakhstan: Somatological Study]*. Nauka.

Ismagulov, O. (1970). *Naselenie Kazahstana Ot Jepohi Bronzy Do Sovremennosti: (Paleoantropologicheskoe Issledovanie) [Population of Kazakhstan from the Bronze Age to the Present: (Paleoanthropological Study)]*. Nauka.

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

Karafet, T. M., Mendez, F. L., Meilerman, M. B., Underhill, P. A., Zegura, S. L., and Hammer, M. F. (2008). New Binary Polymorphisms Reshape and Increase Resolution of the Human Y Chromosomal Haplogroup Tree. *Genome Res.* 18 (5), 830–838. doi:10.1101/gr.7172008

Karafet, T. M., Osipova, L. P., Gubina, M. A., Posukh, O. L., Zegura, S. L., and Hammer, M. F. (2002). High Levels of Y-Chromosome Differentiation Among Native Siberian Populations and the Genetic Signature of a Boreal Hunter-Gatherer Way of Life. *Hum. Biol.* 74 (6), 761–789. doi:10.1353/hub.2003.0006

Keyser, C., Bouakaze, C., Crubézy, E., Nikolaev, V. G., Montagnon, D., Reis, T., et al. (2009). Ancient DNA Provides New Insights into the History of South Siberian Kurgan People. *Hum. Genet.* 126 (3), 395–410. doi:10.1007/s00439-009-0683-0

Kharkov, V. N. (2012). *Structure and Phylogeography of Gene Pools of Aboriginal Peoples of Siberia Based on Y-Chromosomal Markers*. Tomsk: Dr Sci Biol thesis (Research Institute of Medical Genetics).

Khubrani, Y. M., Wetton, J. H., and Jobling, M. A. (2018). Extensive Geographical and Social Structure in the Paternal Lineages of Saudi Arabia Revealed by Analysis of 27 Y-STRs. *Forensic Sci. Int. Genet.* 33 (March), 98–105. doi:10.1016/j.fsigen.2017.11.015

Kozhanuly, M. N. (2018). Nekotorye Drevnie Jetnotoponimy Mangistauskogo Regiona [Some Ancient Ethnotonyms of the Mangystau Region]. *Int. Res. J.* 6 (72), 106–109. doi:10.23670/IRJ.2018.72.6.044

Lacau, H., Bukhari, A., Gayden, T., La Salvia, J., Regueiro, M., Stojkovic, O., et al. (2011). Y-STR Profiling in Two Afghanistan Populations. *Leg. Med.* 13 (2), 103–108. doi:10.1016/j.legalmed.2010.11.004

Lalueza-Fox, C., Sampietro, M. L., Gilbert, M. T. P., Castri, L., Facchini, F., Pettener, D., et al. (2004). Unravelling Migrations in the Steppe: Mitochondrial DNA Sequences from Ancient Central Asians. *Proc. R. Soc. Lond. B* 271 (1542), 941–947. doi:10.1098/rspb.2004.2698

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. Leg. Med* 134 (6), 2015–2027. doi:10.1007/s00414-020-02291-3

Malyarchuk, B., Derenko, M., Denisova, G., Wozniak, M., Grzybowski, T., Dambueva, I., et al. (2010). Phylogeography of the Y-Chromosome Haplogroup C in Northern Eurasia. *Ann. Hum. Genet.* 74 (6), 539–546. doi:10.1111/j.1469-1809.2010.00601.x

Meirmans, P. G. (2006). Using the Amova Framework to Estimate a Standardized Genetic Differentiation Measure. *Evolution* 60 (11), 2399–2402. doi:10.1111/j.0014-3820.2006.tb01874.x

Myres, N. M., Rootsi, S., Lin, A. A., Järve, M., King, R. J., Kutuev, I., et al. (2011). A Major Y-Chromosome Haplogroup R1b Holocene Era Founder Effect in Central and Western Europe. *Eur. J. Hum. Genet.* 19 (1), 95–101. doi:10.1038/ejhg.2010.146

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

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R Language. *Bioinformatics* 20 (2), 289–290. doi:10.1093/bioinformatics/btg412

Paradis, E. (2010). Pegas: An R Package for Population Genetics with an Integrated-Modular Approach. *Bioinformatics* 26, 419–420. doi:10.1093/bioinformatics/btp696

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155 (2), 945–959. doi:10.1093/genetics/155.2.945

Purps, J., Siegert, S., Willuweit, S., Nagy, M., Alves, C., Salazar, R., et al. (2014). A Global Analysis of Y-Chromosomal Haplotype Diversity for 23 STR Loci. *Forensic Sci. Int. Genet.* 12 (September), 12–23. doi:10.1016/j.fsigen.2014.04.008

Robino, C., Crobu, F., Di Gaetano, C., Bekada, A., Benhamamouch, S., Cerutti, N., et al. (2008). Analysis of Y-Chromosomal SNP Haplogroups and STR Haplotypes in an Algerian Population Sample. *Int. J. Leg. Med.* 122 (3), 251–255. doi:10.1007/s00414-007-0203-5

Roewer, L., Krüger, C., Willuweit, S., Nagy, M., Rodig, H., Kokshunova, L., et al. (2007). Y-chromosomal STR Haplotypes in Kalmyk Population Samples. *Forensic Sci. Int.* 173 (2–3), 204–209. doi:10.1016/j.forsciint.2006.11.013

Roewer, L., Nothnagel, M., Gusmão, L., Gomes, V., González, M., Corach, D., et al. (2013). Continent-Wide Decoupling of Y-Chromosomal Genetic Variation from Language and Geography in Native South Americans. *Plos Genet.* 9 (4), e1003460. doi:10.1371/journal.pgen.1003460

Sabitov, Z. (2013). Jetnogenez Kazahov S Tochki Zrenija Populjacionnoj Genetiki [Ethogenesis of Kazakhs: Population Genetics Perspective]. *The Russ. J. Genet. Genealogy* 5 (1), 29–47.

Shan, W., Ablimit, A., Zhou, W., Zhang, F., Ma, Z., and Zheng, X. (2014a). Genetic Polymorphism of 17 Y Chromosomal STRs in Kazakh and Uighur Populations from Xinjiang, China. *Int. J. Leg. Med.* 128 (5), 743–744. doi:10.1007/s00414-013-0948-y

Shan, W., Ren, Z., Wu, W., Hao, H., Abulimiti, A., Chen, K., et al. (2014b). Maternal and Paternal Diversity in Xinjiang Kazakh Population from China. *Russ. J. Genet.* 50 (11), 1218–1229. doi:10.1134/S1022795414110143

Shi, H., Dong, Y.-l., Wen, B., Xiao, C.-J., Underhill, P. A., Shen, P.-d., et al. (2005). Y-chromosome Evidence of Southern Origin of the East Asian-specific Haplogroup O3-M122. *Am. J. Hum. Genet.* 77 (3), 408–419. doi:10.1086/444436

Shi, H., Qi, X., Zhong, H., Peng, Y., Zhang, X., Ma, R. Z., et al. (2013). Genetic Evidence of an East Asian Origin and Paleolithic Northward Migration of Y-Chromosome Haplogroup N. *PLoS ONE* 8 (6), e66102. doi:10.1371/journal.pone.0066102

Shou, W.-H., Qiao, E.-F., Wei, C.-Y., Dong, Y.-L., Tan, S.-J., Shi, H., et al. (2010). Y-chromosome Distributions Among Populations in Northwest China Identify Significant Contribution from Central Asian Pastoralists and Lesser Influence of Western Eurasians. *J. Hum. Genet.* 55 (5), 314–322. doi:10.1038/jhg.2010.30

Tarlykov, P. V., Zholdybayeva, E. V., Akilzhanova, A. R., Nurkina, Z. M., Sabitov, Z. M., Rakhypbekov, T. K., et al. (2013). Mitochondrial and Y-Chromosomal Profile of the Kazakh Population from East Kazakhstan. *Croat. Med. J.* 54 (1), 17–24. doi:10.3325/cmj.2013.54.17

Turrina, S., Atzei, R., and De Leo, D. (2006). Y-chromosomal STR Haplotypes in a Northeast Italian Population Sample Using 17plex Loci PCR Assay. *Int. J. Leg. Med* 120 (1), 56–59. doi:10.1007/s00414-005-0054-x

Underhill, P. A., Jin, L., Lin, A. A., Mehdi, S. Q., Jenkins, T., Vollrath, D., et al. (1997). Detection of Numerous Y Chromosome Biallelic Polymorphisms by Denaturing High-Performance Liquid Chromatography. *Genome Res.* 7 (10), 996–1005. doi:10.1101/gr.7.10.996

Underhill, P. A., Myres, N. M., Rootsi, S., Metspalu, M., Zhivotovsky, L. A., King, R. J., et al. (2010). Separating the Post-Glacial Coancestry of European and Asian Y Chromosomes within Haplogroup R1a. *Eur. J. Hum. Genet.* 18 (4), 479–484. doi:10.1038/ejhg.2009.194

Underhill, P. A., Poznik, G. D., Rootsi, S., Järve, M., Lin, A. A., Wang, J., et al. (2015). The Phylogenetic and Geographic Structure of Y-Chromosome Haplogroup R1a. *Eur. J. Hum. Genet.* 23 (1), 124–131. doi:10.1038/ejhg.2014.50

Unterländer, M., Palstra, F., Lazaridis, I., Pilipenko, A., Hofmanová, Z., Groß, M., et al. (2017). Ancestry and Demography and Descendants of Iron Age Nomads of the Eurasian Steppe. *Nat. Commun.* 8 (March), 14615. doi:10.1038/ncomms14615

Wassily (2014). *Approximate Areas Occupied by the Three Kazakh Zhuzes in the Early 20th century. Image*. Available at: https://upload.wikimedia.org/wikipedia/commons/e/ef/%D0%96%D1%83%D0%B7.svg.

Wei, T., Liao, F., Wang, Y., Pan, C., Xiao, C., and Huang, D. (2018). A Novel Multiplex Assay of SNP-STR Markers for Forensic Purpose. *PLoS ONE* 13 (7), e0200700. doi:10.1371/journal.pone.0200700

Weir, B. S., and Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution* 38 (6), 1358–1370. doi:10.2307/2408641

Wells, R. S., Yuldasheva, N., Ruzibakiev, R., Underhill, P. A., Evseeva, I., Blue-Smith, J., et al. (2001). The Eurasian Heartland: A Continental Perspective on Y-Chromosome Diversity. *Proc. Natl. Acad. Sci.* 98 (18), 10244–10249. doi:10.1073/pnas.171305098

Wen, S.-Q., Sun, C., Song, D.-L., Huang, Y.-Z., Tong, X.-Z., Meng, H.-L., et al. (2020). Y-chromosome Evidence Confirmed the Kerei-Abakh Origin of Aksay Kazakhs. *J. Hum. Genet.* 65 (9), 797–803. doi:10.1038/s10038-020-0759-1

Zegura, S. L., Karafet, T. M., Zhivotovsky, L. A., and Hammer, M. F. (2004). High-Resolution SNPs and Microsatellite Haplotypes Point to a Single, Recent Entry of Native American Y Chromosomes into the Americas. *Mol. Biol. Evol.* 21 (1), 164–175. doi:10.1093/molbev/msh009

Zerjal, T., Xue, Y., Bertorelle, G., Wells, R. S., Bao, W., Zhu, S., et al. (2003). The Genetic Legacy of the Mongols. *Am. J. Hum. Genet.* 72 (3), 717–721. doi:10.1086/367774

Zhabagin, M., Balanovska, E., Sabitov, Z., Kuznetsova, M., Agdzhoyan, A., Balaganskaya, O., et al. (2017). The Connection of the Genetic, Cultural and Geographic Landscapes of Transoxiana. *Sci. Rep.* 7 (1), 3085. doi:10.1038/s41598-017-03176-z

Zhabagin, M., Dibirova, H. D., Frolova, S. A., Sabitov, Z., Yusupov, Y. M., Utevska, O., et al. (2014). The Relation between the Y-Chromosomal Variation and the Clan Structure: The Gene Pool of the Steppe Aristocracy and the Steppe Clergy of the Kazakhs. *Mosc. Univ. Anthropol. Bull.* 1, 96–101.

Zhabagin, M. K., Balanovsky, O. E., Sabitov, Z. M., Temirgaliyev, A. Z., Agdzhoyan, A. T., Koshel, S. M., et al. (2018). Reconstructing the Genetic Structure of the Kazakh from Clan Distribution Data. *Vestn. Vogis* 22 (November), 895–904. doi:10.18699/VJ18.431

Zhabagin, M. K., Sabitov, Z., Agdzhoyan, A., Yusupov, Y. M., Bogunov, Y., Lavryashina, M. B., et al. (2016). *Genezis Krupnejshej Rodoplemennoj Gruppy Kazahov – Argynov – V Kontkste Populjacionnoj Genetiki [Genesis of the Largest Tribal Group of Kazakhs - the Argyns - in the Context of Population Genetics]*. Moscow: Vestnik Moskovskogo Universiteta. Seria XXIII. Antropologia [Moscow University Anthropology Bulletin] 4, 59–68.

Zhabagin, M., Sabitov, Z., Tarlykov, P., Tazhigulova, I., Junissova, Z., Yerezhepov, D., et al. (2020). The Medieval Mongolian Roots of Y-Chromosomal Lineages from South Kazakhstan. *BMC Genet.* 21 (S1), 87. doi:10.1186/s12863-020-00897-5

Zhabagin, M., Sabitov, Z., Tazhigulova, I., Alborova, I., Agdzhoyan, A., Wei, L.-H., et al. (2021). Medieval Super-grandfather Founder of Western Kazakh Clans

from Haplogroup C2a1a2-M48. *J. Hum. Genet.* 66, 707–716. doi:10.1038/s10038-021-00901-5

Zhabagin, M., Sarkytbayeva, A., Tazhigulova, I., Yerezhepov, D., Li, S., Akilzhanov, R., et al. (2019). Development of the Kazakhstan Y-Chromosome Haplotype Reference Database: Analysis of 27 Y-STR in Kazakh Population. *Int. J. Leg. Med* 133 (4), 1029–1032. doi:10.1007/s00414-018-1859-8

Zhong, H., Shi, H., Qi, X.-B., Xiao, C.-J., Jin, L., Ma, R. Z., et al. (2010). Global Distribution of Y-Chromosome Haplogroup C Reveals the Prehistoric Migration Routes of African Exodus and Early Settlement in East Asia. *J. Hum. Genet.* 55 (7), 428–435. doi:10.1038/jhg.2010.40

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: 10 January 2022
doi: 10.3389/fgene.2021.744645

Systematic Evaluation of a Novel 6-dye Direct and Multiplex PCR-CE-Based InDel Typing System for Forensic Purposes

Haoliang Fan[1,2], Yitong He[1], Shuanglin Li[1], Qiqian Xie[1], Fenfen Wang[3], Zhengming Du[3], Yating Fang[1], Pingming Qiu[1]† and Bofeng Zhu[1,4,5]*†*

[1]Guangzhou Key Laboratory of Forensic Multi-Omics for Precision Identification, School of Forensic Medicine, Southern Medical University, Guangzhou, China, [2]School of Basic Medicine and Life Science, Hainan Medical University, Haikou, China, [3]First Clinical Medical College, Hainan Medical University, Haikou, China, [4]Clinical Research Center of Shaanxi Province for Dental and Maxillofacial Diseases, College of Stomatology, Xi'an Jiaotong University, Xi'an, China, [5]Key Laboratory of Shaanxi Province for Craniofacial Precision Medicine Research, College of Stomatology, Xi'an Jiaotong University, Xi'an, China

OPEN ACCESS

Edited by:
Shaoqing Wen,
Fudan University, China

Reviewed by:
Mengge Wang,
Sichuan University, China
Atif Adnan,
China Medical University, China

***Correspondence:**
Pingming Qiu
qiupmfy@126.com
Bofeng Zhu
zhubofeng7372@126.com

†ORCID:
Pingming Qiu
orcid.org/0000-0002-5579-1124
Bofeng Zhu
orcid.org/0000-0002-9038-2342

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

Received: 20 July 2021
Accepted: 29 October 2021
Published: 10 January 2022

Citation:
Fan H, He Y, Li S, Xie Q, Wang F, Du Z, Fang Y, Qiu P and Zhu B (2022) Systematic Evaluation of a Novel 6-dye Direct and Multiplex PCR-CE-Based InDel Typing System for Forensic Purposes. Front. Genet. 12:744645. doi: 10.3389/fgene.2021.744645

Insertion/deletion (InDel) polymorphisms, combined desirable characteristics of both short tandem repeats (STRs) and single nucleotide polymorphisms (SNPs), are considerable potential in the fields of forensic practices and population genetics. However, most commercial InDel kits designed based on non-Asians limited extensive forensic applications in East Asian (EAS) populations. Recently, a novel 6-dye direct and multiplex PCR-CE-based typing system was designed on the basis of genome-wide EAS population data, which could amplify 60 molecular genetic markers, consisting of 57 autosomal InDels (A-InDels), 2 Y-chromosomal InDels (Y-InDels), and Amelogenin in a single PCR reaction and detect by capillary electrophoresis, simultaneously. In the present study, the DNA profiles of 279 unrelated individuals from the Hainan Li group were generated by the novel typing system. In addition, we collected two A-InDel sets to evaluate the forensic performances of the novel system in the 1,000 Genomes Project (1KG) populations and Hainan Li group. For the Universal A-InDel set (UAIS, containing 44 A-InDels) the cumulative power of discrimination (CPD) ranged from $1–1.03 \times 10^{-14}$ to $1–1.27 \times 10^{-18}$, and the cumulative power of exclusion (CPE) varied from 0.993634 to 0.999908 in the 1KG populations. For the East Asia-based A-InDel set (EAIS, containing 57 A-InDels) the CPD spanned from $1–1.32 \times 10^{-23}$ to $1–9.42 \times 10^{-24}$, and the CPE ranged from 0.999965 to 0.999997. In the Hainan Li group, the average heterozygote (He) was 0.4666 (0.2366–0.5448), and the polymorphism information content (PIC) spanned from 0.2116 to 0.3750 (mean PIC: 0.3563 ± 0.0291). In total, the CPD and CPE of 57 A-InDels were $1–1.32 \times 10^{-23}$ and 0.999965, respectively. Consequently, the novel 6-dye direct and multiplex PCR-CE-based typing system could be considered as the reliable and robust tool for human identification and intercontinental population differentiation, and supplied additional information for kinship analysis in the 1KG populations and Hainan Li group.

Keywords: InDel, PCR-CE, East Asian population, Hainan Li group, 1000 Genomes Project, Human identification, Intercontinental population differentiation

INTRODUCTION

Insertion/deletion (InDel) polymorphisms, the length polymorphisms resulting from the insertion or deletion of one or more nucleotides, are gradually becoming a type of alternative genetic markers for forensic purposes (Bus et al., 2016; Sun et al., 2016; Caputo et al., 2017; Sheng et al., 2018; Xie et al., 2018; Zhang et al., 2018; Sun et al., 2019; Zhang et al., 2019; Abel et al., 2020; Cui et al., 2020; Huang et al., 2020; Zhang et al., 2020). Low mutation rates ($\sim 10^{-9}$) and no stutter/noise peaks are the overwhelming superiorities for InDels, which possess desirable properties of both short tandem repeats (STRs) and single nucleotide polymorphisms (SNPs) (Chakraborty et al., 1999; Weber et al., 2002; Bhangale et al., 2005; Mills et al., 2006; Pakstis et al., 2007; Mullaney et al., 2010; Pakstis et al., 2010; da Costa Francez et al., 2012; Kidd et al., 2012). In addition, the relatively small amplicon sizes of InDels enhance discrimination efficiencies in some dated or highly degraded samples from crime scenes when compared with STRs (Golenberg et al., 1996; Brinkmann et al., 1998; Jin et al., 2019). With the relatively uncomplicated chemical and operational approaches to detect the length variations in contrast to the determination methods of SNPs (Kwok 2002; Amoako et al., 2017; Matsuda 2017), the detection method by capillary electrophoresis (CE) for InDels could be extensively applied in distinct forensic scenarios (Zhao et al., 2018; Chen et al., 2019; Jin et al., 2019; Tao et al., 2019; Huang et al., 2020; Song et al., 2020; Zhang et al., 2020).

At present, the commercial and widely-used InDel kits present some issues, 1) they are not always suitable for East Asian (EAS) ancestry populations; 2) the insufficient utilization for the CE system; and 3) the time-consuming procedures for DNA extractions and/or purifications. The shortcomings for most InDel typing systems limit the promotion of forensic system effectiveness and the extension of forensic scenarios for EAS populations. Therefore, based on the underlying genome-wide data of the EAS populations from the 1,000 Genome Project (1KG) Phase 3 (Genomes Project et al., 2010; Genomes Project et al., 2015) and the engineering fundamental logic for the maximum utilization of multiplex PCR-CE system, a novel 6-dye direct and multiplex PCR-CE-based typing system with relatively short amplicons (<230 bp), consisting of 57 autosomal InDels (A-InDels), 2 Y-chromosomal InDels (Y-InDels), and Amelogenin, was studied to expand application scenarios for forensic purposes in EAS populations, especially for different Chinese populations. Moreover, forensic efficiencies and population genetic analyses of the direct and multiplex PCR-CE-based InDel typing system were further evaluated in 26 globally dispersed populations and the Hainan Li (HNL) group, which is a relatively isolated Chinese group revealed by the previous studies (Fan et al., 2018b; Fan et al. 2019b; Fan et al. 2021a; Wang et al., 2021).

MATERIALS AND METHODS

Sample and Data Collections

Bloodstain samples of 279 unrelated healthy Hainan Li individuals were collected after receiving their informed consents. The experiment was conducted in accordance with the guidelines of humane and ethical research of Xi'an Jiaotong University and Southern Medical University, and warranted by the Ethics Committee of Xi'an Jiaotong University (No. 2019–1231). To evaluate the universal applicability of the novel 6-dye direct and multiplex PCR-CE-based typing system, we collected the population data of global 1KG populations from five continents.

Amplification and CE Detection

The amplification of the novel 6-dye direct and multiplex PCR-CE-based typing system was performed in a single multiplex PCR reaction (25 μl in total) using 10 μl of Reaction Mix (AGCU ScienTech Incorporation, Wuxi, Jiangsu, China), 1 μl of U-Taq Enzyme (AGCU ScienTech Incorporation), 5 μl of InDel 60 Primers (AGCU ScienTech Incorporation), and 9 μl of sdH_2O. PCR cocktail was performed on the GeneAmp PCR System 9700 Thermal Cycler (Thermo Fisher Scientific, Waltham, MA, USA) based on the following parameters: initial denaturation at 95°C for 5 min; then 28 cycles of 94°C for 30 s, 60°C for 1 min, and 62°C for 1 min; and the final extension at 72°C for 10 min. Afterward, 1 μl of PCR production was added to the cocktail of 0.5 μl AGCU Marker SIZ-500 (AGCU ScienTech Incorporation) and 12 μl of HiDi™ formamide (Thermo Fisher Scientific). The mixture was denatured at 95°C for 3 min and then immediately chilled on ice for 3 min. Finally, the product was detected on the 3500xL Genetic Analyzer (Thermo Fisher Scientific) using 36-cm capillary arrays (Thermo Fisher Scientific) with the POP-4® Polymer (Thermo Fisher Scientific). The CE parameters were as follows: 10 s injection at 2 kV, and electrophoresis at 15 kV for 1,400 s at 60°C. Genotyping data for each sample were determined by GeneMapper® ID-X software v1.6 (Thermo Fisher Scientific). Control DNA 9948 and deionized water were used as positive and negative controls, respectively.

Statistical Analysis

Allele frequencies of InDels were calculated using SAS® 9.4 software (SAS Institute Inc., Cary, NC, USA). The forensic parameters, match probability (MP), power of discrimination (PD), polymorphism information content (PIC), power of exclusion (PE), typical paternity index (TPI), and heterozygote (He), were conducted by PowerStats software (Promega, Madison, WI, USA). The Hardy–Weinberg equilibrium (HWE) and linkage disequilibrium (LD) were evaluated by Arlequin v3.5 (Excoffier and Lischer 2010). The mean values and standard deviations of forensic relevant parameters were calculated by SAS® 9.4 software (SAS Institute Inc., Cary, NC, USA).

Population pairwise genetic distances (i.e., F_{ST}) and corresponding p-values between different populations were estimated by analysis of molecular variance (AMOVA) using raw data at Arlequin v3.5 (Excoffier and Lischer 2010). Genetic similarities and differences were further visualized by principal component analysis (PCA)

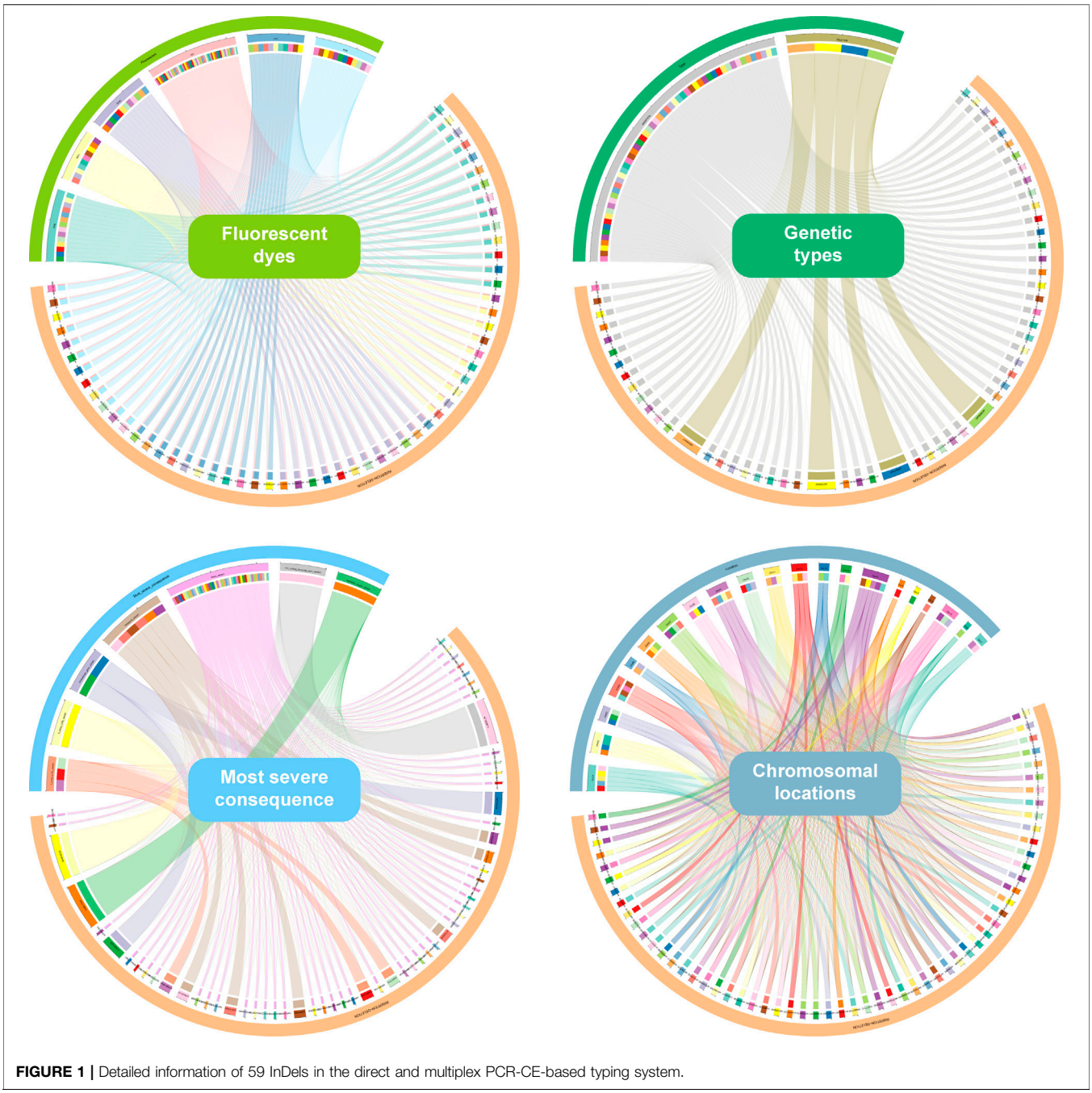


FIGURE 1 | Detailed information of 59 InDels in the direct and multiplex PCR-CE-based typing system.

and multidimensional scaling plot (MDS) using *R* (https://www.r-project.org/) based on insertion allelic frequencies. Additionally, phylogenetic relationships among different populations were depicted in Molecular Evolutionary Genetics Analysis 7.0 (MEGA 7.0) software (Kumar et al., 2016) with neighbor-joining method (Saitou and Nei 1987) and visualized by the Interactive Tree of Life v5 (iTOL) (Letunic and Bork 2019). Other high-quality figures all used *R* to visualize.

RESULTS AND DISCUSSION

Details of the Novel Typing System and Distinct A-InDel Sets

Supplementary Table S1 presents the detailed InDel information of the 6-dye direct and multiplex PCR-CE-based typing system. All genetic markers are autosomal and Y-chromosomal biallelic variations of InDels with the minimum allele frequency (MAF) ≥0.25 in most EAS populations (CHS and CHB in particular). A

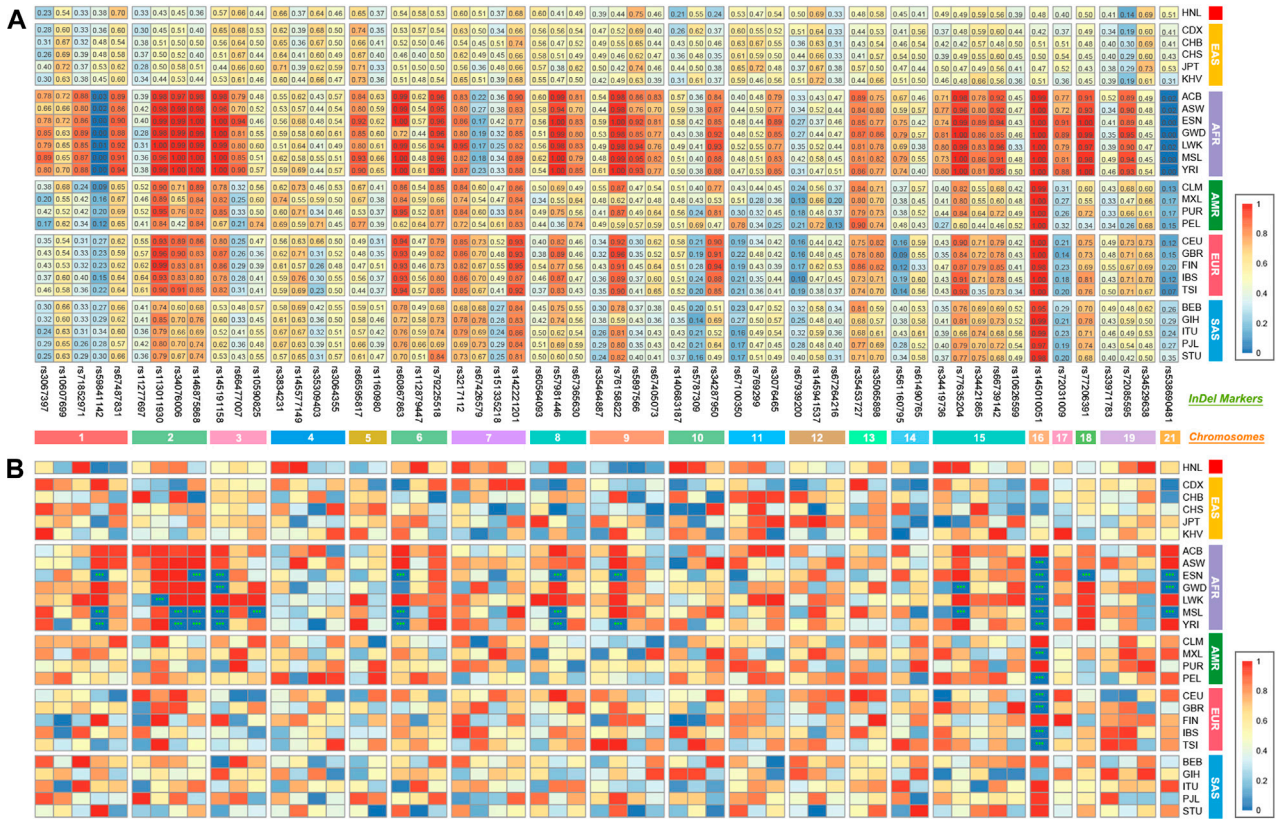


FIGURE 2 | Insertion allelic frequencies and Hardy–Weinberg equilibrium (HWE) tests for EAIS (57 A-InDels) in the 1KG populations and Hainan Li group. **(A)** Heatmap of insertion allelic frequencies; **(B)** results of HWE tests. (***, $p < 0.05/57 \approx 0.0009$ after Bonferroni correction. EAIS, East Asia-based A-InDel set, including 57 A-InDels.)

total of six fluorescent dyes were applied to the direct typing system. The SIZ dye marked the internal reference item, and the HEX, LYN, SUM, PUR, and FAM dyes labeled 10, 11, 12, 13, and 14 InDels in accordance with the amplicon sizes, respectively (**Figure 1** and **Supplementary Table S1**). Besides, the insertion/deletion fragment lengths ranged from 0 to 12 bp, and the amplicon sizes spanned from 76 to 226 bp, respectively. With the consideration of short amplicons for more forensic application scenarios, further validation and evaluation of forensic efficiencies, such as the ability to analyze dated/degraded samples, sensitivity, specificity, stability, and so on, should be performed in our subsequent study. In addition, the 59 InDels were distributed over 20 autosomes and Y chromosome. Of which, 45 InDels were located at the intron variant regions, 6 InDels were located at the intergenic variant regions, 3 InDels were located at the 3′ UTR variant regions, 2 InDels were located at the downstream gene variant regions, 1 InDel was located at the 5′ UTR variant region, 1 InDel was located at the regulatory region variant region, and 1 InDel was located at the non-coding transcript exon variant region (**Figure 1**). Physical distances of pairwise InDels located on the same chromosomes indicated that most pairwise InDels on the same chromosomes were more than 10 Mb apart except for the five pairs (rs3067397 and rs10607699, ~6.3 Mb; rs35464887 and rs76158822, ~4.1 Mb; rs76158822 and rs5897566, ~9.3 Mb; rs34419736 and rs77635204, ~8.8 Mb; and rs72085595 and rs34529638, ~5.9 Mb), but *p*-values (after Bonferroni corrections) of LD analyses (data not shown) were not statistically significant in the 1KG populations and Hainan Li group.

The direct and multiplex PCR-CE typing system (57 A-InDels, 2 Y-InDels, and Amelogenin) was studied based on the genome-wide data from the EAS populations. Thus, the East Asia-based A-InDel set (EAIS, including 57 A-InDels) of the typing system would be performing well for forensic purposes in the EAS populations. The results of HWE tests for 57 A-InDels in the 1KG populations and Hainan Li group are presented in **Figure 2B** and **Supplementary Table S3**. An overwhelming majority of A-InDel loci conformed to HWE in the 1KG populations and Hainan Li group after Bonferroni correction ($0.05/57 = 0.0009$), while 13 A-InDels (rs59841142, rs113011930, rs34076006, rs146875868, rs145191158, rs10590825, rs60867863, rs57981446, rs76158822, rs77635204, rs145010051, rs77206391, and rs538690481) failed to pass the HWE tests, which mainly concentrated on African (AFR) ancestry populations. Therefore, we determined Universal A-InDel set (UAIS, including 44

TABLE 1 | Comparisons of forensic relevant parameters in East Asia-based autosomal insertion/deletion set (EAIS) and universal autosomal insertion/deletion set (UAIS) for the 1,000 Genome Project (1KG) populations and Hainan Li group.

Population		n	Match probability (MP)		Power of discrimination (PD)		Polymorphism information content (PIC)		Power of exclusion (PE)		Typical paternity index (TPI)		Heterozygote (He)	
			EAIS	UAIS	EAIS	UAIS	EAIS	UAIS	EAIS	UAIS	EAIS	UAIS	EAIS	UAIS
AFR	ACB	96	0.5699 ± 0.2096	0.4805 ± 0.1218	0.4301 ± 0.2096	0.5195 ± 0.1218	0.2478 ± 0.1275	0.3005 ± 0.0792	0.0940 ± 0.0753	0.1166 ± 0.0657	0.7722 ± 0.1898	0.8340 ± 0.1570	0.3098 ± 0.1773	0.3771 ± 0.1262
	ASW	61	0.5472 ± 0.2079	0.4549 ± 0.1049	0.4528 ± 0.2079	0.5451 ± 0.1049	0.2614 ± 0.1270	0.3167 ± 0.0703	0.1032 ± 0.0752	0.1294 ± 0.0624	0.7954 ± 0.1891	0.8657 ± 0.1472	0.3310 ± 0.1736	0.4042 ± 0.1092
	ESN	99	0.6112 ± 0.2368	0.5123 ± 0.1457	0.3888 ± 0.2368	0.4877 ± 0.1457	0.2256 ± 0.1429	0.2831 ± 0.0953	0.0916 ± 0.0849	0.1152 ± 0.0801	0.7616 ± 0.2169	0.8291 ± 0.1933	0.2885 ± 0.1987	0.3627 ± 0.1507
	GWD	113	0.5871 ± 0.2186	0.5005 ± 0.1283	0.4129 ± 0.2186	0.4995 ± 0.1283	0.2373 ± 0.1323	0.2875 ± 0.0855	0.0889 ± 0.0778	0.1093 ± 0.0726	0.7588 ± 0.1966	0.8166 ± 0.1728	0.2953 ± 0.1810	0.3598 ± 0.1339
	LWK	99	0.5935 ± 0.2277	0.4980 ± 0.1417	0.4065 ± 0.2277	0.5020 ± 0.1417	0.2346 ± 0.1378	0.2902 ± 0.0919	0.0926 ± 0.0793	0.1158 ± 0.0717	0.7662 ± 0.2019	0.8314 ± 0.1725	0.2988 ± 0.1891	0.3705 ± 0.1385
	MSL	85	0.6057 ± 0.2293	0.5060 ± 0.1408	0.3943 ± 0.2293	0.4940 ± 0.1408	0.2296 ± 0.1406	0.2859 ± 0.0938	0.0955 ± 0.0954	0.1140 ± 0.0813	0.7738 ± 0.2438	0.8272 ± 0.1943	0.2941 ± 0.1992	0.3620 ± 0.1458
	YRI	108	0.6111 ± 0.2377	0.5123 ± 0.1503	0.3901 ± 0.2386	0.4893 ± 0.1511	0.2246 ± 0.1429	0.2816 ± 0.0972	0.0847 ± 0.0786	0.1055 ± 0.0733	0.7454 ± 0.2014	0.8060 ± 0.1767	0.2787 ± 0.1914	0.3486 ± 0.1444
AMR	CLM	94	0.4545 ± 0.1186	0.4155 ± 0.0614	0.5455 ± 0.1186	0.5845 ± 0.0614	0.3194 ± 0.0765	0.3448 ± 0.0433	0.1340 ± 0.0685	0.1530 ± 0.0586	0.8753 ± 0.1638	0.9218 ± 0.1364	0.4062 ± 0.1238	0.4453 ± 0.0848
	MXL	64	0.4452 ± 0.1045	0.4201 ± 0.0591	0.5548 ± 0.1045	0.5799 ± 0.0591	0.3271 ± 0.0681	0.3450 ± 0.0405	0.1467 ± 0.0693	0.1622 ± 0.0634	0.9062 ± 0.1655	0.9439 ± 0.1493	0.4285 ± 0.1127	0.4570 ± 0.0859
	PEL	85	0.4616 ± 0.1216	0.4207 ± 0.0559	0.5384 ± 0.1216	0.5793 ± 0.0559	0.3151 ± 0.0772	0.3418 ± 0.0403	0.1313 ± 0.0657	0.1528 ± 0.0539	0.8685 ± 0.1580	0.9212 ± 0.1253	0.4022 ± 0.1250	0.4465 ± 0.0805
	PUR	104	0.4491 ± 0.1066	0.4309 ± 0.0705	0.5509 ± 0.1066	0.5691 ± 0.0705	0.3202 ± 0.0681	0.3309 ± 0.0485	0.1286 ± 0.0573	0.1331 ± 0.0524	0.8634 ± 0.1364	0.8756 ± 0.1217	0.4045 ± 0.1063	0.4174 ± 0.0845
EAS	CDX	93	0.3960 ± 0.0351	0.4011 ± 0.0341	0.6040 ± 0.0351	0.5989 ± 0.0341	0.3589 ± 0.0201	0.3559 ± 0.0216	0.1639 ± 0.0558	0.1628 ± 0.0566	0.9480 ± 0.1308	0.9453 ± 0.1323	0.4631 ± 0.0701	0.4611 ± 0.0721
	CHB	103	0.3890 ± 0.0305	0.3931 ± 0.0302	0.6110 ± 0.0305	0.6069 ± 0.0302	0.3641 ± 0.0145	0.3614 ± 0.0154	0.1720 ± 0.0451	0.1703 ± 0.0435	0.9663 ± 0.1059	0.9622 ± 0.1014	0.4769 ± 0.0520	0.4751 ± 0.0508
	CHS	105	0.4022 ± 0.0308	0.4027 ± 0.0322	0.5978 ± 0.0308	0.5973 ± 0.0322	0.3646 ± 0.0132	0.3621 ± 0.0140	0.1959 ± 0.0560	0.1883 ± 0.0571	1.0223 ± 0.1330	1.0049 ± 0.1351	0.5029 ± 0.0622	0.4939 ± 0.0640
	JPT	104	0.3948 ± 0.0289	0.3961 ± 0.0296	0.6052 ± 0.0289	0.6039 ± 0.0296	0.3607 ± 0.0174	0.3575 ± 0.0185	0.1700 ± 0.0507	0.1611 ± 0.0491	0.9618 ± 0.1182	0.9413 ± 0.1143	0.4728 ± 0.0598	0.4618 ± 0.0589
	KHV	99	0.4005 ± 0.0331	0.4050 ± 0.0344	0.5995 ± 0.0331	0.5950 ± 0.0344	0.3612 ± 0.0191	0.3584 ± 0.0202	0.1828 ± 0.0518	0.1805 ± 0.0562	0.9914 ± 0.1225	0.9867 ± 0.1333	0.4884 ± 0.0600	0.4846 ± 0.0650
	HNL	279	0.3986 ± 0.0409	0.4042 ± 0.0440	0.6014 ± 0.0409	0.5958 ± 0.0440	0.3563 ± 0.0291	0.3522 ± 0.0318	0.1639 ± 0.0372	0.1597 ± 0.0399	0.9463 ± 0.0856	0.9366 ± 0.0920	0.4666 ± 0.0560	0.4602 ± 0.0610
EUR	CEU	99	0.4903 ± 0.1374	0.4474 ± 0.0903	0.5097 ± 0.1374	0.5526 ± 0.0903	0.2957 ± 0.0881	0.3241 ± 0.0606	0.1183 ± 0.0722	0.1370 ± 0.0674	0.8375 ± 0.1746	0.8837 ± 0.1599	0.3754 ± 0.1365	0.4147 ± 0.1109
	GBR	91	0.4934 ± 0.1491	0.4454 ± 0.1003	0.5066 ± 0.1491	0.5546 ± 0.1003	0.2922 ± 0.0938	0.3222 ± 0.0674	0.1122 ± 0.0668	0.1309 ± 0.0599	0.8224 ± 0.1625	0.8693 ± 0.1413	0.3656 ± 0.1370	0.4082 ± 0.1045
	FIN	99	0.4897 ± 0.1430	0.4522 ± 0.1064	0.5103 ± 0.1430	0.5478 ± 0.1064	0.2986 ± 0.0920	0.3236 ± 0.0713	0.1256 ± 0.0795	0.1453 ± 0.0754	0.8543 ± 0.1924	0.9027 ± 0.1796	0.3830 ± 0.1460	0.4224 ± 0.1232
	IBS	107	0.4825 ± 0.1317	0.4381 ± 0.0824	0.5175 ± 0.1317	0.5619 ± 0.0824	0.2995 ± 0.0842	0.3288 ± 0.0559	0.1173 ± 0.0667	0.1384 ± 0.0590	0.8353 ± 0.1605	0.8870 ± 0.1384	0.3771 ± 0.1296	0.4212 ± 0.0990
	TSI	107	0.4859 ± 0.1398	0.4356 ± 0.0830	0.5141 ± 0.1398	0.5644 ± 0.0830	0.2971 ± 0.0878	0.3285 ± 0.0551	0.1157 ± 0.0630	0.1333 ± 0.0519	0.8312 ± 0.1530	0.8754 ± 0.1218	0.3748 ± 0.1310	0.4163 ± 0.0914
SAS	BEB	86	0.4245 ± 0.0704	0.4093 ± 0.0444	0.5755 ± 0.0704	0.5907 ± 0.0444	0.3394 ± 0.0472	0.3493 ± 0.0317	0.1520 ± 0.0556	0.1580 ± 0.0497	0.9194 ± 0.1306	0.9335 ± 0.1147	0.4447 ± 0.0837	0.4561 ± 0.0681
	GIH	103	0.4335 ± 0.0917	0.4158 ± 0.0528	0.5665 ± 0.0917	0.5842 ± 0.0528	0.3328 ± 0.0580	0.3449 ± 0.0372	0.1426 ± 0.0590	0.1522 ± 0.0565	0.8967 ± 0.1400	0.9201 ± 0.1313	0.4274 ± 0.1001	0.4451 ± 0.0822
	ITU	102	0.4295 ± 0.0862	0.4169 ± 0.0532	0.5705 ± 0.0862	0.5831 ± 0.0532	0.3345 ± 0.0563	0.3429 ± 0.0392	0.1417 ± 0.0568	0.1487 ± 0.0549	0.8948 ± 0.1337	0.9118 ± 0.1267	0.4274 ± 0.0953	0.4407 ± 0.0796
	PJL	96	0.4305 ± 0.0818	0.4146 ± 0.0513	0.5695 ± 0.0818	0.5854 ± 0.0513	0.3312 ± 0.0526	0.3421 ± 0.0362	0.1330 ± 0.0484	0.1418 ± 0.0463	0.8748 ± 0.1138	0.8959 ± 0.1066	0.4176 ± 0.0855	0.4335 ± 0.0714
	STU	102	0.4250 ± 0.0907	0.4124 ± 0.0626	0.5750 ± 0.0907	0.5876 ± 0.0626	0.3326 ± 0.0563	0.3405 ± 0.0408	0.1242 ± 0.0514	0.1270 ± 0.0531	0.8549 ± 0.1225	0.8621 ± 0.1244	0.4028 ± 0.0893	0.4086 ± 0.0808

TABLE 2 | Comparisons of forensic system efficiencies in EAIS and UAIS for the 1KG populations and Hainan Li group.

Population		n	CMP		CPD		CPE	
			EAIS	UAIS	EAIS	UAIS	EAIS	UAIS
AFR	ACB	96	3.50E−16	2.82E−15	1−3.50E−16	1−2.82E−15	0.997060	0.996212
	ASW	61	3.28E−17	3.32E−16	1−3.28E−17	1−3.32E−16	0.998356	0.997997
	ESN	99	1.10E−14	3.25E−14	1−1.10E−14	1−3.25E−14	0.996773	0.996199
	GWD	113	1.74E−15	1.56E−14	1−1.74E−15	1−1.56E−14	0.996016	0.994716
	LWK	99	2.44E−15	1.03E−14	1−2.44E−15	1−1.03E−14	0.996858	0.996168
	MSL	85	8.00E−15	2.01E−14	1−8.00E−15	1−2.01E−14	0.997689	0.995993
	YRI	108	8.96E−15	2.55E−14	1−8.96E−15	1−2.55E−14	0.994830	0.993634
AMR	CLM	94	6.44E−21	1.08E−17	1−6.44E−21	1−1.08E−17	0.999771	0.999396
	MXL	64	2.80E−21	1.81E−17	1−2.80E−21	1−1.81E−17	0.999902	0.999636
	PEL	85	1.59E−20	2.00E−17	1−1.59E−20	1−2.00E−17	0.999721	0.999380
	PUR	104	4.54E−21	4.89E−17	1−4.54E−21	1−4.89E−17	0.999654	0.998282
EAS	CDX	93	9.42E−24	2.99E−18	1−9.42E−24	1−2.99E−18	0.999968	0.999637
	CHB	103	3.59E−24	1.27E−18	1−3.59E−24	1−1.27E−18	0.999981	0.999746
	CHS	105	2.40E−23	3.63E−18	1−2.40E−23	1−3.63E−18	0.999997	0.999908
	JPT	104	8.51E−24	1.78E−18	1−8.51E−24	1−1.78E−18	0.999978	0.999593
	KHV	99	1.85E−23	4.59E−18	1−1.85E−23	1−4.59E−18	0.999991	0.999859
	HNL	279	1.32E−23	3.93E−18	1−1.32E−23	1−3.93E−18	0.999965	0.999549
EUR	CEU	99	3.42E−19	1.96E−16	1−3.42E−19	1−1.96E−16	0.999371	0.998665
	GBR	91	3.69E−19	1.38E−16	1−3.69E−19	1−1.38E−16	0.999037	0.998123
	FIN	99	2.90E−19	2.52E−16	1−2.90E−19	1−2.52E−16	0.999627	0.999160
	IBS	107	1.54E−19	8.72E−17	1−1.54E−19	1−8.72E−17	0.999307	0.998716
	TSI	107	1.92E−19	6.68E−17	1−1.92E−19	1−6.68E−17	0.999220	0.998297
SAS	BEB	86	3.34E−22	6.67E−18	1−3.34E−22	1−6.67E−18	0.999927	0.999521
	GIH	103	8.19E−22	1.23E−17	1−8.19E−22	1−1.23E−17	0.999865	0.999368
	ITU	102	5.27E−22	1.39E−17	1−5.27E−22	1−1.39E−17	0.999855	0.999234
	PJL	96	6.30E−22	1.11E−17	1−6.29E−22	1−1.11E−17	0.999731	0.998879
	STU	102	2.46E−22	7.50E−18	1−2.46E−22	1−7.50E−18	0.999529	0.997665

InDels which excluded the 13 A-InDels unconfirmed to HWE) to evaluate the forensic efficiencies for 26 universal 1KG populations and the Hainan Li group.

Forensic Parameter Evaluations of UAIS (44 A-InDels) in the 1,000 Genomes Project Populations and Hainan Li Group

Allelic frequencies of all 57 A-InDel loci for the 1 KG populations and Hainan Li group are illustrated in **Figure 2A** and **Supplementary Table S4**. For UAIS, including a total of 44 A-InDels conformed by Hardy–Weinberg equilibrium and linkage equilibrium, the insertion allelic frequencies ranged from 0.0934 (GBR, rs561160795) to 0.9861 (YRI, rs79225518). The detailed forensic relevant parameters (MP, PD, PIC, PE, TPI, and He) of UAIS in the 1KG populations and Hainan Li group are calculated and summarized in **Supplementary Tables S5–S10**. In addition, the mean values and standard deviations for all forensic-related parameters in 26 different reference populations from the 1KG and Hainan Li group are shown in **Table 1**. The MP values (Mean ± Standard Deviation) of UAIS ranged from 0.3931 ± 0.0302 (CHB) to 0.5123 ± 0.1457 (ESN) (mean MP: 0.4386 ± 0.0951). The PIC values spanned from 0.2816 ± 0.0972 (YRI) to 0.3621 ± 0.0140 (CHS) (mean PIC: 0.3297 ± 0.0633). The TPI values varied from 0.8060 ± 0.1767 (YRI) to 1.0049 ± 0.1351 (CHS) (mean TPI: 0.8959 ± 0.1523). The average PE value was 0.1423 ± 0.0643 with a range from 0.1055 ± 0.0733 (YRI) to 0.1883 ± 0.0571 (CHS). For He, the mean value was 0.4243 ± 0.1071 ranging from 0.3486 ± 0.1444 (YRI) to 0.4939 ± 0.0640 (CHS). What is more, the cumulative match probability (CMP), cumulative power of discrimination (CPD), and cumulative power of exclusion (CPE) values of UAIS for the 1KG populations and Hainan Li group are illustrated in **Table 2**. The CPD values ranged from $1–1.03 \times 10^{-14}$ (LWK) to $1–1.27 \times 10^{-18}$ (CHB), and the CPE varied from 0.993634 (YRI) to 0.999908 (CHS) in the 1KG populations and Hainan Li group.

In total, for UAIS which showed no evidence of deviation from HWE and LDs in both the 1KG populations and Hainan Li group, the forensic-related parameters were distributed relatively balanced (**Figure 3**), revealing that the UAIS had considerable potential in the field of forensic human identification for universal populations. The 44 A-InDels of UAIS possessed relatively reasonable genetic information (PIC > 0.25) (Botstein et al., 1980), and the UAIS with the CMP range of $1.27 \times 10^{-18}$to 1.03×10^{-14} for 27 universal human populations satisfied the requirements for forensic human identification (10^{-15}–10^{-14}) (Pereira et al., 2009), which indicated that the UAIS could be considered as a powerful tool for human identification. Compared with the CPE provided by the common STR panels (Fan et al., 2019a; Fan et al., 2019b; Li et al., 2020), CPE for UAIS has outclassed 0.993634–0.999908. Therefore, the UAIS could supply additional information for the paternity tests.

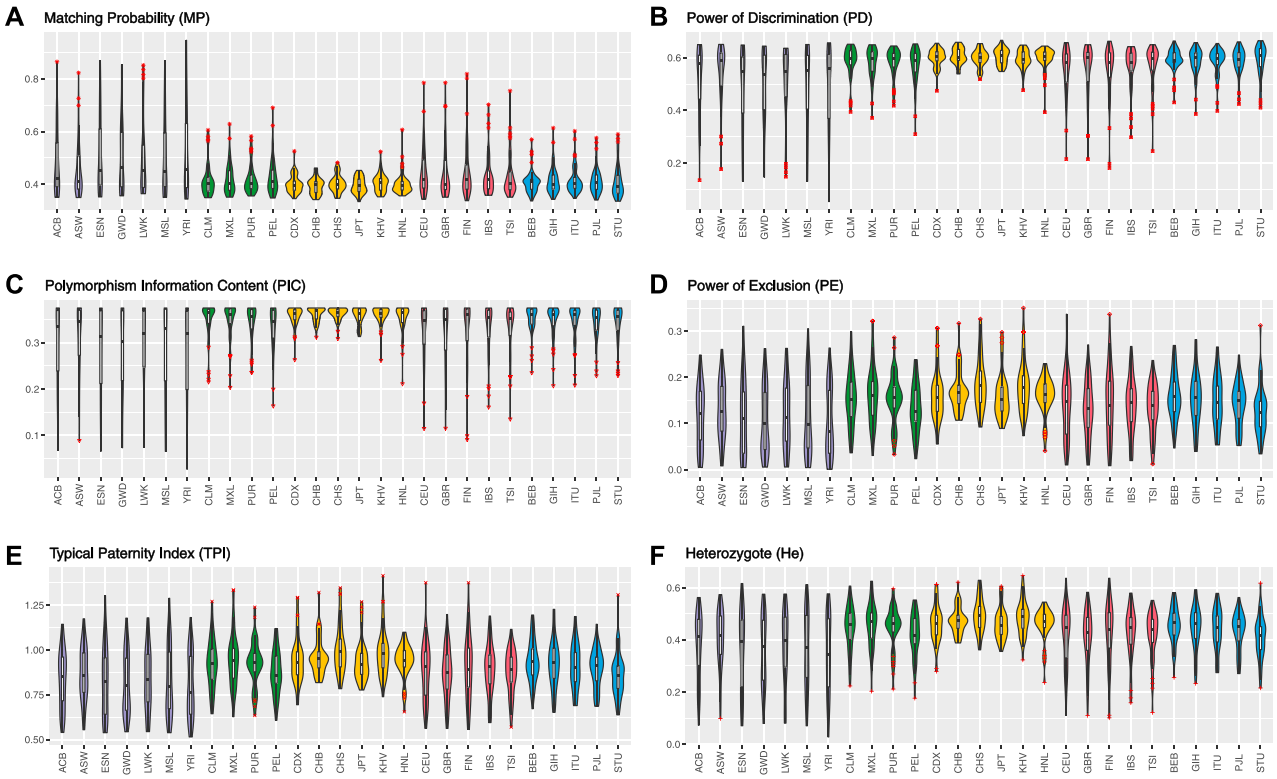


FIGURE 3 | Forensic-related parameters of UAIS (44 A-InDels) for the 1KG populations and Hainan Li group. **(A)** Matching probability; **(B)** power of discrimination; **(C)** polymorphism information content; **(D)** power of exclusion; **(E)** typical paternity index; **(F)** heterozygote.

Forensic Parameter Evaluations of EAIS (57 A-InDels) in the East Asian Populations and Hainan Li Group

The Hainan Li, inhabiting in the south of Hainan island, is a relatively isolated minority group in China, which is beneficial to clarify the exquisite population structure and develop specific genetic markers for subpopulations in the forensic genetic field (Liu et al., 2020b; Fan et al., 2021a). Hence, a total of 279 healthy Hainan Li individuals were collected for forensic evaluations of 57 A-InDel loci in EAIS. The insertion allelic frequencies of the Hainan Li group are demonstrated in **Figures 2A** and **4A** and **Supplementary Table S4**, which were distributed between 0.2079 (rs140683187) and 0.7491 (rs5897566), except for rs72085595 (0.1398). The forensic parameters (MP, PD, PIC, PE, TPI, and He) of the 57 A-InDels in the Hainan Li group are shown in **Supplementary Tables S5–S10** and **Figure 4B**. The MP values of the Hainan Li group ranged from 0.3557 (rs76158822) to 0.6069 (rs72085595) (mean MP: 0.3986 ± 0.0409). The PIC values spanned from 0.2116 (rs72085595) to 0.3750 (rs67939200, rs34419736, rs77206391, and rs538690481) (mean PIC: 0.3563 ± 0.0291). The TPI values varied from 0.6549 (rs72085595) to 1.0984 (rs67405073) (mean TPI: 0.9463 ± 0.0856). The PE values ranged from 0.0405 (rs72085595) to 0.2298 (rs67405073) with an average of 0.1639 ± 0.0372. The He values varied from 0.2366 (rs72085595) to 0.5448 (rs67405073) (mean He: 0.4666 ± 0.0560). Moreover, the CPE and CPD of the 57 A-InDels in the Hainan Li group was 0.999965 and $1–1.32 \times 10^{-23}$, which demonstrated that the EAIS have good performances for individual identification and paternity test in the Hainan Li group (**Table 2**). What is more, compared with the results of 47 A-InDels in 216 Hainan Li (Liu et al., 2020a) and 30 A-InDels in 207 Hainan Li (Liu et al., 2019), the majority of 57 A-InDels showed more balanced frequency distributions in the same population (**Figure 4A**). With the number of analyzed A-InDels increased, the CPD and CPE also increased, while the CMP decreased in the Hainan Li group (**Table 3**).

As shown in **Figure 5** and **Table 1**, the forensic-associated parameters of CDX, CHB, CHS, JPT, and KHV populations, and the Hainan Li group are illustrated. The MP values ranged from 0.3890 ± 0.0305 (CHB) to 0.4022 ± 0.0308 (CHS). The PIC values spanned from 0.3563 ± 0.0291 (HNL) to 0.3646 ± 0.0132 (CHS). The TPI values varied from 0.9463 ± 0.0856 (HNL) to 1.0223 ± 0.1330 (CHS). The PE values ranged from 0.1639 ± 0.0558 (CDX) to 0.1959 ± 0.0560 (CHS). The He values spanned from 0.4631 ± 0.0701 (CDX) to 0.5029 ± 0.0622 (CHS). In addition, the CPE varied from 0.999965 (HNL) to 0.999997 (CHS), and the CPD ranged from $1–1.32 \times 10^{-23}$ (HNL) to $1–9.42 \times 10^{-24}$ (CHB), respectively. The results revealed that the EAIS has sufficient system effectiveness for human identification and kinship analysis in the EAS populations.

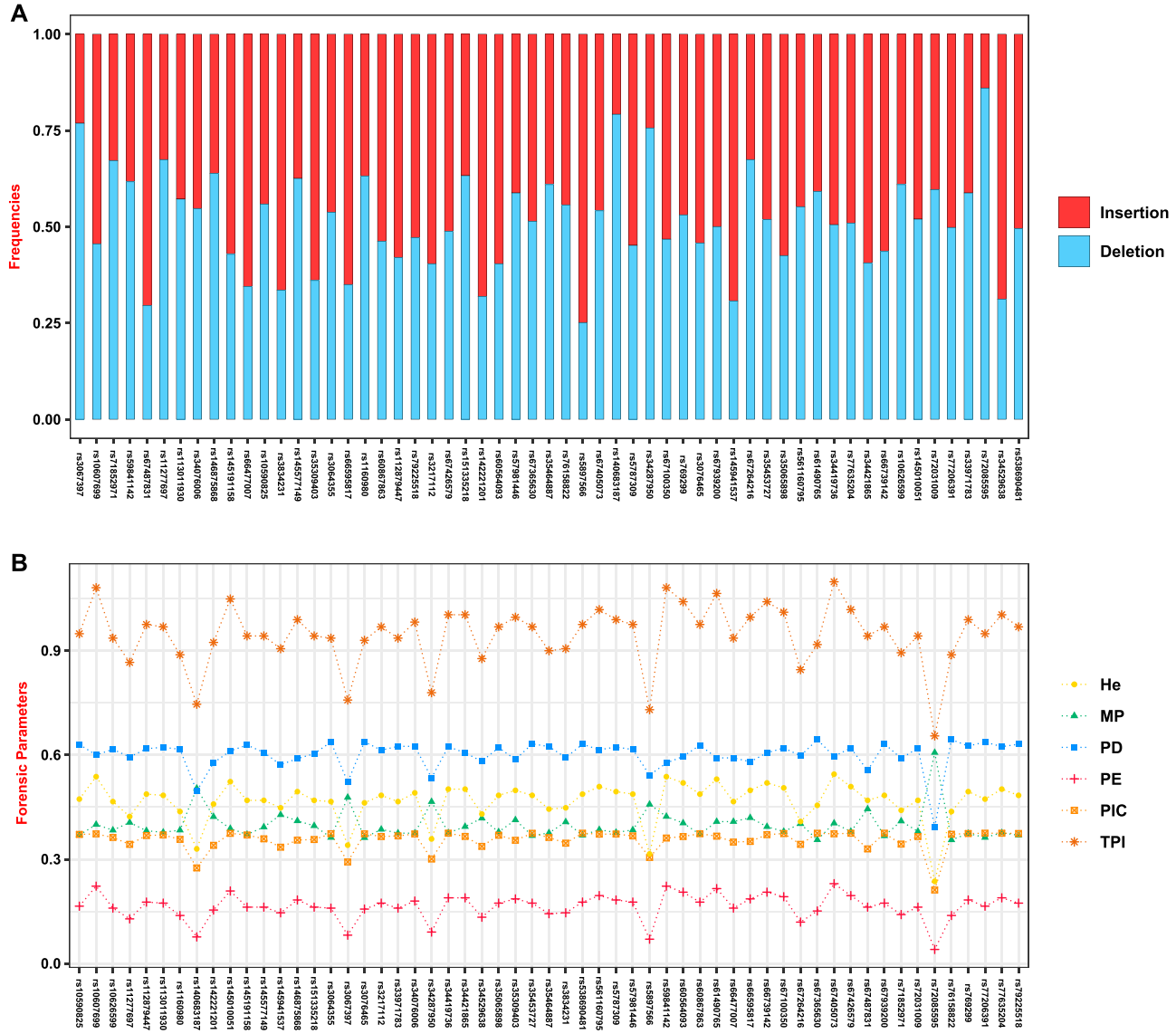


FIGURE 4 | Insertion/deletion allelic frequencies and forensic-associated parameters for EAIS (57 A-InDels) in the Hainan Li group. **(A)** Insertion/deletion allelic frequencies; **(B)** forensic parameters (MP, match probability; PD, power of discrimination; PIC, polymorphism information content; PE, power of exclusion; TPI, typical paternity index; He, heterozygote.

TABLE 3 | Comparisons of forensic system efficiencies in different panels with distinct A-InDels for the Hainan Li group (N, number of A-InDel; n, number of population size).

Population	Panel	N	n	CMP	CPD	CPE
Hainan Li	Investigator DIPplex kit	30	207	2.92E−11	1–2.92E−11	0.986100
	AGCU InDel 50 kit	47	216	7.67E−18	1–7.67E−18	0.999283
	UAIS	44	279	3.93E−18	1–3.93E−18	0.999549
	EAIS	57	279	1.32E−23	1–1.32E−23	0.999965

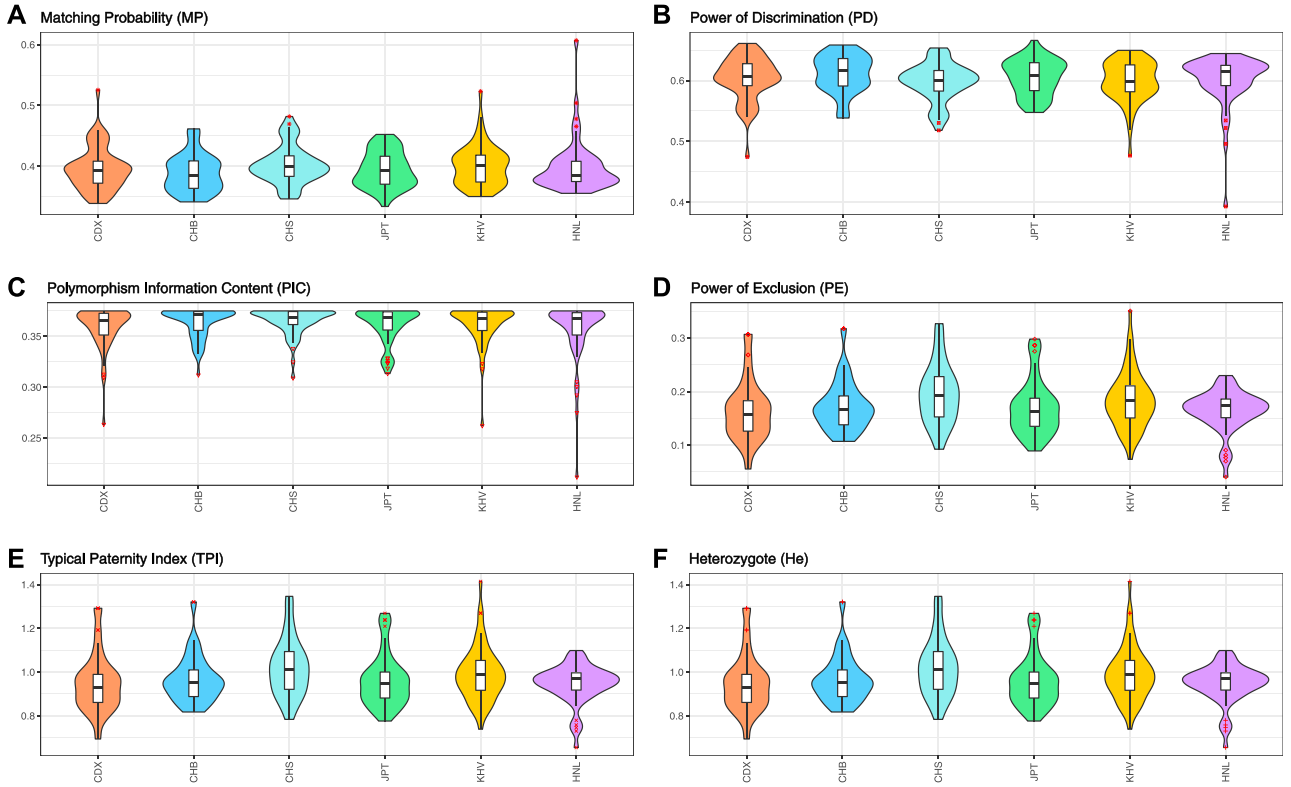


FIGURE 5 | Forensic relevant parameters of EAIS (57 A-InDels) in the East Asian (EAS) populations and Hainan Li group. **(A)** Matching probability; **(B)** power of discrimination; **(C)** polymorphism information content; **(D)** power of exclusion; **(E)** typical paternity index; **(F)** heterozygote.

Population Genetic Analyses Among the Hainan Li Group and 1,000 Genomes Project Populations

To illustrate the genetic landscapes among the 1KG populations and Hainan Li group, the dimensionality reduction analyses (PCA and MDS), which can accelerate the speed of algorithm execution, improve the performance of the analysis model, and reduce the complexity of data at the same time, were conducted based on insertion allelic frequencies of 44 A-InDels, which are illustrated in **Figure 2A** and **Supplementary Table S4**. As shown in **Figures 6B, C**, the first, second, and third components (PC1, PC2, and PC3) accounted for 24.28%, 17.01%, and 7.91% of the total variance observed within these populations, respectively. In the PCA diagrams (**Figures 6B, C**), populations from five different intercontinental ancestries clustered separately, the EAS populations and Hainan Li group clustered together on the upper right. While, the European populations located at the bottom, and the AFR populations distributed on the upper left. In addition, in order to make further confirmation about the genetic relationships between the Hainan Li group and populations from the 1KG conducted by PCA, the Manhattan and Euclidean distance-based MDS were conducted (**Figures 5D, E**), which also depicted the genetic relationships among the Hainan Li group and 1KG populations. The MDS results (**Figures 5D, E**) were in accordance with the genetic patterns of PCAs (**Figures 5B, C**). In brief, the dimensionality reduction analyses (PCA and MDS) made relatively clear distinctions, and the Hainan Li group had the close relationships with the EAS populations.

The results of pairwise F_{ST} and the corresponding p-values between the Hainan Li group and 26 worldwide populations from different continents are listed in **Supplementary Table S11**. The extreme values of F_{ST} were identified at CEU and GBR ($F_{ST} = 0.0001$, $p < 0.0001$), and GWD and PEL ($F_{ST} = 0.2408$, $p < 0.0001$). Phylogenetic relationships between the Hainan Li group and the other 26 reference populations are visualized in the neighbor-joining tree (**Figure 6F**). The EAS populations and Hainan Li group clustered together. For details, the Hainan Li got together with CDX ($F_{ST} = 0.0037$, $p < 0.0001$) and KHV ($F_{ST} = 0.0048$, $p < 0.0001$), and CHB and CHS clustered together with JPT in another inner branch. They all belong to Southeast Asia from the perspective of geography. The pairwise genetic distances indicated by F_{ST} values and the phylogenic relationships based on neighbor-joining tree were consistent with the results of the abovementioned population genetic analyses (PCA and MDS), which manifested that the genetic distances of different populations were consistent with geographic scales in the present study to some degree.

In general, from the perspective of population genetic analyses, compared with paternal Y-STR genetic markers (Fan et al., 2018a; Fan et al., 2018c; Liu et al., 2020b; Ding et al., 2020; Fan et al., 2021b; Fan et al., 2021c; Luo et al., 2021), the novel 6-

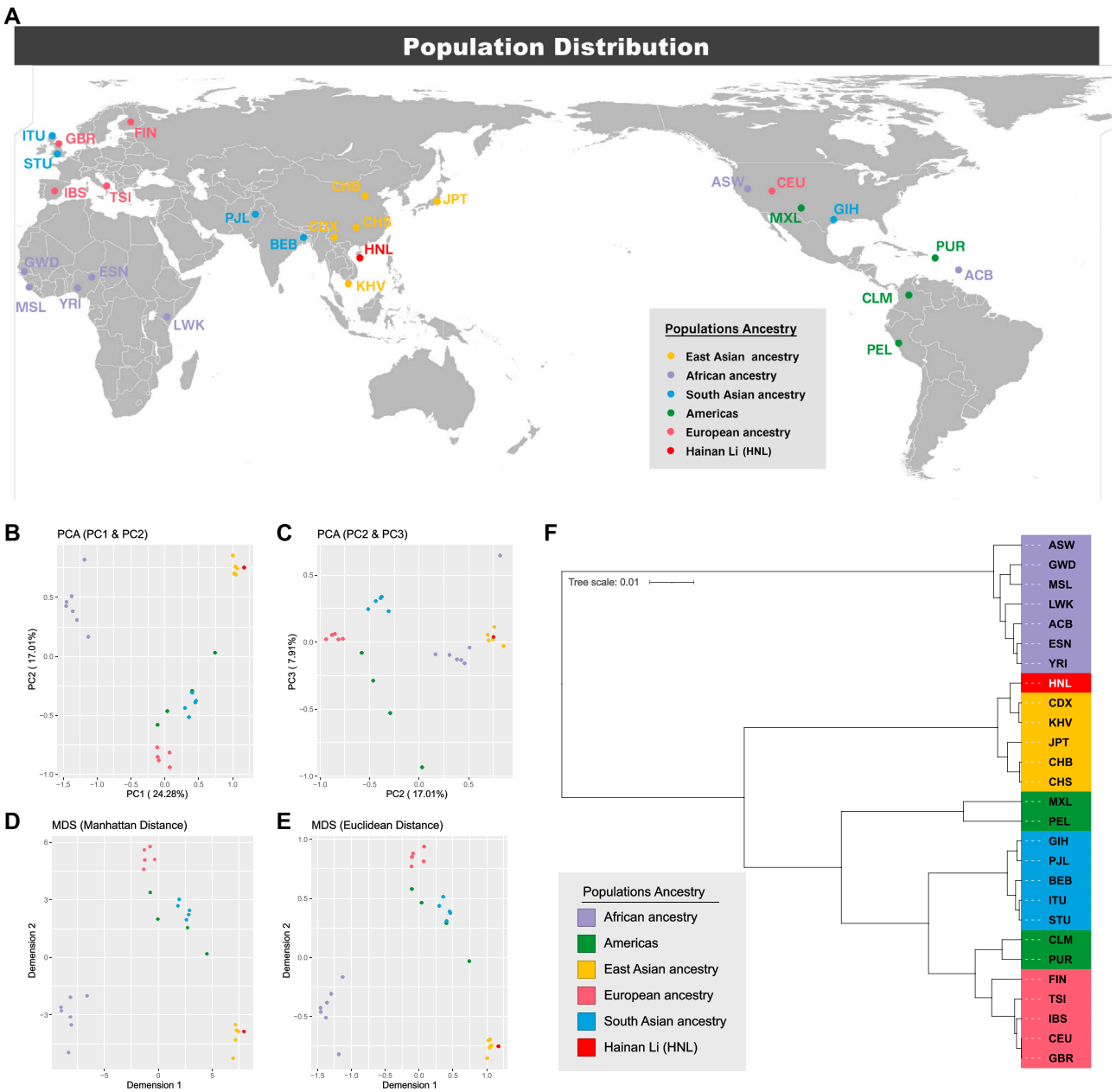


FIGURE 6 | Locations of population distributions and population genetic analyses among the 1KG populations and Hainan Li group. **(A)** Locations of population distributions; **(B)** principal component analysis (PCA) (PC1 and PC2); **(C)** PCA (PC2 and PC3); **(D)** Manhattan-based multidimensional scaling plot (MDS); **(E)** Euclidean-based MDS; **(F)** phylogenetic analysis based on pairwise F_{ST} values.

dye direct and multiplex PCR-CE-based typing system also possessed the ability to differentiate intercontinental populations to a certain extent. The UAIS enabled to make the relatively clear distinctions among populations from five intercontinental ancestries, and the Hainan Li group had the close genetic relationships with EAS populations.

CONCLUSION

In conclusion, the direct and multiplex PCR-CE-based typing system was studied based on genome-wide EAS population data, consisting of 57 A-InDels, 2 Y- InDels, and Amelogenin. We collected two A-InDel sets (EAIS and UAIS) according to the

numbers of A-InDels, which confirmed to HWE and evaluated the forensic system effectiveness for each set from the perspectives of EAS and global 1KG populations, respectively. For UAIS (44 A-InDels), the CPD ranged from $1–1.03 \times 10^{-14}$ to $1–1.27 \times 10^{-18}$, and the CPE varied from 0.993634 to 0.999908. For EAIS (57 A-InDels), the ranges of CPD and CPE values were $1–1.32 \times 10^{-23}$ to $1–9.42 \times 10^{-24}$, and 0.999965–0.999997, respectively. In addition, the CPD and CPE values of EAIS for the Hainan Li group were $1–1.32 \times 10^{-23}$ and 0.999965, respectively. The population genetic analyses clarified the distinctions among the 1KG populations, and the Hainan Li group had close relationships with EAS populations. Consequently, the novel 6-dye direct and multiplex PCR-CE-based typing system should be considered as a reliable and robust tool for human identification and intercontinental population genetics, and supply additional information for kinship analysis in the 1KG populations and Hainan Li group.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of Xi'an Jiaotong University (No. 2019–1231). 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

HF conceptualized the study, performed the formal analysis, wrote the original draft, reviewed and edited the manuscript, and provided the visualization of the study. FW and ZD procured the resources. HF and SL were in charge of the software for the study. YH, SL, YF, and QX performed the investigation. FW conducted the validation. ZD curated the data. PQ and BZ were in charge of the supervision and project administration. FW, ZD, and BZ acquired the funding. All authors have read and agreed to the published version of the manuscript.

FUNDING

This study was supported by the National Undergraduate Innovation and Entrepreneurship Training Program (Nos. 201911810008 and 201911810023) and the Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme (GDUPS, 2017).

ACKNOWLEDGMENTS

We would like to thank all the volunteers who contributed the 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.2021.744645/full#supplementary-material

REFERENCES

Abel, H. J., Larson, D. E., Larson, D. E., Regier, A. A., Chiang, C., Das, I., et al. (2020). Mapping and Characterization of Structural Variation in 17,795 Human Genomes. *Nature* 583, 83–89. doi:10.1038/s41586-020-2371-0

Amoako, K. K., Thomas, M. C., Janzen, T. W., and Goji, N. (2017). Rapid SNP Detection and Genotyping of Bacterial Pathogens by Pyrosequencing. *Methods Mol. Biol.* 1492, 203–220. doi:10.1007/978-1-4939-6442-0_15

Bhangale, T. R., Rieder, M. J., Livingston, R. J., and Nickerson, D. A. (2005). Comprehensive Identification and Characterization of Diallelic Insertion-Deletion Polymorphisms in 330 Human Candidate Genes. *Hum. Mol. Genet.* 14, 59–69. doi:10.1093/hmg/ddi006

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms. *Am. J. Hum. Genet.* 32, 314–331.

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., and Rolf, B. (1998). Mutation Rate in Human Microsatellites: Influence of the Structure and Length of the Tandem Repeat. *Am. J. Hum. Genet.* 62, 1408–1415. doi:10.1086/301869

Bus, M. M., Karas, O., and Allen, M. (2016). Multiplex Pyrosequencing of InDel Markers for Forensic DNA Analysis. *Electrophoresis* 37, 3039–3045. doi:10.1002/elps.201600255

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

Chakraborty, R., Stivers, D. N., Su, B., Zhong, Y., and Budowle, B. (1999). The Utility of Short Tandem Repeat Loci beyond Human Identification: Implications for Development of New DNA Typing Systems. *Electrophoresis* 20, 1682–1696. doi:10.1002/(sici)1522-2683(19990101)20:8<1682:aid-elps1682>3.0.co;2-z

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

Cui, W., Jin, X., Guo, Y., Chen, C., Zhang, W., Kong, T., et al. (2020). Forensic Applicability of Autosomal Insertion/deletion Loci in Chinese Daur Ethnic Group and Genetic Affinity Evaluations between Daur Group and Reference Populations. *Leg. Med.* 47, 101741. doi:10.1016/j.legalmed.2020.101741

da Costa Francez, P. A., Rodrigues, E. M. R., de Velasco, A. M., and dos Santos, S. E. B. (2012). Insertion-deletion Polymorphisms-Utilization on Forensic Analysis. *Int. J. Leg. Med* 126, 491–496. doi:10.1007/s00414-011-0588-z

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.*, 1–8. doi:10.1080/20961790.2020.1810882

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, 564–567. doi:10.1111/j.1755-0998.2010.02847.x

Fan, H., Du, Z., Wang, F., Wang, X., Wen, S.-Q., Wang, L., et al. (2021a). The Forensic Landscape and the Population Genetic Analyses of Hainan Li Based on Massively Parallel Sequencing DNA Profiling. *Int. J. Leg. Med* 135, 1295–1317. doi:10.1007/s00414-021-02590-3

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. Leg. Med* 133, 429–431. doi:10.1007/s00414-018-1828-2

Fan, H., Xie, Q., Li, Y., Wang, L., Wen, S.-Q., and Qiu, P. (2021b). 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

Fan, H., Zeng, Y., Wu, W., Liu, H., Xu, Q., Du, W., et al. (2021c). The Y-STR Landscape of Coastal southeastern Han: Forensic Characteristics, Haplotype Analyses, Mutation Rates, and Population Genetics. *Electrophoresis* 42, 1578–1593. doi:10.1002/elps.202100037

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

Genomes Project, C., 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

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, 68–74. doi:10.1038/nature15393

Golenberg, E. M., Bickel, A., and Weihs, P. (1996). Effect of Highly Fragmented DNA on PCR. *Nucleic Acids Res.* 24, 5026–5033. doi:10.1093/nar/24.24.5026

Huang, Y., Liu, C., Xiao, C., Chen, X., Yi, S., and Huang, D. (2020). Development of a New 32-plex InDels Panel for Forensic Purpose. *Forensic Sci. Int. Genet.* 44, 102171. doi:10.1016/j.fsigen.2019.102171

Jin, X. Y., Wei, Y. Y., Cui, W., Chen, C., Guo, Y. X., Zhang, W. Q., et al. (2019). Development of a Novel Multiplex Polymerase Chain Reaction System for Forensic Individual Identification Using Insertion/deletion Polymorphisms. *Electrophoresis* 40, 1691–1698. doi:10.1002/elps.201800412

Kidd, K. K., Kidd, J. R., Speed, W. C., Fang, R., Furtado, M. R., Hyland, F. C. L., et al. (2012). Expanding Data and Resources for Forensic Use of SNPs in Individual Identification. *Forensic Sci. Int. Genet.* 6, 646–652. doi:10.1016/j.fsigen.2012.02.012

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

Kwok, P.-Y. (2002). SNP Genotyping with Fluorescence Polarization Detection. *Hum. Mutat.* 19, 315–323. doi:10.1002/humu.10058

Letunic, I., and Bork, P. (2019). Interactive Tree of Life (iTOL) V4: Recent Updates and New Developments. *Nucleic Acids Res.* 47, W256–W259. doi:10.1093/nar/gkz239

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., Ye, Z., Wang, Z., Zhou, X., He, G., Mang, M., et al. (2019). Genetic Diversity and Phylogenetic Analysis of Chinese Han and Li Ethnic Populations from Hainan Island by 30 Autosomal Insertion/deletion Polymorphisms. *Forensic Sci. Res.* doi:10.1080/20961790.2019.1672933

Liu, J., Du, W., Wang, M., Liu, C., Wang, S., He, G., et al. (2020a). 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, J., Wang, R., Shi, J., Cheng, X., Hao, T., Guo, J., et al. (2020b). The Construction and Application of a New 17-plex Y-STR System Using Universal Fluorescent PCR. *Int. J. Leg. Med* 134, 2015–2027. doi:10.1007/s00414-020-02291-3

Luo, C., Duan, L., Li, Y., Xie, Q., Wang, L., Ru, K., et al. (2021). Insights from Y-STRs: Forensic Characteristics, Genetic Affinities, and Linguistic Classifications of Guangdong Hakka and She Groups. *Front. Genet.* 12, 676917. doi:10.3389/fgene.2021.676917

Matsuda, K. (2017). PCR-based Detection Methods for Single-Nucleotide Polymorphism or Mutation. *Adv. Clin. Chem.* 80, 45–72. doi:10.1016/bs.acc.2016.11.002

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

Mullaney, J. M., Mills, R. E., Pittard, W. S., and Devine, S. E. (2010). Small Insertions and Deletions (INDELs) in Human Genomes. *Hum. Mol. Genet.* 19, R131–R136. doi:10.1093/hmg/ddq400

Pakstis, A. J., Speed, W. C., Fang, R., Hyland, F. C. L., Furtado, M. R., Kidd, J. R., et al. (2010). SNPs for a Universal Individual Identification Panel. *Hum. Genet.* 127, 315–324. doi:10.1007/s00439-009-0771-1

Pakstis, A. J., Speed, W. C., Kidd, J. R., and Kidd, K. K. (2007). Candidate SNPs for a Universal Individual Identification Panel. *Hum. Genet.* 121, 305–317. doi:10.1007/s00439-007-0342-2

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, 3682–3690. doi:10.1002/elps.200900274

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

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

Song, F., Lang, M., Li, L., Luo, H., and Hou, Y. (2020) Forensic Features and Genetic Background Exploration of a New 47-autosomal InDel Panel in Five Representative Han Populations Residing in Northern China. *Mol. Genet. Genomic Med.* 8, e1224. doi:10.1002/mgg3.1224

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

Sun, K., Yun, L., Zhang, C., Shao, C., Gao, T., Zhao, Z., et al. (2019). Evaluation of 12 Multi-InDel Markers for Forensic Ancestry Prediction in Asian Populations. *Forensic Sci. Int. Genet.* 43, 102155. doi:10.1016/j.fsigen.2019.102155

Tao, R., Zhang, J., Sheng, X., Zhang, J., Yang, Z., Chen, C., et al. (2019). Development and Validation of a Multiplex Insertion/deletion Marker Panel, SifaInDel 45plex System. *Forensic Sci. Int. Genet.* 41, 128–136. doi:10.1016/j.fsigen.2019.04.008

Wang, F., Du, Z., Han, B., Cao, S., Fu, F., Luo, Z., et al. (2021). Genetic Diversity, Forensic Characteristics and Phylogenetic Analysis of the Qiongzhong Aborigines Residing in the Tropical Rainforests of Hainan Island via 19 Autosomal STRs. *Ann. Hum. Biol.* 48, 335–342. doi:10.1080/03014460.2021.1951352

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

Xie, T., Guo, Y., Chen, L., Fang, Y., Tai, Y., Zhou, Y., et al. (2018). A Set of Autosomal Multiple InDel Markers for Forensic Application and Population Genetic Analysis in the Chinese Xinjiang Hui Group. *Forensic Sci. Int. Genet.* 35, 1–8. doi:10.1016/j.fsigen.2018.03.007

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, S., Zhu, Q., Chen, X., Zhao, Y., Zhao, X., Yang, Y., et al. (2018). Forensic Applicability of Multi-Allelic InDels with Mononucleotide Homopolymer Structures. *Electrophoresis* 39, 2136–2143. doi:10.1002/elps.201700468

Zhang, W., Jin, X., Wang, Y., Kong, T., Cui, W., Chen, C., et al. (2020). Genetic Polymorphisms and Forensic Efficiencies of a Set of Novel Autosomal InDel

Markers in a Chinese Mongolian Group. *Biomed. Res. Int.* 2020, 3925189. doi:10.1155/2020/3925189

Zhao, X., Chen, X., Zhao, Y., Zhang, S., Gao, Z., Yang, Y., et al. (2018). Construction and Forensic Genetic Characterization of 11 Autosomal Haplotypes Consisting of 22 Tri-allelic Indels. *Forensic Sci. Int. Genet.* 34, 71–80. doi:10.1016/j.fsigen.2018.02.001

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 MW declared a past coauthorship with the authors HF, FW, ZD, YF, PQ, and BZ at the time of review.

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.

GLOSSARY

InDel insertion/deletion
STRs short tandem repeats
SNPs single nucleotide polymorphisms
1KG 1,000 Genomes Project
AFR African
AMR Americas
EAS East Asian
EUR European
SAS South Asian
HNL Hainan Li
ESN Esan in Nigeria
GWD Gambian in Western Division, Mandinka
LWK, Luhya in Webuye, Kenya
MSL Mende in Sierra Leone
YRI Yoruba in Ibadan, Nigeria
ACB African Caribbean in Barbados
ASW people with African Ancestry in Southwest United States
CLM Colombians in Medellin, Colombia
MXL people with Mexican Ancestry in Los Angeles, CA, United States
PEL Peruvians in Lima, Peru
PUR Puerto Ricans in Puerto Rico
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
CEU Utah residents (CEPH) with Northern and Western European ancestry
GBR British in England and Scotland
FIN Finnish in Finland
IBS Iberian Populations in Spain
TSI Toscani in Italia
BEB Bengali in Bangladesh
GIH Gujarati Indians in Houston, TX, United States
ITU Indian Telugu in the United Kingdom
PJL Punjabi in Lahore, Pakistan
STU Sri Lankan Tamil in the United Kingdom
AMOVA analysis of molecular variance
F_{ST} fixation index
MAF minimum allele frequency
MP match probability
PD power of discrimination
PIC polymorphism information content
PE power of exclusion
TPI typical paternity index
He heterozygote
CMP cumulative match probability
CPD cumulative power of discrimination
CPE cumulative power of exclusion
HWE Hardy–Weinberg equilibrium
LD linkage disequilibrium
EAIS East Asia-based A-InDel Set (57 A-InDels)
UAIS Universal A-InDel Set (44 A-InDels)
PCA principal component analysis
MDS multidimensional scaling plot
CE capillary electrophoresis
MEGA Molecular Evolutionary Genetics Analysis
iTOL Interactive Tree of Life
SAS® Statistical Analysis System
R *R* project for statistical computing

frontiers in Genetics

ORIGINAL RESEARCH
published: 10 March 2022
doi: 10.3389/fgene.2022.827277

Sex-Biased Population Admixture Mediated Subsistence Strategy Transition of Heishuiguo People in Han Dynasty Hexi Corridor

Jianxue Xiong[1†], Panxin Du[1†], Guoke Chen[2], Yichen Tao[1], Boyan Zhou[3], Yishi Yang[2], Hui Wang[4,5], Yao Yu[4], Xin Chang[4], Edward Allen[4], Chang Sun[1], Juanjuan Zhou[1], Yetao Zou[1], Yiran Xu[4], Hailiang Meng[1], Jingze Tan[1], Hui Li[1]* and Shaoqing Wen[4,5]**

[1]Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Fudan University, Shanghai, China, [2]Institute of Cultural Relics and Archaeology in Gansu Province, Lanzhou, China, [3]Division of Biostatistics, Department of Population Health, School of Medicine, New York University, New York, NY, United States, [4]Institute of Archaeological Science, Fudan University, Shanghai, China, [5]Center for the Belt and Road Archaeology and Ancient Civilizations (BRAAC), Fudan University, Shanghai, China

OPEN ACCESS

Edited by:
Marc Via,
University of Barcelona, Spain

Reviewed by:
Levon Yepiskoposyan,
Armenian National Academy of Sciences, Armenia
Jiang Huang,
Guizhou Medical University, China

***Correspondence:**
Jingze Tan
jztan@fudan.edu.cn
Hui Li
lihui.fudan@gmail.com
Shaoqing Wen
wenshaoqing@fudan.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: 01 December 2021
Accepted: 10 January 2022
Published: 10 March 2022

Citation:
Xiong J, Du P, Chen G, Tao Y, Zhou B, Yang Y, Wang H, Yu Y, Chang X, Allen E, Sun C, Zhou J, Zou Y, Xu Y, Meng H, Tan J, Li H and Wen S (2022) Sex-Biased Population Admixture Mediated Subsistence Strategy Transition of Heishuiguo People in Han Dynasty Hexi Corridor. Front. Genet. 13:827277. doi: 10.3389/fgene.2022.827277

The Hexi Corridor was an important arena for culture exchange and human migration between ancient China and Central and Western Asia. During the Han Dynasty (202 BCE–220 CE), subsistence strategy along the corridor shifted from pastoralism to a mixed pastoralist-agriculturalist economy. Yet the drivers of this transition remain poorly understood. In this study, we analyze the Y-chromosome and mtDNA of 31 Han Dynasty individuals from the Heishuiguo site, located in the center of the Hexi Corridor. A high-resolution analysis of 485 Y-SNPs and mitogenomes was performed, with the Heishuiguo population classified into Early Han and Late Han groups. It is revealed that (1) when dissecting genetic lineages, the Yellow River Basin origin haplogroups (i.e., Oα-M117, Oβ-F46, Oγ-IMS-JST002611, and O2-P164+, M134-) reached relatively high frequencies for the paternal gene pools, while haplogroups of north East Asian origin (e.g., D4 and D5) dominated on the maternal side; (2) in interpopulation comparison using PCA and *Fst* heatmap, the Heishuiguo population shifted from Southern-Northern Han cline to Northern-Northwestern Han/Hui cline with time, indicating genetic admixture between Yellow River immigrants and natives. By comparison, in maternal mtDNA views, the Heishuiguo population was closely clustered with certain Mongolic-speaking and Northwestern Han populations and exhibited genetic continuity through the Han Dynasty, which suggests that Heishuiguo females originated from local or neighboring regions. Therefore, a sex-biased admixture pattern is observed in the Heishuiguo population. Additionally, genetic contour maps also reveal the same male-dominated migration from the East to Hexi Corridor during the Han Dynasty. This is also consistent with historical records, especially excavated bamboo slips. Combining historical records, archeological findings, stable isotope analysis, and paleoenvironmental studies, our uniparental genetic investigation on the Heishuiguo population reveals how male-dominated migration accompanied with lifestyle adjustments brought by these eastern groups may be the main factor affecting the subsistence strategy transition along the Han Dynasty Hexi Corridor.

Keywords: Hexi Corridor, subsistence strategy, sex-biased admixture, Y chromosome, mitogenome, ancient DNA

INTRODUCTION

Human history can be seen as a history of dealing with new challenges caused by changes of factors including resource distribution and social relationships. To overcome them, human beings have been using extrasomatic ways, including subsistence-, socio-, and ideo-technologies to create new niches for survival (Binford, 1962; Zhang 2021). Among the studies of adaptive changes, the evaluation of the factors leading to significant changes in subsistence strategy in human prehistory and history is a fascinating topic. Climate change, technological innovation, rapidly growing population, trans-continental cultural exchange, human migration, and geopolitics are clear potential candidates (Cohen, 1975; Bonsall et al., 2002; Gao et al., 2007; Pokharia et al., 2017; Cheung et al., 2019; Petraglia et al., 2020; Yang et al., 2020; Li et al., 2021). Climate change is well-studied and considered one of the most guiding factors of shifting subsistence strategies. A suitable climate will be conducive to the promotion of agricultural development, as rainy weather conditions facilitated the development of oasis agriculture in South-east Arabia around 5,100 cal BP (Preston, 2011), to take one example. On the other hand, an extraordinarily harsh climate can result in civilizational collapse. Humans are required to adopt different strategies, as well as adjust subsistence strategies, even migrate in search of better conditions, in order to cope with and adapt to such environmental changes (Polyak and Asmerom, 2001; Nunez et al., 2002; Haug et al., 2003; An et al., 2005; Preston et al., 2012; King et al., 2013; d'Alpoim Guedes et al., 2015; Jia et al., 2016; Pokharia et al., 2017).

The Hexi Corridor, located centrally on the eastern Silk Road, once played a crucial role in cultural exchange between east and west. The area was also a crossroads of agricultural and nomadic populations within China. Multiple lines of evidence showed that a significant shift in subsistence strategies along the Hexi Corridor occurred during the Han Dynasty (Ma et al., 2016; Yang et al., 2019a; Yang et al., 2020). Historical records from the *Shiji* (Records of the Grand Historian, 史记) to the *Hanshu* (Book of Han, 汉书) stated that the Hexi Corridor was occupied by nomadic pastoralists (Xiongnu, Yuezhi, and Wusun) before the Han Dynasty, a claim supported by excavated relics and faunal remains from Shajing culture (2,900–2,100 cal yr BP) and Shanma culture (2,700–2,100 cal yr BP) (Yang et al., 2019b) sites. The importance of domesticated pig declined precipitously during this same period (GPICRA, 2001), while sheep/goat, cattle, horse, and camel emerged as the major domesticated animals along the Hexi Corridor (GPICRA and SAMPU, 2011). A wealth of leather and woolen products had also been found at Shajing culture sites, suggesting how residents had initiated their own secondary products revolution with these domesticated pastoral animals (Sherratt, 1981). These were all indications suggesting that Hexi populations lived highly mobile nomadic lifestyles. During and after the Han Dynasty (202 BCE–220 CE), however, agriculture developed rapidly to become the subsistence strategy in this region. Numerous domesticated crop types had been found at archaeological sites, including barley, wheat, millet, highland barley, and pea (Sun and Liu, 2014), while advanced iron implements including iron plough, iron sickle, and iron spade had been excavated from numerous Han Dynasty sites (Yang, 2015). This strongly suggested that an advanced agricultural technology became widespread in the Hexi Corridor during this period of time. Remains of chicken, pigs, dogs, sheep/goat, cattle, and horse identified at the Heishuiguo Han Dynasty tombs should be placed in this context. Chicken would emerge as the most common domestic animal in this period, followed by pig (Li, 2021). This Heishuiguo population of domesticates resembled what we find with Central Plain farmers while being vastly different from that of nomads (the assemblage of domestic animals was sheep/goat, horse, and cattle) (Deng, 2015), who generally fed the camel and horse for long-distance migration, while pig and chicken may be more likely to appear in settled-peoples' homes (Yang et al., 2019a). Finally, large numbers of painted murals depicting farming and animal-grazing had been found at Hexi Corridor sites, revealing this predominant Han-Jin Dynasties mixed economy (Zheng and Gao, 2019). We can be sure, therefore, that subsistence strategy in the Hexi Corridor shifted from nomadic style (pre-Han Dynasty) to a mixed style (during and post-Han Dynasty). The argument was also supported by stable isotope data (Li, 2021).

Located in an arid climatic transitional belt, the Hexi Corridor is extremely sensitive to changes in its environment. Previous studies have primarily attempted to explain shifting subsistence strategies through the mirror of environmental change (Zhou et al., 2016; Shi et al., 2018; Yang et al., 2019a; Yang et al., 2020). But we believe that the influence of population migration should not be overlooked, especially when considering the Hexi Corridor's unique geographical position. Uniparentally inherited markers (Y chromosome or mitochondria) have been widely used in human population migration studies. The non-recombining portion of the Y chromosome (NRY) is strictly inherited paternally, while mtDNA is inherited maternally (Calafell and Larmuseau, 2017). MtDNA and Y chromosome therefore provide a matrilineal and patrilineal demographic history, respectively, revealing pictures of sex-specific processes in the past. Notably, genetic history revealed by Y chromosomes need not be identical to that by mtDNA. By using both mtDNA and Y chromosome inherited markers, sex-biased migration has been frequently found in studies of human populations and therefore might truly reflect the influence of social behaviors (Oota et al., 2001). So as to discuss the relationship between the change of subsistence strategies in the Hexi Corridor and population migrations during the Han Dynasty, here we analyze the Y-chromosome (including 485 Y-SNP markers) and mitochondrial genomes from 31 samples from the Heishuiguo site covering the whole Han Dynasty.

MATERIALS AND METHODS

Materials

The Heishuiguo site is located in Ganzhou county, Zhangye city, Gansu province, China (**Figure 1**). The site was excavated by the Institute of Cultural Relics and Archaeology of Gansu Province in

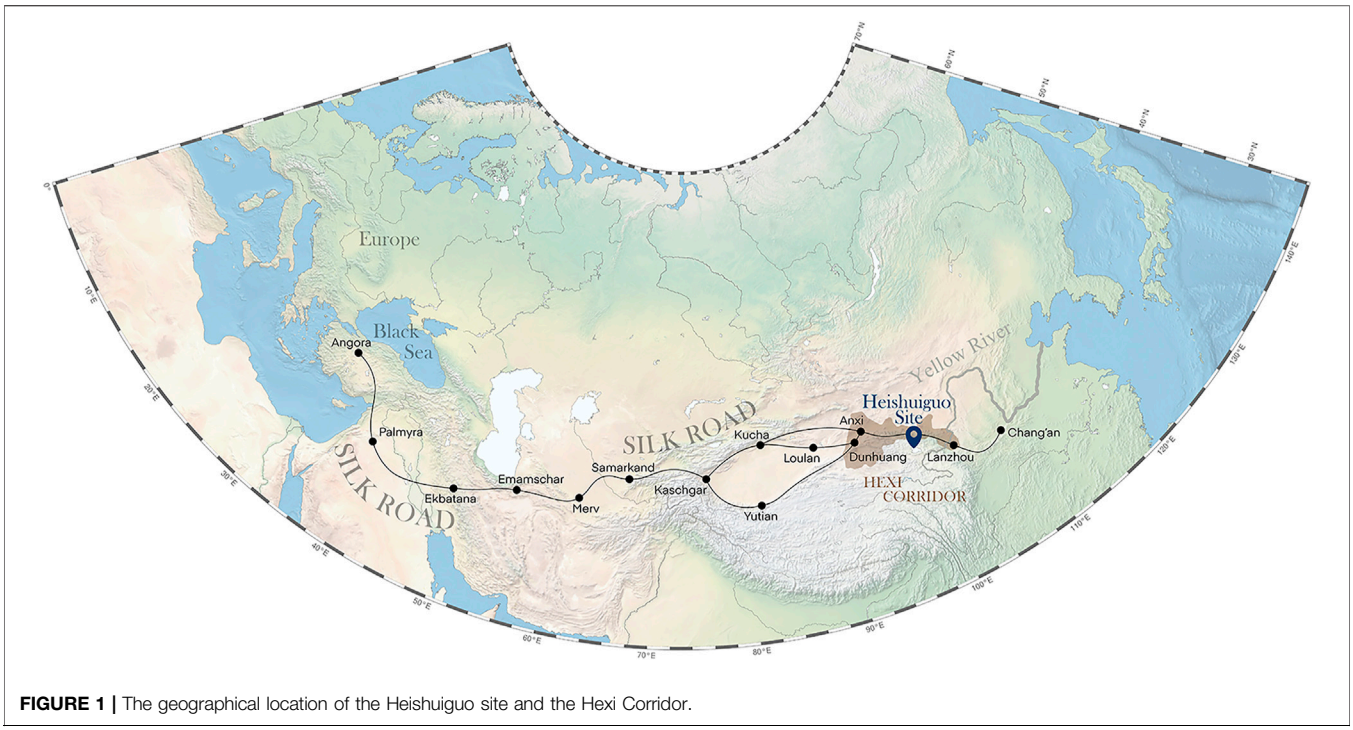


FIGURE 1 | The geographical location of the Heishuiguo site and the Hexi Corridor.

2018. Tomb distribution at Heishuiguo reveals a large cemetery consisting of family burial grounds and scattered burial groups (Chen et al., 2019). Based on ^{14}C dating, tomb morphology and grave good assemblages (Li, 2021), burials at Heishuiguo have been divided into four phases: Phase 1, during the middle Western Han Dynasty (118–49 BCE); Phase 2, during the late Western Han Dynasty (48 BCE–6CE); Phase 3, from the Wangmang Xin Dynasty through the early Eastern Han Dynasty (7–67 CE); and Phase 4, from the middle to the late Eastern Han Dynasty (67–191 CE) (Zhang, 2017; Chen et al., 2019).

This study classified 31 individuals into two groups, the Early Han (Phases 1–2) and Late Han Dynasty (Phases 3–4), respectively. Heishuiguo individual sex was determined by pelvic (Klales et al., 2012) and skull morphology (Buikstra, 1994). Details are provided in **Table 1** and **Supplementary Table S1**.

Methods

DNA Extraction

We extracted DNA from 31 samples using a dedicated aDNA facility at Fudan University, Shanghai, following the established precautions for working with ancient human DNA (Knapp et al., 2012; Sun et al., 2021). Negative extraction control samples (no sample powder used) were included, to monitor against contamination, as well as library negative controls (extract supplemented by water) in every batch of samples, which were processed and carried through the entire wet laboratory processing.

Prior to sampling, all samples were irradiated with UV-light for 30 min on each side and wiped with a 5% bleach solution. Teeth and other osseous materials were sandblasted to remove the outer surface before being ground to fine power with a mixer mill (Retsch, Germany). A dense section of temporal bones was cut around the cochlea by first removing the outer part, then grinding the clean inner part into fine powder. A total of 100 mg of bone powder was then used to extract DNA. Pre-lysis methods included the addition of a 1 ml extraction buffer, containing 0.5 M EDTA, 0.25 mg/ml Proteinase K (Merck, Germany), pH 8.0, after which samples were rotated for 1 hour at a temperature of 37°C. After centrifugation, the supernatant was discarded and 2.5 ml extraction buffer added, followed by overnight rotation at 37°C. We mixed 20 µl magnetic beads (Enlighten Biotech, China) with 12.5 ml binding buffer containing 5 M GuHCl, 40% Isopropanol, 25 mM sodium acetate, and 0.05% Tween-20 (Merck, Germany), (PH 5.2). We then transferred the supernatant (~2.5 ml) to a binding buffer/bead mixture prior to the robotic extraction (Enlighten Biotech, China) procedure. At last, the DNA underwent elution through a 50 µl TET buffer (QIAGEN, Germany).

Library Preparation

We prepared double-stranded libraries in accordance with Meyer's protocols (Meyer and Kircher, 2010), but with minor alterations. Libraries were amplified with indexing primers in two parallel polymerase chain reactions (PCR) using Q5 High-Fidelity DNA Polymerase (New England Biolabs, USA). Indexed products from the same library were pooled and purified using Agencourt AMPure XP beads (Beckman Coulter, Germany) and then eluted in 20 µl TET buffer. We quantitated the clean-up libraries using a Qubit 2.0 Fluorometer (Thermo Fisher, USA). We then sequenced the libraries on an

TABLE 1 | Ancient individuals sampled in this study.

Sample ID	Archaeological ID	Periods	Skeletal element	5 C-T%	Sex (Genetic)	Contamination	Mt_Depth	Mt Haplogroup	Y-SNPs	Y Haplogroup
FA0211	M57 east	Early Han	Temporal bone	—	Male	—	0.109	—	480	O-F1759
G10105	M18 west	Early Han	Fibula	—	—	—	—	—	372	O-F325
FA0212	M54 east	Early Han	Temporal bone	12	Male	0.071	32.1589	—	480	O*-F996
EA1102	M4	Early Han	Temporal bone	10	Male	0.087	5.9048	B5a*	480	O-F8
EA1110	M23 west	Early Han	Clavicle	10	—	0.010	211.9229	B5b2a2*	480	O-F325
G10103	M15 south	Early Han	Fibula	5	—	0.089	20.7101	D4	413	C*-M217
FA0213	M30 east	Early Han	Temporal bone	10	—	0.061	30.1299	D4	480	O-F1736
EA1132	M25-2	Early Han	Tooth	11	Male	0.174	5.411	D4a6	478	N*-CTS439
EA1107	M19 west	Early Han	Temporal bone	10	Male	0.295	1.8525	D5a2a	477	O-F325
FA0205	M59	Early Han	Temporal bone	15	Male	0.019	54.9705	D5a2a1+@ 16172*	480	O*-F2924
FA0209	M57west	Early Han	Temporal bone	12	Female	0.066	17.3689	M11d	—	—
G30401	M33	Early Han	Tibia	14	Male	0.400	1.7359	M33c	445	O-F325
EA1101	M6①	Early Han	Temporal bone	14	Male	0.150	6.4597	R11a	454	O-F8
EA1130	M62 east	Late Han	Limb bone	—	—	—	0.0918	—	479	O*-F1365
G30705	M115 east	Late Han	Temporal bone	15	Male	0.018	59.0987	B4a1c3b	480	O-F1759
F11325	M84	Late Han	Fibula	17	—	0.009	89.4457	B5a2a1a	464	O-F1266
FA0210	M116 west	Late Han	Temporal bone	16	Male	0.023	41.4219	C4a1a2*	481	O-F1736
G10101	M13	Late Han	Limb bone	12	—	0.050	18.6234	D4a3b*	408	N-F710
EA0420	M90 west	Late Han	Tooth	19	—	0.002	361.4428	D4b2b*	480	C*-M217
EA1106	M98①	Late Han	Tooth	18	Male	0.001	730.8106	D4b2b*	480	N-F710
G30704	M113 north	Late Han	Temporal bone	16	Male	0.026	35.4385	D4j*	480	C-F5477
FA0206	M58	Late Han	Temporal bone	12	Male	0.018	30.8458	D5a2*	480	Q-1827
G40801	M79	Late Han	Limb bone	25	—	0.015	59.4633	D5b1b*	239	O-F141
FA0215	M38	Late Han	Temporal bone	14	Male	0.061	21.1436	D5c*	480	O-F8
G30703	M93	Late Han	Temporal bone	17	Male	0.016	60.0148	F1a1a*	480	O*-F1365
G30701	M5 west	Late Han	Temporal bone	11	Male	0.005	129.8333	G1c*	480	O-F4068
G30702	M9 north	Late Han	Temporal bone	13	Male	0.027	1.7359	G3	481	N-F710
FA0201	M80 north	Late Han	Temporal bone	13	Male	0.065	17.1204	M9a1a1b	480	O*-F46
EA1116	M92 east	Late Han	Limb bone	21	—	0.001	901.0238	N9a1*	251	O-F60
EA1104	M90 east	Late Han	Tooth	11	Male	0.069	7.898	R11b	480	C*-F3967
FA0214	M108 west	Late Han	Temporal bone	12	Male	0.033	29.8907	R11b*	481	O-F1759

Illumina HiSeq X10 instrument at the Annoroad Company (Beijing, China) using the 150-bp paired-end sequencing design.

Mitochondrial Capture and Sequencing

Target enrichment of the mitogenome was performed on each amplified library using a MyGenostics Human Mitochondria Capture Kit (MyGenostics Company, Beijing, China) as described by Sun et al. (2021), with hybridization and wash temperatures lowered to 60°C and 55°C to facilitate the enrichment of our short library molecules, in line with Dabney et al. (2013). A final post-enrichment amplification was performed for 15 cycles. The post-enrichment amplified

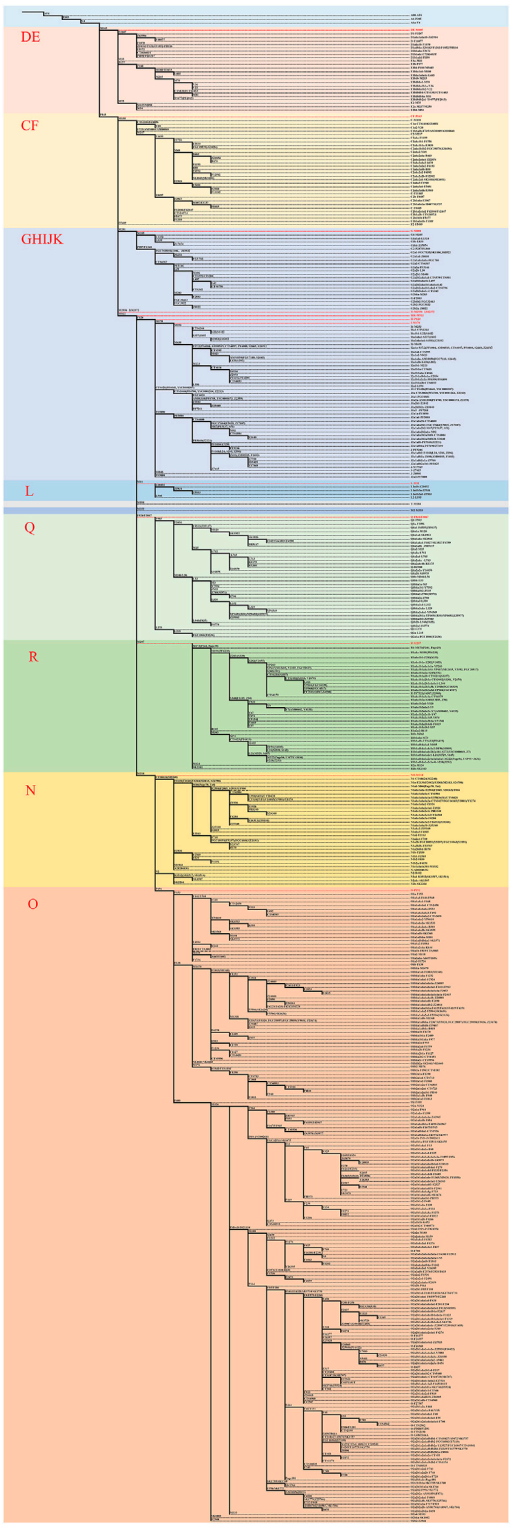


FIGURE 2 | The phylogenetic tree of 485 Y-SNPs in this study.

product was then quantified via qPCR. Sequencing was performed using a Novaseq 6,000 platform at Mingma Technologies Company (Shanghai, China). Next, 150 bp paired-end reads were generated according to the manufacturer's instructions.

Multiplex Polymerase Chain Reactions Targeted Amplification and Sequencing for Y Chromosome

Ancient DNA fragment lengths are generally skewed toward short fragments, the vast majority of which are typically shorter than 100 bp (Sawyer et al., 2015). Furthermore, in order to preliminary screen ancient samples quickly and inexpensively, we opted for multiplex PCR targeting enrichment with short amplicons based on the NGS (Next Generation Sequencing) platform. After amplification enrichment, a large number of samples could be detected and analyzed in parallel using the NGS platform. In view of the characteristics of highly degraded DNA involved in this study, we designed a more sensitive short amplifier primer system (Wen et al., 2019) and conducted tests on samples from the Heishuiguo site. The system comprises 485 Y-SNPs (**Figure 2**; **Supplementary Tables S2, S3**), covering common lineages in East Asia. Details can be found in the **Supplementary Material**. PCR amplification, sequencing, and data analysis can also be found in the **Supplementary Material**s.

Sequence Data Processing

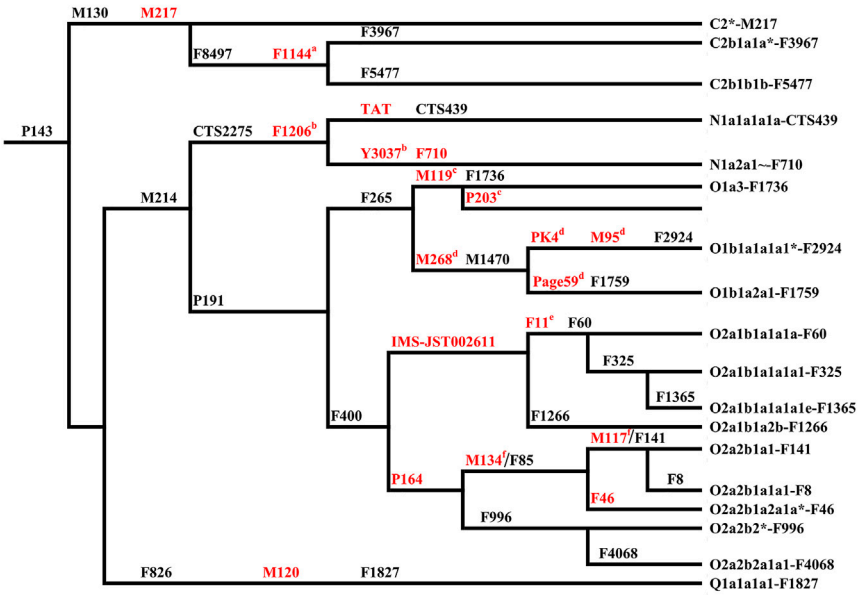
For shotgun and mtDNA capture data, we clipped sequencing adapters and merged using sequences by ClipAndMerge v1.7.8 (Peltzer et al., 2016). Following this, we mapped merged reads to the human reference genome (hs37d5; GRCh37 with decoy sequences) using BWA v0.7.17 (Li and Durbin, 2010). PCR duplicates were removed using Dedup v0.12.3 (Peltzer et al., 2016). Using trimBam implemented in BamUtil v1.0.14 (https://github.com/statgen/bamUtil), we clipped four bases from both ends of each read to avoid an excess of remaining C->T and G->A transitions at the ends of the sequences.

Ancient DNA Authentication

The authenticity of the ancient genome sequence was mainly determined by the combination of two observations of the same specimen. Firstly, deep sequencing of the mitochondrial genome would show that the vast majority of the DNA fragments had come from a single individual. Second, the patterns of DNA degradation (**Supplementary Figure S1**), in particular nucleotide misincorporations resulting from deamination of cytosine residues at the ends of DNA fragments, would indicate that the mtDNA is ancient. We first checked DNA damage pattern and estimated the 5′ C>T and 3′ G>A misincorporation rate using mapDamage v 2.0.61 (Jónsson et al., 2013). We then made use of a Schmutzi program to test mitochondrial contamination rates for all individuals (Renaud et al., 2015).

Uniparental Haplogroup Assignment

For mtDNA, we employed a log2fasta program implemented in Schmutzi (Renaud et al., 2015) in order to call the mtDNA consensus sequences. Mutations that appeared when checked against rCRS were also re-checked in BAM (Binary Alignment Map) files through visual inspection with IGV software (Helga



	Early Han		Late Han	
	N	Frequency	N	Frequency
C2*-M217	1	8.3%	1	5.6%
C2b1a1a*-F3967			1	5.6%
C2b1b1b-F5477			1	5.6%
N1a1a1a1a-CTS439	1	8.3%		
N1a2a1~-F710			3	16.7%
O1a3-F1736	1	8.3%	1	5.9%
O1b1a1a1a1*-F2924	1	8.3%		
O1b1a2a1-F1759	1	8.3%	2	11.1%
O2a1b1a1a1a-F60			1	5.6%
O2a1b1a1a1a1-F325	4	33.3%		
O2a1b1a1a1a1e-F1365			2	11.1%
O2a1b1a2b-F1266			1	5.6%
O2a2b1a1-F141			1	5.6%
O2a2b1a1a1-F8	2	16.7%	1	5.6%
O2a2b1a2a1a*-F46			1	5.6%
O2a2b2*-F996	1	8.3%		
O2a2b2a1a1-F4068			1	5.6%
Q1a1a1a1-F1827			1	5.6%

FIGURE 3 | The phylogenetic relationship of Y-chromosome haplogroups in this study and their haplogroup-based frequencies in the sampled populations (Early Han, the early Han group; Late Han, the late Han group). Marker names are shown along the branches, and haplogroup names are shown to the right, based on ISOGG Y-DNA Haplogroup Tree 2019. Asterisks distinguish potentially paraphyletic undefined subgroups from recognized haplogroups. The markers in red are key Y-SNPs. a-f: These markers are not designed in Y-SNP panel but very common in Y phylogenetic trees.

et al., 2013). Then, we used Haplogrep 2 (Weissensteiner et al., 2016) to assign the haplogroups.

Y chromosome haplogroups were examined by aligning a set of positions in the ISOGG (International Society of Genetic Genealogy, http://isogg.org/) and Y-full (https://www.yfull.com/tree/) databases, and analysis performed in the case of a base and mapping quality exceeding 30. Haplogroup determination was performed with the script Yleaf.py in Yleaf software (Ralf et al., 2018), which provides outputs for allele counts of ancestral and derived SNPs along a path of branches of the Y-chromosome tree. Finally, we re-checked the SNPs by visual inspection with IGV software (Helga et al., 2013).

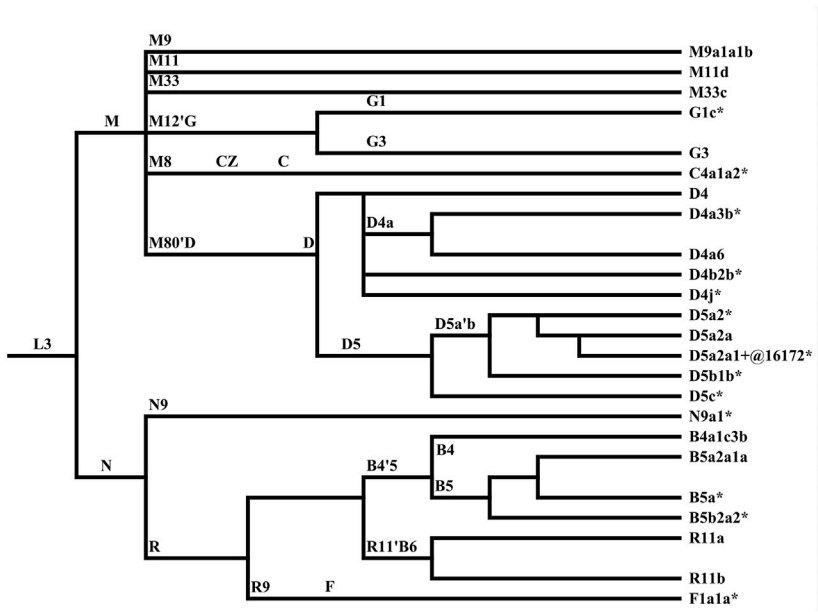
Statistical Analyses

Principal component analysis (PCA) was performed using FactoMineR package of R 3.6.3. Reference populations are listed in **Supplementary Tables S4–S6**. The pairwise genetic *Fst* was calculated using Arlequin 3.5 software. Heatmaps were also used to illustrate the clusters based on *Fst* by stats package of R 3.6.3. At last, to visualize the origins, we used PC1 values from PCA plot to generate contour maps with Surfer 12.0 software (Golden Software, https://www.goldensoftware.com/).

RESULTS

Y-Chromosome and mtDNA Haplogroup Profile

Y chromosome haplogroups of 30 Heishuiguo males were determined according to the ISOGG's Y-DNA Haplogroup Tree 2019 (**Figure 3**; **Supplementary Table S1**). Overall, the Heishuiguo population was comprised of the haplogroups Oγ-IMS-JST002611(26.7%), Oα-M117 (13.3%), C2-M217(13.3%), N-F1206 (13.3%), O1b1a2-Page59 (10%), O1a-M119+, P203- (6.7%), O2-P164+, M134- (6.7%), Oβ-F46 (3.3%), O1b1a1a-M95 (3.3%), and Q-M120 (3.3%). Haplogroups Oγ-IMS-JST002611 (26.7%), Oα-M117 (13.3%), and Oβ-F46(3.3%) represent three major founder paternal lineages (Yan et al., 2014; Wen et al., 2016) and encompass more than 40% of the present-day Han Chinese (estimated 16% for Oα, 11% for Oβ, and 14% for Oγ) (Yan et al., 2011). These three lineages have been considered to be derived from Neolithic farmers in Yellow River Basin (Yan et al., 2014). The three haplogroups are also the most frequent among the Heishuiguo population, representing 43.3% of the total, similar to modern Han populations. Haplogroup O2-P164+, M134-, may have expanded similarly with the above three haplogroups. Haplogroups C2-M217, N-F1206, and Q-M120 are common in Chinese, Altaic, Uralic, and Northern Eurasian Indo-European populations. The vast majority of haplogroup C in China belongs to C2-M217 (Zhong et al., 2010), constituting ~10% of Han Chinese, as well as a great portion of Altaic-speaking populations, i.e., Mongol, Manchu, and Kazakh people. In the Heishuiguo population, the haplogroup C2-M217 can be further classified into C2*-M217 and C-F1144. The latter is a southern clade (Yan et al., 2014) and one of the six major paternal lineages of the Han Chinese (Wen et al., 2016; Wu et al., 2020). N-F1206 is called the northern clade of N haplogroup and widely distributed across Northern Eurasia (Hu et al., 2015). N-F1206 can be sub-divided into N-TAT and N-F710. The highest frequency of N-TAT is found in



Early Han		Late Han	
N	Frequency	N	Frequency
		1	5.88%
1	10.00%		
1	10.00%		
		1	5.88%
		1	5.88%
		1	5.88%
2	20.00%		
		1	5.88%
1	10.00%		
		2	11.76%
		1	5.88%
		1	5.88%
1	10.00%		
1	10.00%		
		1	5.88%
		1	5.88%
		1	5.88%
		1	5.88%
		1	5.88%
1	10.00%		
1	10.00%		
1	10.00%		
		2	11.76%
		1	5.88%

FIGURE 4 | The phylogenetic relationship of mtDNA haplogroups surveyed in this study and their haplogroup-based frequencies in the sampled populations. Haplogroup names are shown to the right, according to the PhyloTree mtDNA tree Build 17.

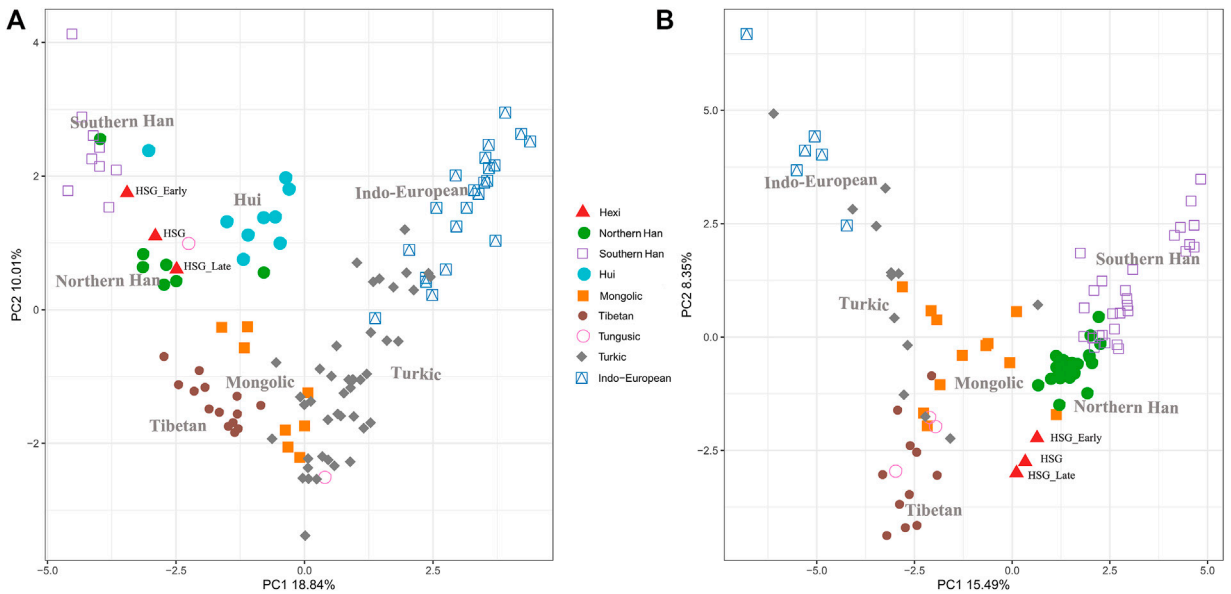


FIGURE 5 | Principal component plots of Heishuiguo population and reference populations for Y-chromosome haplogroups **(A)** and mtDNA haplogroups **(B)**.

Altaic, Uralic, and Indo-European speaking populations (Ilumäe et al., 2016), such as Vilyuy Yakuts (91.525%), Evenks (50.877%), Buryats (41.441%), Udmurts (66.667%), Finns (53.846%), and Latvians (43.023%), while the N-F710 is considered as having migrated from Northeast Asia southward into the Yellow River region at no later than 2.7 kya (Ma et al., 2021). Haplogroup Q-M120 originated in South Siberia and expanded across northwestern China between 5–3 kya (Sun et al., 2019). The lineage was absorbed into ancient Huaxia (Han Chinese) populations before 2 kya and would eventually become one of the six founder lineages in modern Han populations (Wen et al., 2016; Sun et al., 2019). Finally, the O1b1a2-Page59, O1a-M119+, P203-, and O1b1a1a-M95 haplogroups, making up 20% of the Heishuiguo haplogroup proportions, are of minor southern East Asia origin (Karafet et al., 2010; Cai et al., 2011; Trejaut et al., 2014; Zhang et al., 2015; Luo et al., 2020). The paternal diversity

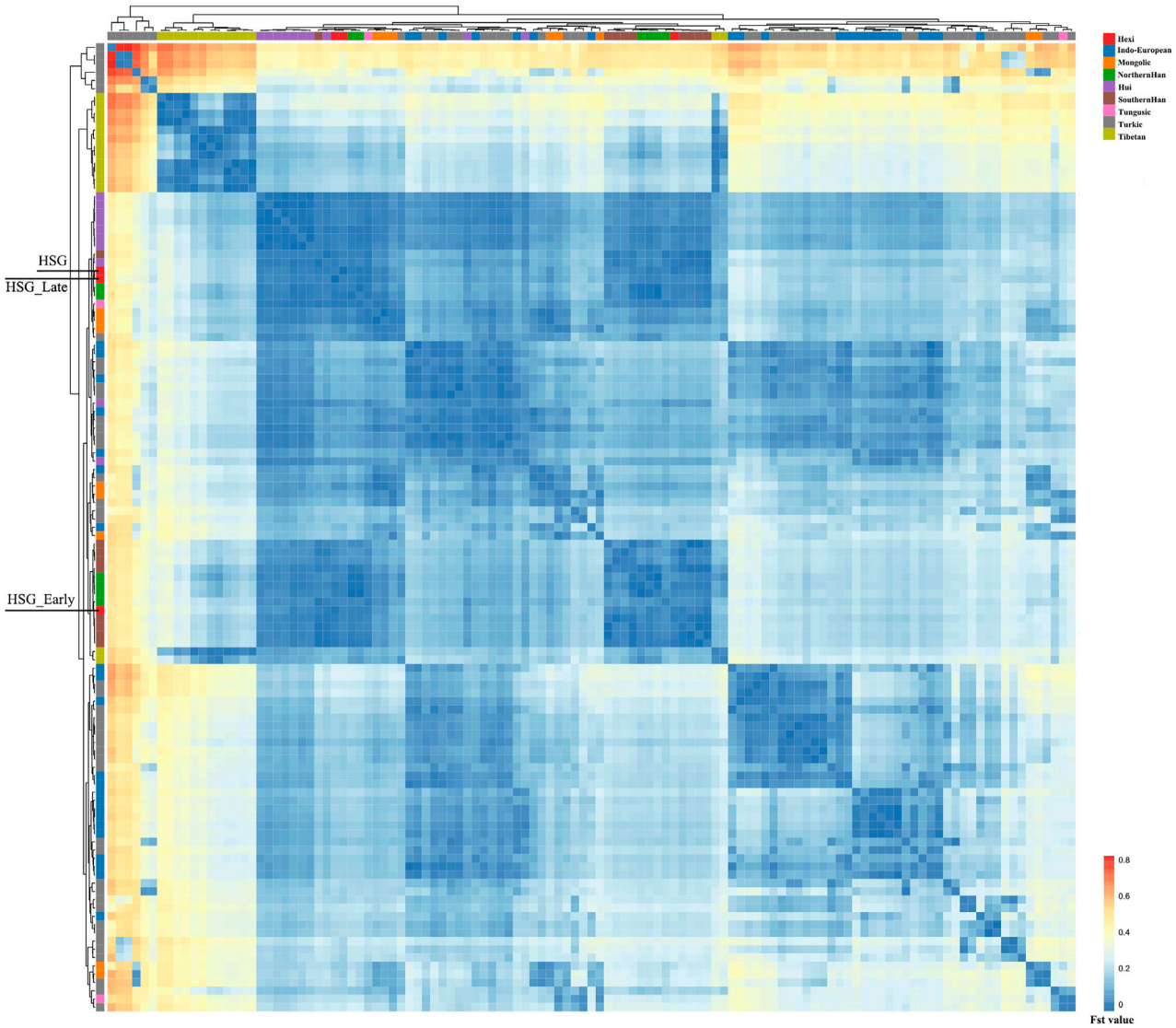


FIGURE 6 | The genetic distance (*Fst*) heatmap plot of Heishuiguo population and reference populations for Y-chromosome.

of the Late Han Heishuiguo group exceeds that of the Early Han group, and features the addition of haplogroups N-F710, C-F1144, Oβ-F46, and Q-M120.

MtDNA haplogroups of 27 Heishuiguo samples were determined based on PhyloTree mtDNA tree Build 17 (**Figure 4**; **Supplementary Table S1**). Different from the Y chromosome, the mtDNA gene pool is more heterogeneous. In the Heishuiguo population, these haplogroups consisted of D4 (25.93%), D5(18.52%), B5 (11.11%), R11(11.11%), B4 (3.7%), C4 (3.7%), F1(3.7%), G1(3.7%), G3 (3.7%), M11 (3.7%), M33 (3.7%), M9 (3.7%), and N9 (3.7%). Among these, haplogroups D4, D5, C4, G1, G3, M11, and M9 have origins in north East Asia. Haplogroup D4 occurs with the greatest frequency (25.93%) in the Heishuiguo population, and also retains a very high frequency in northern Asian (average 16.7%), central Asian (average 15.3%), and eastern Asian populations (average 22.5%) (Derenko et al., 2010). In this study, D4 could be further divided into sub-clades D4, D4a3b*, D4a6, D4b2b*, and D4j*. As for D5, this haplogroup spread across northern East Asia with moderate to low frequency and reaches its highest levels in Tubalar (25%) (Volodko et al., 2008), Beijing Han (15.38%) (Jin et al., 2009), Shannan Tibetan (15%) (Xu and Hu, 2015), Henan Han (11.4%) (Xu and Hu, 2015), Orochens (11.4%) (Kong et al., 2003), Koreans (10.4%) (Kong et al., 2003) from China, and Shandong Han (10%) (Yao et al., 2002). Here, D5 includes the sub-clades D5a2*, D5a2a, D5a2a1+@16172*, D5b1b*, and D5c*. These two dominant haplogroups, D4 and D5, with 44.44% of the Heishuiguo total, have a Northern Asian origin and distribution (Derenko et al., 2010). The remaining C4, G1, G3, M11, and M9 haplogroups are more common in northern East Asia than southern Eastern Asia. The B5, B4, and F1 haplogroups (occurring with 18.52% frequency at Heishuiguo) are relatively common in southern

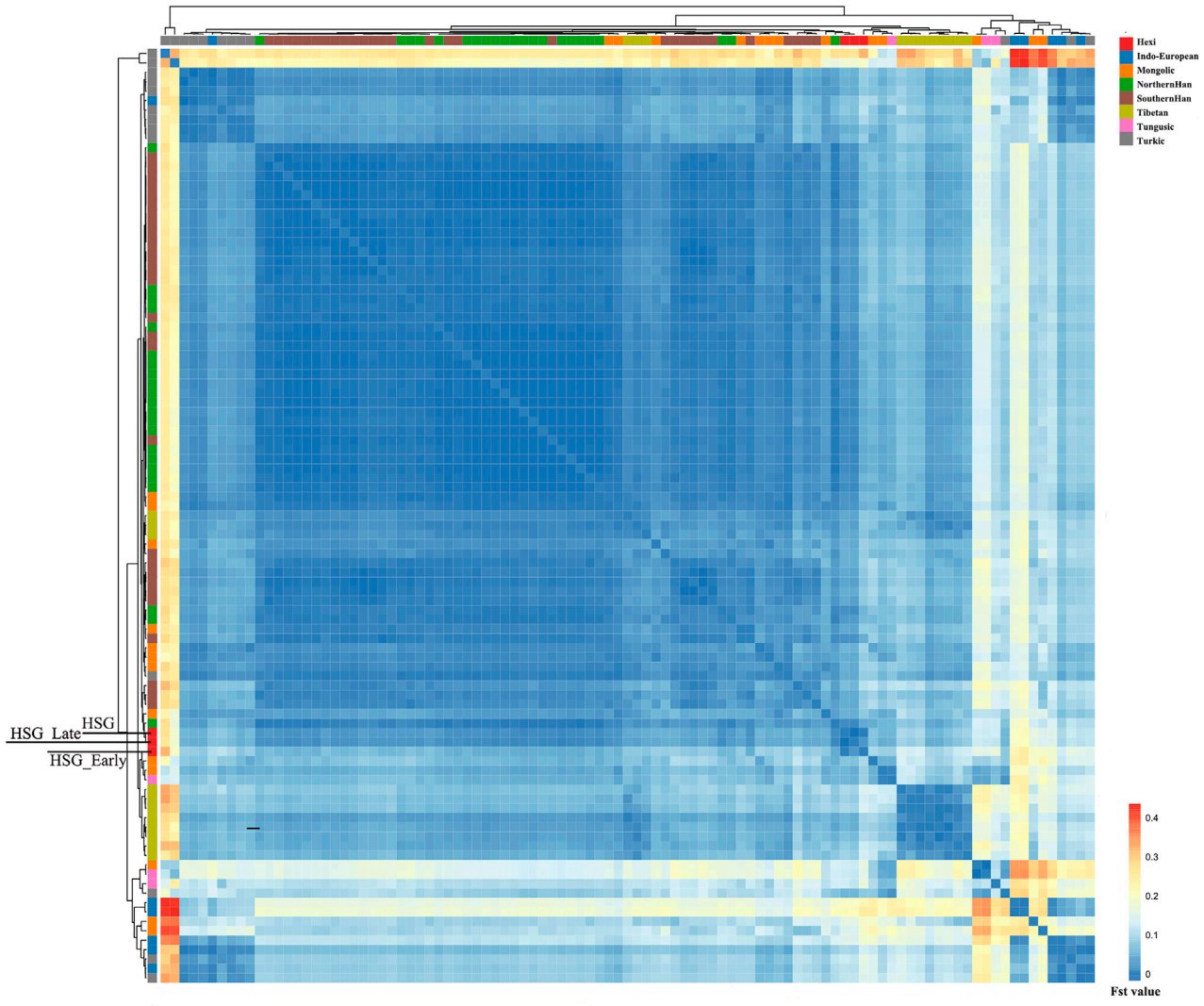


FIGURE 7 | The genetic distance (*Fst*) heatmap plot of Heishuiguo population and reference populations for mtDNA.

East Asia and may indicate gene influx from these regions. The haplogroup B5 retains high frequency in southern East Asia, especially in Tai-Kadai populations such as the Seak (69.23%) (Kutanan et al., 2017) and Kalueng (40%) (Kutanan et al., 2017) in Northeastern Thailand, Austro-Asiatic population like the Bru (45.83%) (Kutanan et al., 2017) from Northeastern Thailand, and Hmong-Mien population such as Hunan Yao (29.17%) (Wen et al., 2005) and Guizhou Han (10.19%) (Li et al., 2019). B5 is comprised of the haplogroups B5a*, B5a2a1a, and B5b2a2*. Likewise, the haplogroups F1 and B4 were common in southern East Asia and have southern Asian origins (Wen et al., 2005; Ko et al., 2014; Kutanan et al., 2017). Finally, the haplogroups R11, N9, and M33 occur sporadically in East Asian populations. We noted a similar degree of genetic diversity among both the Early Han and Late Han groups. In summary, the dominant mtDNA haplogroups at Heishuiguo, such as D4 and D5, are more frequent in northern East Asian populations and exhibit a northern Asian origin, while haplogroups B5, B4, and F1 might reflect gene flows from southern China.

Population Comparisons

In order to investigate the genetic relationships between Heishuiguo population and reference populations, we conducted a principal component analysis (PCA) based on haplogroup frequencies (**Figure 5**, **Supplementary Figure S2**; **Supplementary Tables S4–S6**). A genetic distance (*Fst*) heatmap was also visualized, to further explore population relationships (**Figures 6**, **7**; **Supplementary Figure S3**). In this study, the Heishuiguo population was classified into three groups: Early Han, Late Han, and Overall group merging the former two groups. The PCA map (**Figure 5**) shows a division between the Early Han and Late Han groups according to PC1. Our Y-chromosome principal component plot (**Figure 5A**) reveals an overall Heishuiguo

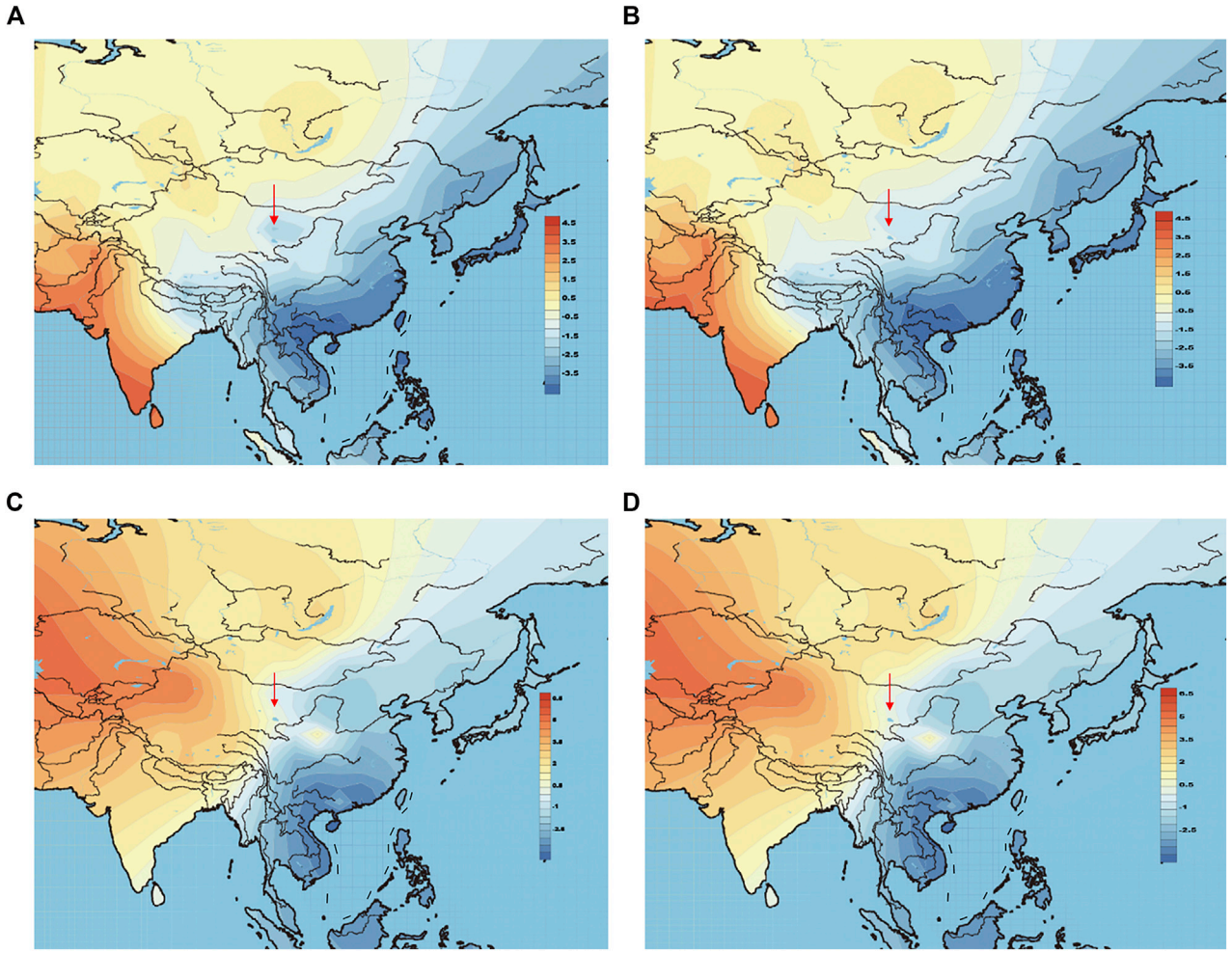


FIGURE 8 | Genetic contour maps of Heishuiguo population and reference populations for Y-chromosome with early Han group **(A)**, Y-chromosome with late Han group **(B)**, mtDNA with early Han group **(C)**, mtDNA with late Han group **(D)**.

population is projected on to the Southern Han and Northern Han cline. More specifically, the Early Group appears to cluster around Southern Han populations, especially Fujian Han and Guangdong Han. According to previous studies, these two latter populations were descended from Northern China immigrations beginning in the Han Dynasty (202 BCE–220 CE) (Ge et al., 1997; Wen et al., 2004a). Our Early Heishuiguo population may therefore also reflect migration, in this case from the Central Plains to Hexi Corridor. The Late Group clusters closely with the Northern Han population in PCA plot and additionally clusters with Northern Han and Northwestern Hui in our heatmap (**Figure 6**), suggesting the genetic admixture with indigenous people over time. On the maternal side, the three Heishuiguo populations are tightly congregated (**Figure 5B**), mirroring genetic continuity from the Early Han to Late Han. Meanwhile, the Heishuiguo population maps out close to Mongolian-speaking (e.g. Buryat) and Northwestern Han (e.g., Ningxia Han) populations, which indicates an indigenous maternal origin.

In summary, an interpopulation comparison reveals that the Heishuiguo population shows close affinity with Han Chinese in terms of paternal structure and Mongolic and Northwestern Han populations in terms of maternal structure. This is a clear indication of sex-biased population admixture.

The Origins of the Heishuiguo Population

The above-mentioned PCA plot shows the eastern and western population cline according to PC1. We can make further use of PC1 values to generate genetic contour maps (**Figure 8**) and further visualize the possible origins of Heishuiguo population. On the paternal side, the contour map with early Heishuiguo group (**Figure 8A**) shows PC1 values gradually increasing from east-to-west across northwest China before dropping abruptly in the Hexi Corridor, a clue to the non-local and likely eastern origins of the population. Such a pattern does not repeat for the Late Han Heishuiguo group (**Figure 8B**). On the other hand, the maternal side paints a different picture (**Figure 8C** and **Figure 8D**), where we do not observe significant fluctuations

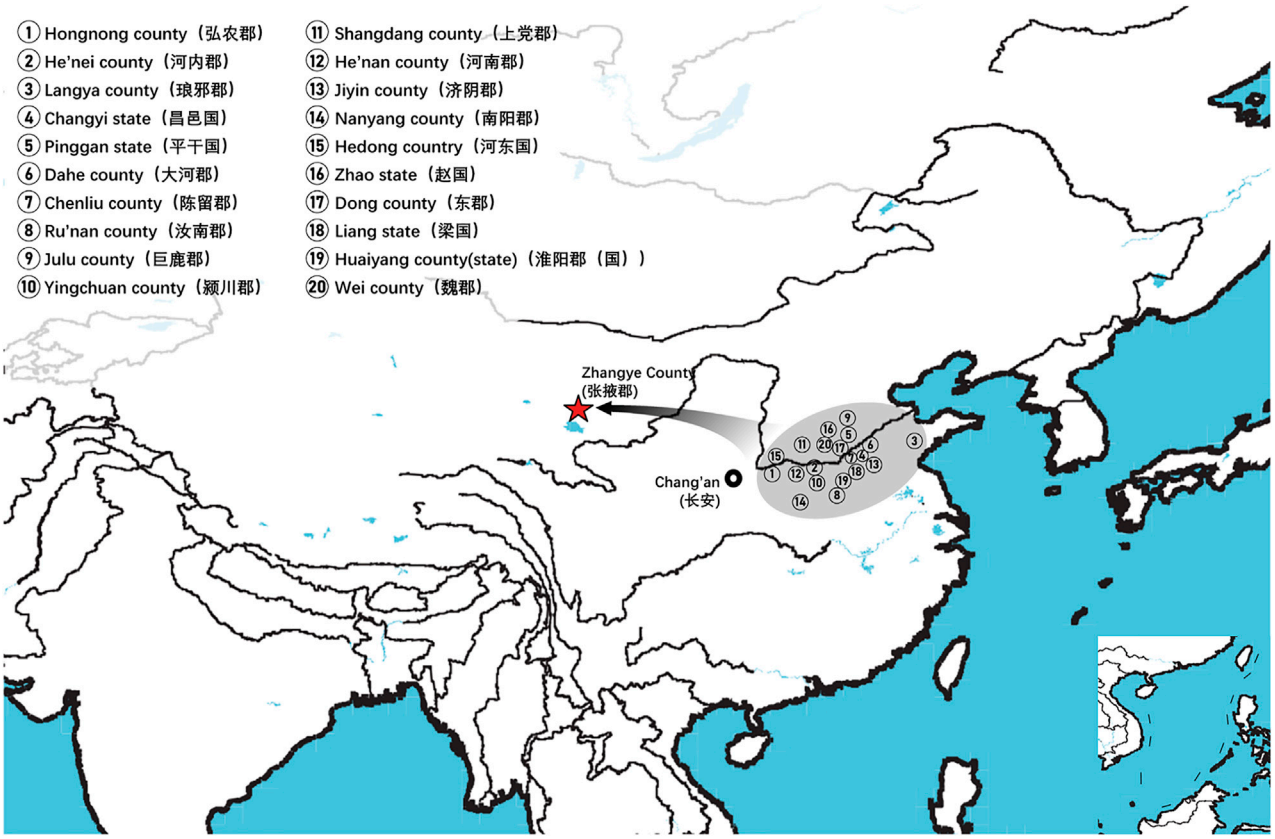


FIGURE 9 | Immigrants from 20 counties in middle and lower Yellow River to Hexi Corridor according to the contents of unearthed bamboo slips.

between the Early Han and Late Han groups. This result points to the local roots of Heishuiguo females and their genetic continuity covering the whole Han Dynasty, coinciding with our PCA plot.

DISCUSSION

Multiple lines of evidence (i.e., historical records, archeological finds, and stable isotope analysis) have demonstrated how subsistence strategies along the Hexi Corridor shifted from a nomadic economy to mixed economy (i.e., pastoralism and farming) from the Han Dynasty (Sun and Liu, 2014; Li 2021). Prior research has focused on exploring the relationship between climate change and subsistence strategy transition in this region during the Han, suggesting a cold and arid climate had dried up considerable stretches of river and impoverished fertile lands prior to the Han Dynasty, with resulting prevalent nomadism. A warmer and wetter climate from the Han onwards promoted the thriving of agriculture (Yang et al., 2020). Up till now, it has seemed farfetched to link this transition to climate factor simply.

On basis of Y chromosome and mtDNA profiles of Heishuiguo population, we have located a Heishuiguo patrilineal population consisting of Yellow River Basin origin haplogroups (i.e., Oα-M117, Oβ-F46, Oγ-IMS-JST002611, and O2-P164+, M134-) at an overall rate of over 50%, along with southern East Asian origin haplogroups (O1a and O1b) at ~20% and northern East Asian origin haplogroups (C2-M217, N-F1206, and Q-M120) at ~30%. The matrilineal Heishuiguo population consisted of northern East Asian haplogroups (e.g., haplogroups D4, D5, and C4), which accounted for ~62.95%, alongside southern East Asian haplogroups (such as B5, B4, and F1) at 18.52%. By way of interpopulation comparisons (PCA and *Fst* heatmap) with reference populations (e.g., Southern Han, Northern Han, Hui, Mongolic, and Tibetan), we have been able to show closer paternal genetic affinity with Northern Han and Hui populations among the groups at Heishuiguo. Via PCA, we observed genetic structure changes from Southern-Northern Han cline to Northern-Northwestern Han/Hui cline with time (**Figure 5A**), indicating genetic admixture between Yellow River immigrants and natives. Historical records and archaeological finds add further credence to our results. According to historical documents (Qi, 1983; Si, 2002; Ban, 2008; Bamboo Slips Museum, 2013), the Han Dynasty government migrated population on a large scale from about 20 counties in Yellow River Basin to four counties (i.e. Zhangye 张掖郡, Jiuquan 酒泉郡, Wuwei 武威郡, and Dunhuang 敦煌郡) in Hexi to strengthen administration and control over this region. Bamboo Slips (简牍) provide more detail and even give us the possible source of these male immigrants in 21 counties from the Middle and Lower Yellow River (Hongnong

弘农郡, He'nei 河内郡, Langya琅邪郡, Changyi 昌邑国, Pinggan State平干国, Dahe 大河郡, Chenliu 陈留郡, Runan 汝南郡, Julu 巨鹿郡, Yingchuan 颍川郡, Shangdang 上党郡, Henan 河南郡, Jiyin 济阴郡, Nanyang 南阳郡, Hedong 河东郡, Zhao State赵国, Dong 东郡, Liang State梁国, Zhangye张掖郡, Huaiyang 淮阳郡, and Wei 魏郡) (Liu, 2012; Li, 2018, **Figure 9**). Our view from the maternal mtDNA (**Figure 5B**), however, shows the Heishuiguo population closely clustered with certain Mongolic and Northwestern Han populations and exhibiting a genetic continuity covering the whole Han Dynasty, suggesting a possible local origin for Heishuiguo females. This is in accordance with historical records (Liu 2012), where major migration events were often male-dominated migratory and frequently involved migration for garrison building, political migration, and migration of minority groups. Young males were usually the ones building garrisons, and most couldn't bring their families outside of small local garrisons. Political migration involved political prisoners, ordinary crimes, and victims of natural calamities. Migration of minority groups targeted rebels from border areas such as Di and Qiang peoples in the Upper Yellow River Basin. Among them, the military migration was in the majority. This is likely why a sex-biased admixture pattern can be clearly observed in the Heishuiguo population. Such sex-biased admixture patterns have also been observed in population expansion of Han and Tibeto-Burman-speaking (Wen et al., 2004a; Wen et al., 2004b) people. We plotted genetic contour maps to visualize the possible origins of Heishuiguo population. As shown in **Figure 8**, we can easily observe that the primary eastern East Asian origin of male ancestry and native origins of female ancestry.

In this study, by means of a uniparental genetic analysis, we have observed a male-dominated admixture event occurring during this period, one additionally supported by historical records and archaeological findings. That shifting subsistence strategy along the Hexi Corridor residents that kept pace with human migration cannot be coincidental. Mass migration of individuals and transplantation of subsistence lifestyles would have impacted the former subsistence strategy in Hexi Corridor during the Han Dynasty. This study provides new insights and possibilities into how population admixture serves as a key factor in changes of subsistence strategy.

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

Approval for the use of ancient human individuals was curated by co-authors and obtained with permission from the respective provincial archaeology institutes or universities that managed the samples. The permission and oversight were also provided by the Ethics Committee of Fudan University of Life Sciences to study their ancient genomes.

AUTHOR CONTRIBUTIONS

JT, HL, and SW designed this study. HL and SW supervised the study. GC, YiY, and HW provided materials and resources. HM and JT collected the samples. YiY performed archaeological data analysis. JX, PD, JZ, YZ, and YX performed genetic laboratory work. YT, BZ, CS, XC, and YaY performed genetic data analysis. JX and PD integrated the genetic data. JX, PD, and SW wrote the paper, with contributions from EA.

FUNDING

This work was supported by the National Key R&D Program of China (2020YFE0201600 and 2020YFC1521607), B&R Joint Laboratory of Eurasian Anthropology (18490750300), National Natural Science Foundation of China (31771325 and 32070576), Major Research Program of National Natural Science Foundation of China (91731303), Major Project of National Social Science Foundation of China (20&ZD212), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), the 111 Project (B13016), and European Research Council (ERC) grant to Dan Xu (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.2022.827277/full#supplementary-material

Supplementary Figure S1 | Ancient DNA damage patterns. Nucleotide misincorporation patterns cause by C deamination in ancient DNA sequences.

Supplementary Figure S2 | Principal component plot of Heishuiguo population and published ancient populations for mtDNA haplogroups.

Supplementary Figure S3 | The genetic distance (*Fst*) heatmap plot of Heishuiguo population and published ancient populations for mtDNA.

REFERENCES

An, C.-B., Tang, L., Barton, L., and Chen, F.-H. (2005). Climate Change and Cultural Response Around 4000 Cal Yr B.P. In the Western Part of Chinese Loess Plateau. *Quat. Res.* 63 (3), 347–352. doi:10.1016/j.yqres.2005.02.004

Ban, G. (2008). *Hanshu. The Biography of Zhaochongguo*. Taiyuan: Sanjin publishing House, 143.

Binford, L. R. (1962). Archaeology as Anthropology. *Am. Antiq.* 28 (2), 217–225. doi:10.2307/278380

Bonsall, C., Macklin, M. G., Anderson, D. E., and Payton, R. W. (2002). Climate Change and the Adoption of Agriculture in north-west Europe. *Eur. J. Archaeol.* 5 (1), 9–23. doi:10.1179/eja.2002.5.1.9

Buikstra, J. E. (1994). *Standards for Data Collection from Human Skeletal Remains*. Fayetteville, Arkansas: Arkansas archeological survey. Research series no. 44.

Cai, X., Qin, Z., Wen, B., Xu, S., Wang, Y., Lu, Y., et al. (2011). Human Migration through Bottlenecks from Southeast Asia into East Asia during Last Glacial Maximum Revealed by Y Chromosomes. *PloS one* 6 (8), e24282. doi:10.1371/JOURNAL.PONE.0024282

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

Chen, G. K., Yang, Y., and Liu, B. (2019). *Ganzhou, Zhangye: The Excavation Report of Han Dynasty Cemetery in Heishuiguo Site (Volume II)*. Lanzhou: Gansu Education Publishing House.

Cheung, C., Zhang, H., Hepburn, J. C., Yang, D. Y., and Richards, M. P. (2019). Stable Isotope and Dental Caries Data Reveal Abrupt Changes in Subsistence Economy in Ancient China in Response to Global Climate Change. *PLOS ONE* 14 (7), e0218943. doi:10.1371/journal.pone.0218943

Cohen, M. N. (1975). Archaeological Evidence for Population Pressure in Pre-agricultural Societies. *Am. Antiq.* 40 (4), 471–475. doi:10.2307/279335

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., et al. (2013). Complete Mitochondrial Genome Sequence of a Middle Pleistocene Cave bear Reconstructed from Ultrashort DNA Fragments. *Proc. Natl. Acad. Sci.* 110 (39), 15758–15763. doi:10.1073/PNAS.1314445110

d'Alpoim Guedes, J., Jin, G., and Bocinsky, R. K. (2015). The Impact of Climate on the Spread of Rice to North-Eastern China: A New Look at the Data from Shandong Province. *PloS one* 10 (6), e0130430. doi:10.1371/journal.pone.0130430

Deng, H. (2015). Animal Remains in Han Danasty Tomb. *Cult. Relics Southern China* 3, 58–69.

Derenko, M., Malyarchuk, B., Grzybowski, T., Denisova, G., Rogalla, U., Perkova, M., et al. (2010). Origin and post-glacial Dispersal of Mitochondrial DNA Haplogroups C and D in Northern Asia. *PloS one* 5 (12), e15214. doi:10.1371/journal.pone.0015214

Gansu Bamboo Slips Museum (GBSM) (2013). *Han Dynasty Wooden Slips from Jianshuijinguan (Third)*. Shanghai: Zhongxi Book Company, 52.

Gao, H., Zhu, C., and Xu, W. (2007). Environmental Change and Cultural Response Around 4200 Cal. Yr BP in the Yishu River Basin, Shandong. *J. Geogr. Sci.* 17, 285–292. doi:10.1007/s11442-007-0285-5

Ge, J. X., Wu, S. D., and Chao, S. J. (1997). *Zhongguo Yimin Shi (The Migration History of China)*. Fuzhou: Fuzhou: Fujian People's Publishing House.

GPICRA and SAMPU (Gansu Provincial Institute of Cultural Relics and Archaeology and School of Archaeology and Museology of Peking University) (2011). *The Report of Prehistoric Archaeology Survey in the Hexi Corridor*. Beijing: Cultural Relics Publishing House.

GPICRA (Gansu Provincial Institute of Cultural Relics and Archaeology) (2001). *Excavation Report of Xigangchaiwangang Tomb in Yongchang*. Lanzhou: Gansu People's Publishing House.

Haug, G. H., Günther, D., Peterson, L. C., Sigman, D. M., Hughen, K. A., and Aeschlimann, B. (2003). Climate and the Collapse of Maya Civilization. *Science* 299, 1731–1735. doi:10.1126/science.1080444

Hu, K., Yan, S., Liu, K., Ning, C., Wei, L. H., Li, S. L., et al. (2015). The Dichotomy Structure of Y Chromosome Haplogroup N. *arXiv*. arXiv:1504.06463.

Ilumäe, A.-M., Reidla, M., Chukhryaeva, M., Järve, M., Post, H., Karmin, M., et al. (2016). Human Y Chromosome Haplogroup N: A Non-trivial Time-Resolved Phylogeography that Cuts across Language Families. *Am. J. Hum. Genet.* 99 (1), 163–173. doi:10.1016/j.ajhg.2016.05.025

Jia, X., Sun, Y., Wang, L., Sun, W., Zhao, Z., Lee, H. F., et al. (2016). The Transition of Human Subsistence Strategies in Relation to Climate Change during the Bronze Age in the West Liao River Basin, Northeast China. *The Holocene* 26 (5), 781–789. doi:10.1177/0959683615618262

Jin, H.-J., Tyler-Smith, C., and Kim, W. (2009). The Peopling of Korea Revealed by Analyses of Mitochondrial DNA and Y-Chromosomal Markers. *PloS one* 4 (1), e4210. doi:10.1371/journal.pone.0004210

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., and Orlando, L. (2013). mapDamage2.0: Fast Approximate Bayesian Estimates of Ancient DNA Damage Parameters. *Bioinformatics (Oxford, England)* 29 (13), 1682–1684. doi:10.1093/bioinformatics/btt193

Karafet, T. M., Hallmark, B., Cox, M. P., Sudoyo, H., Downey, S., Lansing, J. S., et al. (2010). Major East-West Division Underlies Y Chromosome Stratification across Indonesia. *Mol. Biol. Evol.* 27 (8), 1833–1844. doi:10.1093/molbev/msq063

King, C. L., Bentley, R. A., Tayles, N., Viðarsdóttir, U. S., Nowell, G., and Macpherson, C. G. (2013). Moving Peoples, Changing Diets: Isotopic Differences Highlight Migration and Subsistence Changes in the Upper Mun River Valley, Thailand. *J. Archaeological Sci.* 40, 1681–1688. doi:10.1016/J.JAS.2012.11.013

Klales, A. R., Ousley, S. D., and Vollner, J. M. (2012). A Revised Method of Sexing the Human Innominate Using Phenice's Nonmetric Traits and Statistical Methods. *Am. J. Phys. Anthropol.* 149 (1), 104–114. doi:10.1002/ajpa.22102

Knapp, M., Clarke, A. C., Horsburgh, K. A., and Matisoo-Smith, E. A. (2012). Setting the Stage - Building and Working in an Ancient DNA Laboratory. *Ann. Anat. - Anatomischer Anzeiger* 194 (1), 3–6. doi:10.1016/j.aanat.2011.03.008

Ko, A. M.-S., Chen, C.-Y., Fu, Q., Delfin, F., Li, M., Chiu, H.-L., et al. (2014). Early Austronesians: into and Out of Taiwan. *Am. J. Hum. Genet.* 94 (3), 426–436. doi:10.1016/j.ajhg.2014.02.003

Kong, Q.-P., Yao, Y.-G., Sun, C., Bandelt, H.-J., Zhu, C.-L., and Zhang, Y.-P. (2003). Phylogeny of East Asian Mitochondrial DNA Lineages Inferred from Complete Sequences. *Am. J. Hum. Genet.* 73 (3), 671–676. doi:10.1086/377718

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 (1), 85–98. doi:10.1007/s00439-016-1742-y

Li, H., and Durbin, R. (2010). Fast and Accurate Long-Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 26 (5), 589–595. doi:10.1093/bioinformatics/btp698

Li, H., Liu, Z., James, N., Li, X., Hu, Z., Shi, H., et al. (2021). Agricultural Transformations and Their Influential Factors Revealed by Archaeobotanical Evidence in Holocene Jiangsu Province, Eastern China. *Front. Earth Sci.* 9, 661–684. doi:10.3389/feart.2021.661684

Li, T. Y. (2018). *A Study about the Garrisons' Native Place at Zhangye Area in Bamboo Slips of Han Dynasty*. Master's thesis. Changchun: Jilin University.

Li, X. (2021). *Human Diets and its Influencing Factors during Han and Jin Periods in the Hexi Corridor and its Adjacent Areas*. Dissertation. Lanzhou: Lanzhou University.

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

Liu, Y. J. (2012). *Study of Immigrants of Hexi in Han Dynasty*. Master's thesis. Lanzhou: Northwest Normal University.

Luo, X. Q., Du, P. X., Wang, L. X., Zhou, B. Y., Li, Y. C., Zheng, H. X., et al. (2019). Uniparental Genetic Analyses Reveal the Major Origin of Fujian Tanka from Ancient Indigenous Daic Populations. *Hum. Biol.* 91 (4), 257–277. doi:10.13110/humanbiology.91.4.05

Ma, M., Dong, G., Jia, X., Wang, H., Cui, Y., and Chen, F. (2016). Dietary Shift after 3600 Cal Yr BP and its Influencing Factors in Northwestern China: Evidence from Stable Isotopes. *Quat. Sci. Rev.* 145, 57–70. doi:10.1016/j.quascirev.2016.05.041

Ma, P., Yang, X., Yan, S., Li, C., Gao, S., Han, B., et al. (2021). Ancient Y-DNA with Reconstructed Phylogeny Provide Insights into the Demographic History of Paternal Haplogroup N1a2-F1360. *J. Genet. Genomics* 48 (12), 1130–1133. doi:10.1016/j.jgg.2021.07.018

Meyer, M., and Kircher, M. (2010). Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harb Protoc.* 2010 (6), pdb.prot5448–prot5448. doi:10.1101/pdb.prot5448

Nu´ñez, L., Grosjean, M., and Cartajena, I. (2002). Human occupations and climate change in the Puna de Atacama, Chile. *Science* 298, 821–824. doi:10.1126/science.1076449

Oota, H., Settheetham-Ishida, W., Tiwawech, D., Ishida, T., and Stoneking, M. (2001). Human mtDNA and Y-Chromosome Variation Is Correlated with Matrilocal versus Patrilocal Residence. *Nat. Genet.* 29 (1), 20–21. doi:10.1038/ng711

Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., et al. (2016). EAGER: Efficient Ancient Genome Reconstruction. *Genome Biol.* 17, 60. doi:10.1186/s13059-016-0918-z

Petraglia, M. D., Groucutt, H. S., Guagnin, M., Breeze, P. S., and Boivin, N. (2020). Human Responses to Climate and Ecosystem Change in Ancient Arabia. *Proc. Natl. Acad. Sci. USA* 117, 8263–8270. doi:10.1073/pnas.1920211117

Pokharia, A. K., Agnihotri, R., Sharma, S., Bajpai, S., Nath, J., Kumaran, R. N., et al. (2017). Altered Cropping Pattern and Cultural Continuation with Declined prosperity Following Abrupt and Extreme Arid Event at ~4,200 Yrs BP: Evidence from an Indus Archaeological Site Khirsara, Gujarat, Western India. *PloS one* 12 (10), e0185684. doi:10.1371/journal.pone.0185684

Polyak, V. J., and Asmerom, Y. (2001). Late Holocene Climate and Cultural Changes in the Southwestern United States. *Science* 294, 148–151. doi:10.1126/science.1062771

Preston, G. W. (2011). *From Nomadic Herder-hunters to Sedentary Farmers: The Relationship between Climate, Environment and Human Societies in the united arab emirates from the Neolithic to the Iron Age*. Dissertation. Oxford: Oxford Brookes University.

Preston, G. W., Parker, A. G., Walkington, H., Leng, M. J., and Hodson, M. J. (2012). From Nomadic Herder-hunters to Sedentary Farmers: the Relationship between Climate Change and Ancient Subsistence Strategies in South-Eastern Arabia. *J. Arid Environments* 86, 122–130. doi:10.1016/j.jaridenv.2011.11.030

Qi, C. (1983). *The Rise and Fall of Ancient Hexi. The Archaeology Team of Silk Road*. Lanzhou: Gansu people's Publishing House.

Ralf, A., Montiel González, D., Zhong, K., and Kayser, M. (2018). Yleaf: Software for Human Y-Chromosomal Haplogroup Inference from Next-Generation Sequencing Data. *Mol. Biol. Evol.* 35 (5), 1291–1294. doi:10.1093/molbev/msy032

Renaud, G., Slon, V., Duggan, A. T., and Kelso, J. (2015). Schmutzi: Estimation of Contamination and Endogenous Mitochondrial Consensus Calling for Ancient DNA. *Genome Biol.* 16, 224. doi:10.1186/s13059-015-0776-0

Sawyer, S., Renaud, G., Viola, B., Hublin, J. J., Gansauge, M. T., Shunkov, M. V., et al. (2015). Nuclear and Mitochondrial DNA Sequences from Two Denisovan Individuals. *Proc. Nat. Acad. Sci. USA* 112 (51), 15696–15700. doi:10.1073/pnas.1519905112

Shao-Qing, W., Ruo-Yu, B., Bo-Yan, Z., Pan-Xin, D., Chang, S., Liang, C., et al. (2019). China National DNA Martyry: a beacon of hope for the Martyrs' Coming home. *J. Hum. Genet.* 64 (10), 1045–1047. doi:10.1038/s10038-019-0649-6

Sherratt, A. (1981). "Plough and Pastoralism: Aspects of the Secondary Products Revolution," in *Pattern of the Past: Studies in Honour of David Clarke*. Editors I. Hodder, G. Isaac, and N. Hammond (Cambridge: Cambridge University Press), 261–305.

Shi, Z., Chen, T., Storozum, M. J., and Liu, F. (2019). Environmental and Social Factors Influencing the Spatiotemporal Variation of Archaeological Sites during the Historical Period in the Heihe River basin, Northwest China. *Quat. Int.* 507, 34–42. doi:10.1016/j.quaint.2018.12.016

Si, M. Q. (2002). *Shiji. The Biographies of Dawan*. Changsha: Yuelu Publishing House, 697–698. doi:10.1007/s00106-002-0694-9

Sun, N., Ma, P.-C., Yan, S., Wen, S.-Q., Sun, C., Du, P.-X., et al. (2019). Phylogeography of Y-Chromosome Haplogroup Q1a1a-M120, a Paternal Lineage Connecting Populations in Siberia and East Asia. *Ann. Hum. Biol.* 46 (3), 261–266. doi:10.1080/03014460.2019.1632930

Sun, X.-f., Wen, S.-q., Lu, C.-q., Zhou, B.-y., Curnoe, D., Lu, H.-y., et al. (2021). Ancient DNA and Multimethod Dating Confirm the Late Arrival of Anatomically Modern Humans in Southern China. *Proc. Natl. Acad. Sci. USA* 118 (8), e2019158118. doi:10.1073/pnas.2019158118

Sun, Z., and Liu, S. (2014). Study of Hexi Diet from Unearthed Slips on the Han Dynasty. *Gansu Soc. Sci.* 000 (006), 90–94.

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration. *Brief. Bioinform.* 14, 178–192. doi:10.1093/bib/bbs017

Trejaut, J. A., Poloni, E. S., Yen, J.-C., Lai, Y.-H., Loo, J.-H., Lee, C.-L., et al. (2014). Taiwan Y-Chromosomal DNA Variation and its Relationship with Island Southeast Asia. *BMC Genet.* 15 (1), 77–23. doi:10.1186/1471-2156-15-77

Volodko, N. V., Starikovskaya, E. B., Mazunin, I. O., Eltsov, N. P., Naidenko, P. V., Wallace, D. C., et al. (2008). Mitochondrial Genome Diversity in Arctic Siberians, with Particular Reference to the Evolutionary History of Beringia and Pleistocenic Peopling of the Americas. *Am. J. Hum. Genet.* 82 (5), 1084–1100. doi:10.1016/j.ajhg.2008.03.019

Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.-J., et al. (2016). HaploGrep 2: Mitochondrial Haplogroup Classification in the Era of High-Throughput Sequencing. *Nucleic Acids Res.* 44 (W1), W58–W63. doi:10.1093/nar/gkw233

Wen, B., Li, H., Gao, S., Mao, X., Gao, Y., Li, F., et al. (2005). Genetic Structure of Hmong-Mien Speaking Populations in East Asia as Revealed by mtDNA Lineages. *Mol. Biol. Evol.* 22 (3), 725–734. doi:10.1093/molbev/msi055

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

Wen, B., Xie, X., Gao, S., Li, H., Shi, H., Song, X., et al. (2004b). Analyses of Genetic Structure of Tibeto-Burman Populations Reveals Sex-Biased Admixture in Southern Tibeto-Burmans. *Am. J. Hum. Genet.* 74 (5), 856–865. doi:10.1086/386292

Wen, S.-Q., Tong, X.-Z., and Li, H. (2016). Y-chromosome-based Genetic Pattern in East Asia Affected by Neolithic Transition. *Quat. Int.* 426, 50–55. doi:10.1038/s10038-020-0775-110.1016/j.quaint.2016.03.027

Wu, Q., Cheng, H.-Z., Sun, N., Ma, P.-C., Sun, J., Yao, H.-B., et al. (2020). Phylogenetic Analysis of the Y-Chromosome Haplogroup C2b-F1067, a Dominant Paternal Lineage in Eastern Eurasia. *J. Hum. Genet.* 65, 823–829. doi:10.1038/s10038-020-0775-1

Xu, K., and Hu, S. (2015). Population Data of Mitochondrial DNA HVS-I and HVS-II Sequences for 208 Henan Han Chinese. *Leg. Med.* 17 (4), 287–294. doi:10.1016/j.legalmed.2015.02.003

Yan, S., Wang, C. C., Zheng, H. X., Wang, W., Qin, Z. D., Wei, L. H., et al. (2014). Y Chromosomes of 40% Chinese Descend from Three Neolithic Super-grandfathers. *PloS one* 9 (8), e105691. doi:10.1371/journal.pone.0105691

Yan, S., Wang, C. C., Wang, C.-C., Li, H., Li, S.-L., and Jin, L. (2011). An Updated Tree of Y-Chromosome Haplogroup O and Revised Phylogenetic Positions of Mutations P164 and PK4. *Eur. J. Hum. Genet.* 19 (9), 1013–1015. doi:10.1038/ejhg.2011.64

Yang, J. (2015). *Study on Food Culture in Hexi Areas of Han to Jin Dynasties from Archaeological Perspective* (Lanzhou: Northwest Normal University). [Dissertation].

Yang, L., Shi, Z., Zhang, S., and Lee, H. F. (2020). Climate Change, Geopolitics, and Human Settlements in the Hexi Corridor over the Last 5,000 Years. *Acta Geologica Sinica - English Edition* 94, 612–623. doi:10.1111/1755-6724.14529

Yang, Y., Ren, L., Dong, G., Cui, Y., Liu, R., Chen, G., et al. (2019a). Economic Change in the Prehistoric Hexi Corridor (4800-2200bp), North-West China. *Archaeometry* 61, 957–976. doi:10.1111/arcm.12464

Yang, Y., Zhang, S., Oldknow, C., Qiu, M., Chen, T., Li, H., et al. (2019b). Refined Chronology of Prehistoric Cultures and its Implication for Re-evaluating Human-Environment Relations in the Hexi Corridor, Northwest China. *Sci. China Earth Sci.* 62, 1578–1590. doi:10.1007/S11430-018-9375-4

Yao, Y.-G., Kong, Q.-P., Bandelt, H.-J., Kivisild, T., and Zhang, Y.-P. (2002). Phylogeographic Differentiation of Mitochondrial DNA in Han Chinese. *Am. J. Hum. Genet.* 70 (3), 635–651. doi:10.1086/338999

Zhang, H. L. (2017). *A Study on Han and Jin Tombs in the Luoyang Area* (Zhengzhou: Zhengzhou University). [Dissertation].

Zhang, M. (2021). *Late Pleistocene and Early Holocene Microblade-Based Industries in Northeastern Asia: A Macroecological Approach to Foraging Societies*. Oxford: British Archaeological Reports.

Zhang, X., Liao, S., Qi, X., Liu, J., Kampuansai, J., Zhang, H., et al. (2015). Y-chromosome Diversity Suggests Southern Origin and Paleolithic Backwave Migration of Austro-Asiatic Speakers from Eastern Asia to the Indian Subcontinent. *Sci. Rep.* 5 (1), 1–8. doi:10.1038/srep15486

Zheng, B., and Gao, G. (2019). *The Murals Unearthed from Wei, Jin and Tang Tombs in Gansu Province*. Lanzhou: Gansu University Press.

Zhong, H., Shi, H., Qi, X.-B., Xiao, C.-J., Jin, L., Ma, R. Z., et al. (2010). Global Distribution of Y-Chromosome Haplogroup C Reveals the Prehistoric Migration Routes of African Exodus and Early Settlement in East Asia. *J. Hum. Genet.* 55 (7), 428–435. doi:10.1038/jhg.2010.40

Zhou, X., Li, X., Dodson, J., and Zhao, K. (2016). Rapid Agricultural Transformation in the Prehistoric Hexi Corridor, China. *Quat. Int.* 426, 33–41. doi:10.1016/j.quaint.2016.04.021

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 | Frontiers in Genetics

ORIGINAL RESEARCH
published: 27 May 2022
doi: 10.3389/fgene.2022.909267

Ancient Mitogenomes Reveal the Origins and Genetic Structure of the Neolithic Shimao Population in Northern China

Jiayang Xue[1,2†], Wenjun Wang[1,3†], Jing Shao[4†], Xiangming Dai[5†], Zhouyong Sun[4], Jacob D. Gardner[1], Liang Chen[6], Xiaoning Guo[4], Nan Di[4], Xuesong Pei[4], Xiaohong Wu[7], Ganyu Zhang[1], Can Cui[1,2], Peng Cao[1], Feng Liu[1], Qingyan Dai[1], Xiaotian Feng[1], Ruowei Yang[1], Wanjing Ping[1,2], Lizhao Zhang[1], Nu He[8] and Qiaomei Fu[1,2,9]**

[1]Key Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology, Center for Excellence in Life and Paleoenvironment, Chinese Academy of Sciences, Beijing, China, [2]University of Chinese Academy of Sciences, Beijing, China, [3]Science and Technology Archaeology, National Centre for Archaeology, Beijing, China, [4]Shaanxi Academy of Archaeology, Xi'an, China, [5]Archaeology Institute of National Museum of China, Beijing, China, [6]School of Cultural Heritage, Northwest University, Xi'an, China, [7]School of Archaeology and Museology, Peking University, Beijing, China, [8]Institute of Archaeology, Chinese Academy of Social Sciences, Beijing, China, [9]Shanghai Qi Zhi Institute, Shanghai, China

Shimao City is considered an important political and religious center during the Late Neolithic Longshan period of the Middle Yellow River basin. The genetic history and population dynamics among the Shimao and other ancient populations, especially the Taosi-related populations, remain unknown. Here, we sequenced 172 complete mitochondrial genomes, ranging from the Yangshao to Longshan period, from individuals related to the Shimao culture in northern Shaanxi Province and Taosi culture in southern Shanxi Province, Middle Yellow River basin. Our results show that the populations inhabiting Shimao City had close genetic connections with an earlier population in the Middle Neolithic Yangshao period of northern Shaanxi Province, revealing a mostly local origin for the Shimao Society. In addition, among the populations in other regions of the Yellow River basin, the Shimao-related populations had the closest maternal affinity with the contemporaneous Taosi populations from the Longshan period. The Shimao-related populations also shared more affinity with present-day northern Han populations than with the minorities and southern Han in China. Our study provides a new perspective on the genetic origins and structure of the Shimao people and the population dynamics in the Middle Yellow River basin during the Neolithic period.

Keywords: Shimao, mitochondrial genome, Ancient DNA, Yellow River, Neolithic

OPEN ACCESS

Edited by:
Shaoqing Wen,
Fudan University, China

Reviewed by:
Yinqiu Cui,
Jilin University, China
Hong-Xiang Zheng,
Fudan University, China

****Correspondence:***
Qiaomei Fu
fuqiaomei@ivpp.ac.cn
Nu He
henu@cass.org.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: *31 March 2022*
Accepted: *25 April 2022*
Published: *27 May 2022*

Citation:
Xue J, Wang W, Shao J, Dai X, Sun Z, Gardner JD, Chen L, Guo X, Di N, Pei X, Wu X, Zhang G, Cui C, Cao P, Liu F, Dai Q, Feng X, Yang R, Ping W, Zhang L, He N and Fu Q (2022) Ancient Mitogenomes Reveal the Origins and Genetic Structure of the Neolithic Shimao Population in Northern China. Front. Genet. 13:909267. doi: 10.3389/fgene.2022.909267

Abbreviations: BP, before present; CWE, central-west Eurasians; DM, Dongmen; HCT, Huangchengtai; LN, late Neolithic; MN, middle Neolithic; MZZL, Muzhuzhuliang; mtDNA, Mitochondrial DNA; NC, Neicheng; NEAs, Northeastern Asians; PCA, principal component analysis; SEAs, Southeastern Asians; SGDL, Shengedaliang; TS, Taosi; XH, Xinhua; YR, Yellow River; ZS, Zhaishan; ZJZ, Zhoujiazhuang.

INTRODUCTION

Northern China is a large geographic region that encompasses the Yellow River (YR) basin, in which the residing Neolithic cultures (such as the Yangshao and Longshan cultures) laid an important foundation for the origin of Chinese civilization (Wang, 1989; Dong et al., 2016; Zhou, 2017). The Middle Neolithic (MN) Yangshao period (~7,000–5,000 years before present, BP) was a stage of rapid development and expansion, giving rise to the Majiayao culture (~5,700 BP) in the Upper YR (Yang, 2016), the Dahecun culture in the Middle YR (~5,700 BP) (Xu, 2004), and the Beixin (~5,400 BP) and Dawenkou cultures (~6,000 BP) in the Lower YR (Zhang, 1996; Wang, 2005; Zhang, 2015; Dong et al., 2016; Duan, 2019). This cultural development and expansion coincided with the Holocene Climatic Optimum period in northern China (Hou et al., 2019). In the Late Neolithic (LN) Longshan period (~4,500–3,800 BP), the cultural features in different regions (~4,300 BP in Shaanxi Province; ~4,400 BP in Henan Province; ~4,500 BP in Shanxi Province) of the YR basin varied spatially, increased in social complexity, and formed distinct settlements with different levels of social hierarchy (He, 2004; Chang, 2005; Sun, 2016). The influences of different archaeological cultures on various regions of the YR basin changed dynamically over time, which might have been accompanied by population flow and interaction (Sun et al., 2020a).

The Shimao site (~4,300–3,800 BP), also called "Shimao City", is considered an important political and religious center during the Middle YR's Longshan period (~4,500–3,800 BP) (Sun et al., 2013; Rawson, 2017; Sun et al., 2020a). It is currently the largest Neolithic settlement known in China, covering 4 km^2 with a triple structure made of stone-reinforced walls (**Figure 1**), and was selected as one of the world's top 10 archaeological discoveries in the past decade (Archaeology Institute of America, 2021). The center of Shimao City, Huangchengtai, has many high-grade buildings and relics (Sun et al., 2020a; Sun et al., 2020b). The Neicheng (or "inner city") surrounds Huangchengtai and consists of multiple grave sites (e.g., Hanjiagedan, Houyangwan, and Mahuangliang). The Dongmen (or "East Gate") is located along the northeastern wall of Waicheng (or "outer city") and exhibits complex fortifications (Sun et al., 2020a). According to archaeological records, these distinct sites within Shimao City showed clear differences in social hierarchy and inequality. For example, Hanjiagedan (Sun et al., 2016), Houyangwan (Sun and Shao, 2015), and those closer to Huangchengtai, yielded more high-status graves than Dongmen (Sun and Shao, 2015). Archaeologists named the "Shimao culture" based on artifacts unearthed in Shimao City (Dai, 1998; Sun et al., 2020b). The sites neighboring Shimao City in northern Shaanxi Province, such as the Muzhuzhuliang (Wang et al., 2015), Shengedaliang (Guo et al., 2016), Xinhua (Xing et al., 2002), and Zhaishan sites

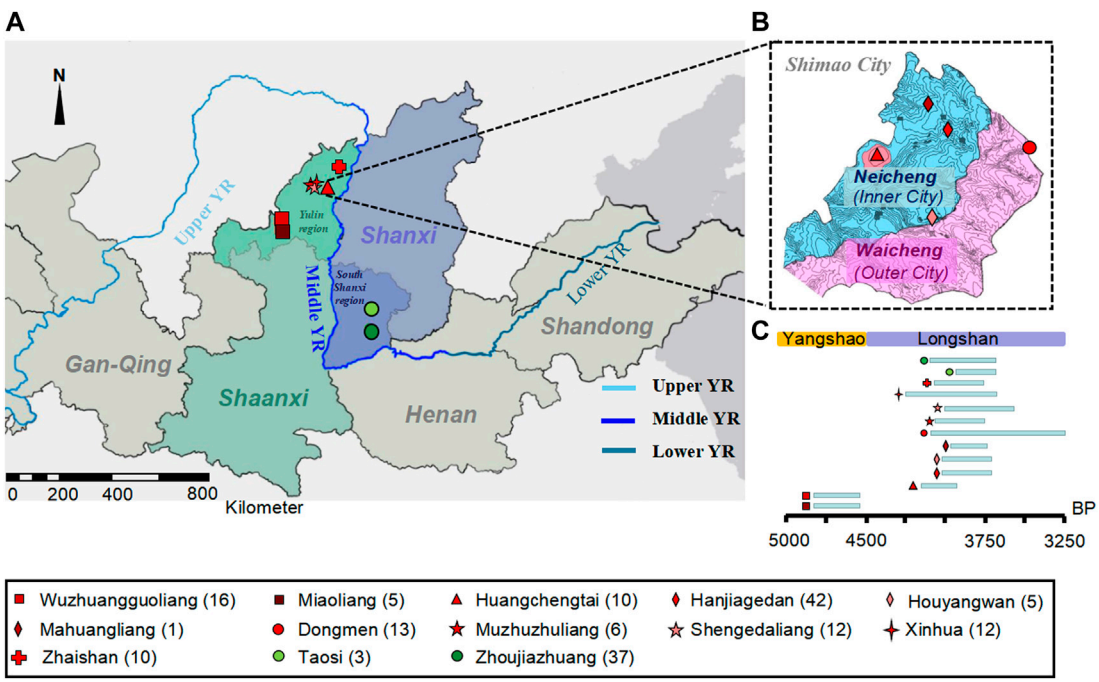


FIGURE 1 | The geographical and temporal distributions of the new samples. **(A)** The geographical location of ancient individuals from 13 archaeological sites in northern Shaanxi and southern Shanxi Province of Middle YR. The different colored and shaped labels represent different sites. The dark red represents those from the MN Yangshao sites, and red represents Shimao-related populations in and neighboring Shimao City in LN Longshan period in northern Shaanxi Province; green represents the Taosi-related populations in southern Shanxi Province. **(B)** The geographical distribution of ancient individuals in Shimao City. **(C)** Timeline of archaeological sites.

(Shao et al., 2021), were all attributed to the Shimao culture. However, the origin of Shimao City was still uncertain. The Shimao culture was considered to have developed from local populations with an influence from surrounding cultures; however, it may have alternatively originated from the migration of populations from the Central Plain or other regions (Dai, 1998; Zhang, 2004; Sun, 2005). In addition, recent studies revealed that the Shimao culture interacted frequently with other regions in the YR basin outside northern Shaanxi Province during the Neolithic period (Guo, 2013; Chen et al., 2016; Chen et al., 2017; Sun et al., 2020b), especially the Taosi culture in southern Shanxi Province of the Middle YR (Yan and He, 2005; Tian and Dai, 2018; Shao, 2020). The links between these two cultures may have been political, economic, cultural, or through shared population connections (Shao, 2020). However, the interactions between the Shimao and Taosi people remain ambiguous from the perspective of archaeology and physical anthropology (Yan and He, 2005; Guo, 2013; Chen et al., 2016; Chen et al., 2017; Tian and Dai, 2018; Sun et al., 2020b; Shao, 2020). Although there were some genomic analyses that included samples from the Shengedaliang (Ning et al., 2020) and Wuzhuangguoliang sites in northern Shaanxi Province (Zhao et al., 2017; Wang et al., 2021), the large-scale genetic affinities among the populations related to the Shimao culture and their predecessors, along with other populations in different regions of the YR basin, is still unclear.

In the current study, we sequenced the complete mitochondrial genomes of 172 samples from various archaeological sites, particularly the individuals related to Shimao and Taosi cultures in the Middle YR. Our research presents large-scale mitogenomic data and new perspectives for exploring the maternal genetic history and dynamics of Shimao-related populations and the populations in the Middle YR basin during the Neolithic period.

MATERIALS AND METHODS

Ancient DNA Extraction and Library Preparation

We collected samples from a total of 172 ancient human individuals from 13 sites. We describe their archaeological details in the Supplementary Materials (**Supplementary Table S1**). A total of 172 DNA extractions were obtained from less than 100 mg of bones or dental remains of ancient samples. All ancient DNA work was conducted in the aDNA clean room at the Institute of Vertebrate Paleontology and Paleoanthropology, following strict aDNA standards (Gilbert et al., 2005).

We prepared single-stranded (denoted as "SS") and double-stranded (denoted as "DS" in **Supplementary Table S1**) libraries and partially treated them with uracil-DNA glycosylase ("UDG") to remove deaminated cytosine (Kircher et al., 2021; Meyer et al., 2012). Libraries were amplified for 35 cycles using AccuPrime Pfx DNA polymerase under conditions described in a previous study (Dabney and Meyer, 2012). P5 and P7 adapters were added to limit the contamination rate (Kircher et al., 2012). And the NanoDrop2000 spectrometer was used for monitoring the DNA concentration.

Ancient DNA Capture and Sequencing

To enrich endogenous ancient DNA from the high levels of background environmental DNA, we used a DNA capture technique (Fu et al., 2013; Fu et al., 2015; Haak et al., 2015). The in-solution capture of the mitochondrial DNA (mtDNA) was accomplished by overlapping probes with DNA fragments and enriching the resulting libraries (Fu et al., 2013). The probes were synthesized based on the human mitochondrial genome.

After enrichment, the Illumina Miseq platform was used to generate 2 × 76 bp paired-end reads. The leeHom software (https://github.com/grenaud/leeHom) was used to trim adapters and merge sequences, with paired-end reads overlapping by at least 11 bp (Gabriel et al., 2014). Sequenced and merged reads with lengths of at least 30 bp were then mapped to the revised Cambridge Reference Sequence version 17 (rCRS, Genbank accession number NC_012920) for mtDNA, using the same command in the BWA v0.6.1 aligner (arguments used: -n 0. 01 and –l16500) (Li and Durbin, 2009; Renaud et al., 2014). We removed duplicate sequences and retained the one with highest mapping quality. After removing sequences with a mapping quality below 30, we constructed the whole mitochondrial sequence (**Supplementary Table S1**).

Test for Contamination

We evaluated the contamination rate using the ContamMix software and compared the mtDNA fragments with the consensus mitochondrial genome for our newly sampled individuals and 311 present-day world-wide sequences (Fu et al., 2013). We treated the libraries as contaminated if over 4% of the fragments matched with other sequences that are better than the consensus (Fu et al., 2016). For the libraries with substantial contamination (contamination rate > 4%), we excluded them (Briggs et al., 2007; Rohland et al., 2015). Of the 172 new mtDNA samples, 166 of them have lower contamination rates (< 4%, average 0.95%) (**Supplementary Table S1**).

Kinship Analysis

We treated the mtDNA sequences as kinship-related individuals if they shared identical mitochondrial genomic sequences and were from the same tombs. The Bioedit software (version 7.2.5) was used for testing the kinship. Finally, we found four pairs of sequences with probable matrilineal kinships (Hall, 1999). We excluded the samples in each pair that had relatively lower coverage (**Supplementary Table S1**).

In total, we sequenced the complete mitochondrial genomes of 172 individuals. After removing those with higher contamination rates and kinships, 162 individuals sequenced to between 20.74-fold and 827.53-fold coverage (average 253.95-fold) were used for analysis (**Supplementary Table S1**).

Haplogroup Analysis

MUSCLE (MUSCLE 3.8.31) and Bioedit software were used to align and edit the complete sequences of mtDNA with rCRS

(Andrews et al., 1999; Edgar, 2004; Weissensteiner et al., 2016). Haplogrep2, built on Phylotree Build 17, was used to call haplogroups for each sample (van Oven, 2015; Weissensteiner et al., 2016) (**Supplementary Table S1**). We grouped all haplogroups absent from the populations in and around Shimao City as "Other". Since the haplogroups R and N distribute in both East and West Eurasian, we used $R^{\#}$ and $N^{\#}$ (such as haplotype R+16189, sub-haplogroups R11 and N9, which were also observed in the Shimao-related populations) to represent the haplogroups carried by East Eurasian in our dataset. The other sub-haplogroups R and N (such as sub-haplogroups R1, R2, N1, N2) found in West Eurasians, were assigned to "Other."

Principal Component Analysis and Haplogroup Sharing

In addition, we calculated haplogroup frequencies for every group and conducted a PCA using the built-in function "prcomp" in R (version 4.1.2) software (Venables and Ripley, 2002). We plotted PC1 and PC2 to illustrate the haplogroup differences among populations and explore the maternal genetic relationship among the populations.

We calculated haplogroup sharing as a pairwise matrix of proportions of haplogroups shared among populations. In the matrix, the entries represent the proportions of haplogroups shared in the two populations, which are calculated by summing the shared frequencies of all the same haplogroups between them (Ko et al., 2014; Miao et al., 2021). The proportions of each population were normalized by dividing by the total count and summing to one (Ko et al., 2014).

Discriminant Analysis of Principal Components

We also used discriminant analysis of principal components (DAPC), which maximizes the inter-variation between groups while minimizing the intra-variation, to show the maternal genetic relationships among ancient populations in the YR basin (Pritchard et al., 2000). We used the DAPC function of the "adegenet" package in R (version 4.1.2) software to conduct a sequence–based DAPC (Jombart et al., 2010).

Genetic Distance Analysis

The Arlequin software package (version 3.5.2.2) was used to calculate genetic distances (F_{ST}) between populations (Excoffier et al., 2007), which were visualized with the package "pheatmap" in R. Generally, a lower F_{ST} represents a closer maternal genetic relationship between two groups. Heatmaps were also drawn to illustrate the statistical significance of clusters based on F_{ST}.

Haplotype Network Construction

To explore the genetic relationship of certain haplotypes in samples, we used DNASP6 (version 6.12.03) and PopArt 1.7 to conduct a median-joining network analysis on all the samples in the same haplogroup (sub-haplogroup or haplotype) dataset (Bandelt et al., 1999; Leigh and Bryant, 2015) and construct the haplotype network graph. This helped us understand the inflow or diffusion process of a haplotype population.

RESULTS

Sample and Ancient DNA Generation

We captured mitochondrial DNA (mtDNA) from 172 ancient individuals from 13 archaeological sites in the northern Shaanxi and southern Shanxi Provinces of the Middle YR (**Figure 1A**; **Supplementary Table S1**), with dates ranging from 4,836 to 3,253 calibrated BP (cal BP). Removing six individuals with high contamination rates (> 4%) and four individuals owning close relatives (defined as the same mtDNA sequences), resulted in a final dataset of 162 individuals with coverage between 20.74- and 827.53-fold (**Supplementary Table S1**).

Among these new samples, we obtained 21 samples from the Miaoliang and Wuzhuangguoliang sites (referred to as the "preShimao_MW" group) in northern Shaanxi Province, dating to 4,836–4,530 cal BP in the MN Yangshao period. Additionally, we obtained 91 samples from the LN Longshan period of northern Shaanxi Province, of which 66 were from Shimao City and 35 were from sites neighboring Shimao City (**Supplementary Table S1**). We grouped the individuals in Shimao City based on their archaeological cultures, dates, and geographical locations within Shimao: 10 individuals were excavated from the political and religious center, Huangchengtai site, which we named "Shimao_HCT" (4,148–3,895 cal BP); 44 individuals were from the Neicheng (or "inner city"), which we grouped as "Shimao_NC" (3,977–3,699 cal BP) and contained individuals from the Hanjiagedan, Houyangwan, and Mahuangliang sites; and 12 individuals from the Dongmen site in Waicheng (also called the "East Gate in outer city"), which we named "Shimao_DM" (4,144–3,253 cal BP). The individuals excavated from the Xinhua (XH, n = 9, 4,231–3,650 cal BP), Muzhuzhuliang (MZZL, n = 4, 4,082–3,722 cal BP), Shengedaliang (SGDL, n = 12, 3,969–3,570 cal BP), and Zhaishan (ZS, n = 10, ~4,050–3,750 BP) sites neighboring Shimao City, we named as the abbreviations of their site names. We integrated MZZL and SGDL into "MZZSGDL" (n = 16) for their similar archaeological cultures, locations, and dates, and the small population size (n = 4) from the MZZL site.

In addition, we sequenced mtDNA from 40 LN individuals excavated from the "TSZJZ" group (~4,150–3,696 BP) related to Taosi culture (containing Taosi and Zhoujiazhuang sites) in the southern Shanxi Province of the Middle YR.

We also collected 801 previously published mtDNA sequences for the ancient individuals from East and West Eurasia, ranging from the Early Neolithic (EN) to Historic Era (HE). These included populations from Xinjiang (~5,000–500 BP), Gansu and Qinghai Provinces (~5,040–411 BP), Henan Province (~5,500–5,000 BP, Qingtai site), Shandong Province (~9,600–2,000 BP), the Tibetan Plateau (~3,000–100 BP), southern East Asia (~4,600–300 BP), the Baikal River in southern Siberia (~7,123–6,319 BP and ~4,860–3,760 BP), Mongolia (~3,330–2,950 BP and ~2,147–2,007 BP), and the

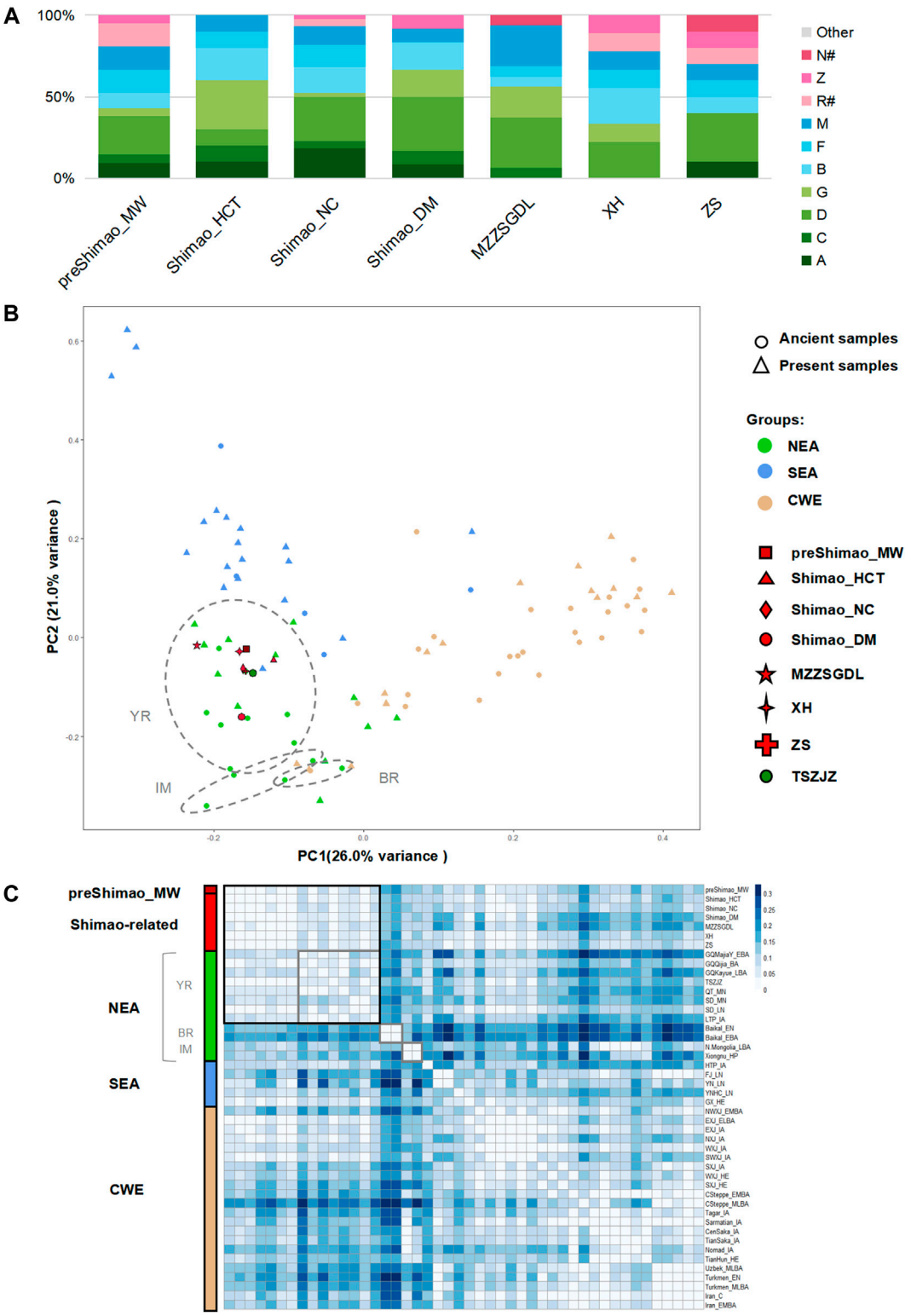


FIGURE 2 | The genetic analysis of ancient populations in northern Shaanxi Province. **(A)** Haplogroup frequency. The haplogroups with green are those common in northeastern Asians (NEAs), and those with blue are common in southeastern Asians (SEAs). The haplogroups absent from Shimao-related populations are grouped in "Other." The haplogroups $R^{\#}$ and $N^{\#}$ represent the haplotypes shown in East Eurasians (such as haplotype R+16189, sub-haplogroups R11 and N9, which were observed in the Shimao-related populations). **(B)** Principal Component Analysis (PCA) based on the haplogroup frequencies. The circle and triangle shapes represent ancient and present-day populations, respectively. The colored and shaped symbols correspond to **Figure 1A**. The grey circles represent the ancient

(*Continued*)

FIGURE 2 | populations from the Yellow River basin (YR), Mongolia and Inner Mongolia (IM), and the Baikal Lake region (BR) in NEA. CWE: Central and Western Eurasian. **(C)** The genetic distance (F_{ST}) heatmap of Shimao-related populations and other ancient populations. The different labels and colors correspond to the PCA plot, and different shades of color are used to mark different regional populations. Values with F_{ST} = 0.00 are in white, representing a close genetic relationship. SD_EN were excluded in the heatmap because of the significantly large genetic distance (F_{ST} > 0.10) between them and other populations. The grey squares represent the ancient populations from the Yellow River basin (YR), Mongolia and Inner Mongolia (IM), and the Baikal Lake region (BR) in NEA.

Steppe and West Eurasia (~5,450–1,500 BP). Meanwhile, we also obtained the haplogroup information from individuals in Inner Mongolia (~4,500 BP, Halahaigou site) (**Supplementary Table S2**).

For the present-day populations, we collected 7,641 individuals from northeastern Asians (NEAs, including North Asians and northern East Asians), southeastern Asians (SEAs, Southeast Asians and southern East Asians), and central-west Eurasians (CWEs). Among these populations, 2,102 individuals from China, including 388 Han individuals from northern China and 168 individuals from southern China, which we named "NChina_Han" and "SChina_Han", respectively. We also collected 548 individuals from 16 ethnic minorities, which covers the vast majority of ethnic minorities in China (**Supplementary Table S3**). We served them as different groups following their minorities. The present-day populations also contained the populations in Tibet ("SChina_Tibet") and Taiwan ("SChina_Taiwan") (Ko et al., 2014; Kang et al., 2016).

The Mostly Local Genetic Origin of Shimao Populations From Earlier Populations in Northern Shaanxi Province

To understand the genetic connection between the LN Shimao populations and the preceding populations in the MN period, we collected 21 individuals from the MN preShimao_MW sites and 66 individuals from the LN Shimao City in northern Shaanxi Province (**Figure 1**).

The haplogroup analysis found that ancient and present-day NEAs, showed a high proportion of haplogroups A (maximum, 71.43%), C (maximum, 55.00%), D (maximum, 60.00%), and G (maximum, 37.50%) with a north-south declining trend (**Supplementary Figure S1**; **Supplementary Tables S4, S5**). Haplogroups B (maximum, 36.36%, B4′5), F (maximum, 40.00%), and M (maximum, 83.33%) were common in ancient and present-day SEA and showed a north-south increasing trend (**Supplementary Figure S1**; **Supplementary Tables S4, S5**). The earlier population in the MN period of northern Shaanxi Province (4,836–4,530 cal BP), preShimao_MW, carried the haplogroups A (9.52%), C (4.76%), D (23.81%), G (4.76%), B (9.52%, B4′5), F (14.29%), M (14.29%), Z (4.76%), and $R^{\#}$ (14.29%), and showed a higher proportion of NEA (rather than SEA) haplogroups (**Figure 2A**). Our Principal Component Analysis (PCA) based on haplogroup frequency showed that the PC1 explains population variation from east to west geographically and that PC2 explains the variation from north to south (**Figure 2B**). In general, all the populations are genetically divided into three clusters: NEA, SEA, and CWEs. The preShimao_MW was distributed among the NEA populations and clustered with the populations in the YR (**Figure 2B**). In addition, this MN population showed the highest proportion of haplogroup D (23.81%), which was also found in relatively higher proportions in the YR populations (18.18–44.83%) (**Figure 2A**, **Supplementary Table S4**). The F_{ST} heatmap based on genetic distance also showed that the preShimao_MW clustered with the YR populations (**Figure 2C**). Thus, the MN Yangshao populations from northern Shaanxi Province (preShimao_MW) were more related to the NEA populations in the YR basin than to populations from other regions in East Asia. Although there were no significant genetic affinities between the preShimao_MW and the EN and MN YR populations (F_{ST} > 0.06, $p < 0.01$ with QT_MN and SD_MN; and $F_{ST} = 0.31$, $p > 0.07$ with SD_EN), the DAPC shows some overlap between the preShimao_MW and the QT_MN from the middle YR (**Figure 3A**; **Supplementary Table S6**). The same haplotypes G3a2, D5a2a1, and F1a1c were also observed in both preShimao_MW and QT_MN, suggesting some connections between them (**Supplementary Figure S2**; **Supplementary Tables S1, S2**).

For the LN Longshan populations (4,148–3,253 cal BP) in Shimao City, including Shimao_HCT, Shimiao_NC, and Shimao_DM, the haplogroup analysis showed that they carried similar haplogroups: A (8.33–18.18%), C (4.55–10.00%), D (10.00–33.33%), G (2.27–30.00%), B (15.91–20.00%, B4′5), and M (8.33–11.36%) (**Supplementary Table S4**). The Shimao populations also displayed higher proportions of NEA dominating haplogroups than SEA dominating haplogroups, and Shimao_DM (66.67%) showed a higher ratio of NEA dominating haplogroups than Shimao_HCT (60.00%) and Shimao_DM (52.27%). Moreover, Shimao_NC additionally carried haplogroup $R^{\#}$ (4.55%). Shimao_NC and Shimao_DM also had the highest proportions of haplogroup D (27.27–33.33%), similar to preShimao_MW and most YR populations (**Figure 2A**; **Supplementary Table S4**). The PCA shows that the three Shimao populations clustered and plot among the NEA populations in the YR region, consistent with the haplogroup analysis results (**Figures 2A,B**). Also, we found that the genetic distances (F_{ST} values) among the three Shimao populations were all about zero ($F_{ST} < 0.01$, $p > 0.05$), revealing a close genetic affinity among them (**Figure 2C**; **Supplementary Table S6**). The same haplotypes (B4a4, C4a2, G2a1, and G1c) were found in these three populations, further suggesting a close relationship among them (**Supplementary Figure S2**). These results all suggest that the populations in different regions of the LN Shimao City shared close affinities with each other.

We also explored the genetic connections between the LN Longshan populations (4,148–3,253 cal BP) in Shimao City and the earlier populations (before 4,500 BP) in and outside of

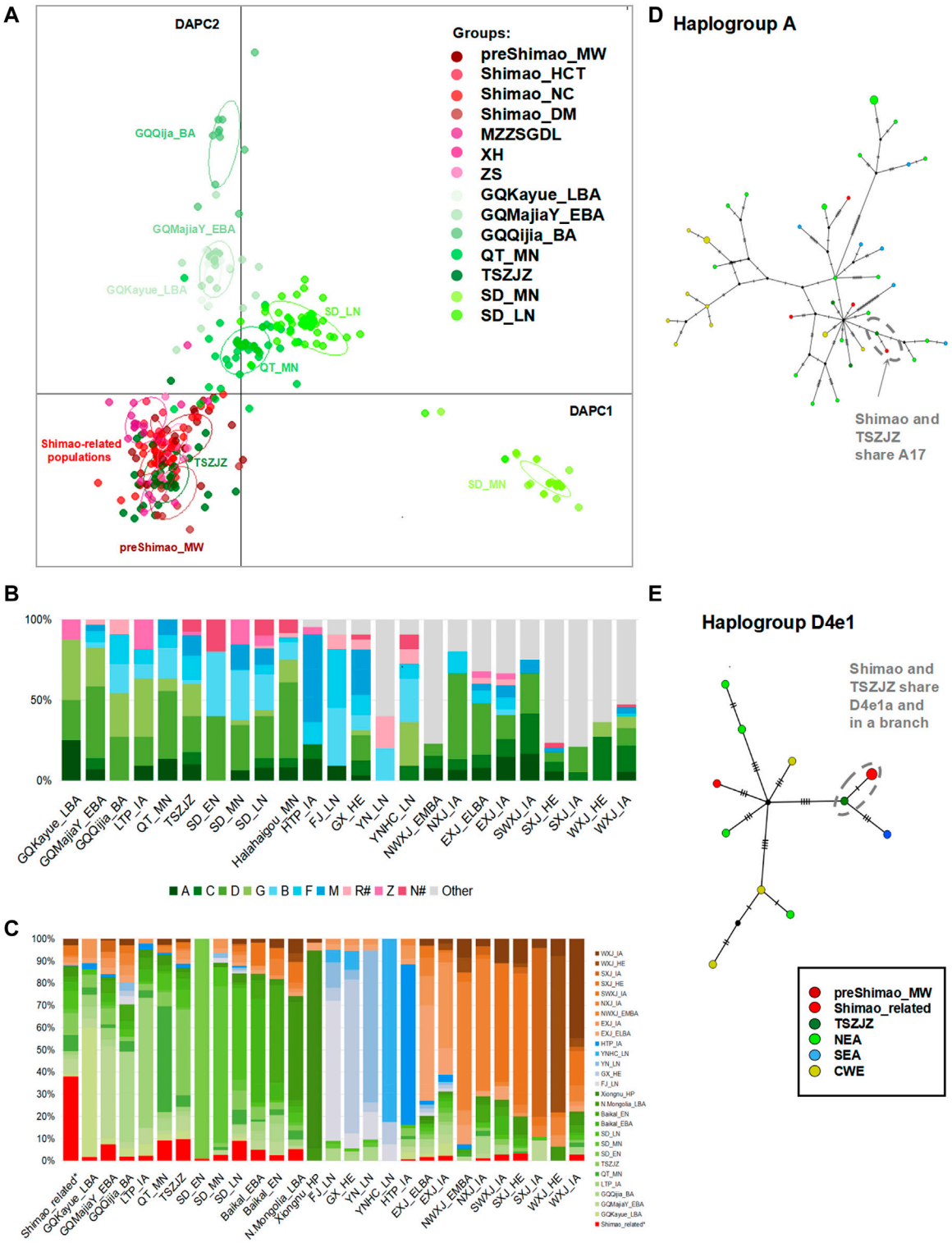


FIGURE 3 | The genetic analysis between Shimao-related populations and other ancient populations. **(A)** The discriminant analysis of principal components (DAPC) of ancient populations in YR. The red points represent the Shimao-related populations. The dark green points represent the TSZJZ individuals, and the other green points represent the individuals in YR. **(B)** Haplogroup frequency of ancient populations. The haplogroups with green are those common in NEAs, and those with blue are common in SEAs. The haplogroups absent from Shimao-related populations are grouped in "Other." The haplogroups $R^{\#}$ and $N^{\#}$ represent the haplotypes found in East Eurasians (such as haplotype R+16189, sub-haplogroups R11 and N9, which were also observed in the Shimao-related populations). The other sub-haplogroups R and N found in West Eurasians were assigned to "Other." **(C)** Haplogroup sharing analysis. The different colors correspond to F_{ST} heatmap and networks. (*Continued*)

FIGURE 3 | Median-joining networks of haplotypes A17 **(D)** and D4e1a **(E)** related to ancient northern Chinese populations. The different population groups are shown in different colors that are consistent with those groups in the F_{ST} heatmap.

northern Shaanxi Province. We found that the MN preShimao_MW and LN Shimao populations in northern Shaanxi carried similar haplogroups (A, C, D, G, Z, B, F, and M) and that Shimao_HCT and Shimao_DM both have the highest proportions of haplogroup D, in common with preShimao_MW (**Figure 2A**). Some of those haplogroups were absent from the early populations outside Shaanxi Province. For example, the QT_MN (~5,500–5,000 BP), SD_EN (~9,600–7,700 BP), and SD_MN (~5,500–4,600 BP) lacked haplogroup C, and SD_EN and SD_MN lacked haplogroup F (**Figure 3B**; **Supplementary Table S4**). The DAPC also indicates that the Shimao populations clustered with the preceding population (preShimao_MW) in (but not outside) northern Shaanxi Province (**Figure 3A**). Moreover, the three Shimao populations clustered with the preceding MN populations (preShimao_MW) in the F_{ST} heatmap and showed the smallest genetic differentiation between them ($F_{ST} < 0.01$, $p > 0.05$), while showing larger F_{ST} values with the early populations (QT_MN, SD_EN, and SD_MN) outside northern Shaanxi ($F_{ST} > 0.05$) (**Supplementary Table S6**). These results indicate that the LN Longshan populations in Shimao City showed the closest genetic affinity with the earlier MN Yangshao populations (preShimao_MW, 4,836–4,530 cal BP) in (but not outside) northern Shaanxi Province. This close relationship was also demonstrated by the shared haplotypes between preShimao_MW and Shimao-related populations, including D4j3 and D4b2b of D4 and haplotypes A+152 + 16362, F1a1c, and R11, and by plotting on the same branches of the median-joining network (**Supplementary Figure S2**). However, we also found some connections between the LN Shimao populations and the earlier QT_MN from the middle YR. This is supported by the appearance of haplotypes M9a1a1 and M10a1b, which differed in these two populations by only one and four mutations, respectively, in the network analysis (**Supplementary Figure S2**).

Therefore, the populations in different regions of the Longshan period's Shimao City (4,148–3,253 cal BP) shared close affinities with each other and with the preceding MN Yangshao populations (4,836–4,530 cal BP) in (rather than outside) northern Shaanxi Province. The results reveal that the MN Yangshao populations in northern Shaanxi Province were largely not replaced with the foundation of Shimao City, supporting a hypothesis of a mostly local genetic origin for the Shimao people. However, given the shared haplotypes with other YR populations (i.e., the MN Qingtai), we cannot rule out additional genetic contribution from populations outside northern Shaanxi province.

The Genetic Affinities Among Populations in and Around Shimao City in Northern Shaanxi

The archaeological reports indicated that the LN Longshan sites around Shimao City in northern Shaanxi Province, such as the MZZSGDL, XH, and ZS, were all related to the Shimao culture. To explore the genetic affinities among the contemporaneous LN populations related to Shimao culture (containing the populations in and around Shimao City), we sequenced 35 new samples from the sites neighboring Shimao City.

These three populations (MZZSGDL, XH, and ZS) around Shimao City primarily carried haplogroups D (22.22–31.25%), B (6.25–22.22%, B4′5), F (6.25–11.11%), and M (10.00–25.00%). Moreover, MZZSGDL also carried haplogroups C (6.25%) and G (18.75%), XH carried haplogroups G (11.11%), Z (11.11%), and $R^{\#}$ (11.11%), and ZS had haplogroups A (10.00%) and $R^{\#}$ (10.00%). All three populations displayed the highest proportions of haplogroup D (31.25% in MZZSGDL, 22.22% in XH, and 30.00% in ZS) (**Figure 2A**; **Supplementary Table S4**). In the PCA plot, these populations neighboring Shimao City plot among the NEA populations in YR region and closer to each other (**Figure 2B**). This close relationship was also revealed by the smaller genetic differentiation between them ($F_{ST} < 0.01$) (**Figure 2C**; **Supplementary Table S6**). In the median-joining networks, these populations shared the same branches in haplotype Z3, F2g, and M10a1a1b (**Supplementary Figure S2**). Thus, these three populations neighboring Shimao City were closely related.

In addition, we found that these populations neighboring Shimao City showed the same haplogroups (haplogroups D, B (B4′5), F, and M) as the populations in Shimao City, and some of them share haplotypes (A+152 + 16362, B4a4, D4j3, F2g, and Z3) of the same branch (**Figures 2A,D**; **Supplementary Figure S2**). We also found this close affinity in the DAPC and F_{ST} heatmap, as well as through the smaller genetic distance between them ($F_{ST} = 0.00$, $p > 0.05$ in most of them, and $F_{ST} = 0.04$, $p < 0.05$ between Shimao_NC and MZZSGDL) (**Figures 2C**, **3A**; **Supplementary Table S6**). Among the earlier populations (before 4,500 BP), the individuals neighboring Shimao City showed the closest relationship with the MN Yangshao populations (preShimao_MW) in (rather than outside) northern Shaanxi Province, similar to the populations in Shimao City, indicated by their shared haplogroups, close distributions in the DAPC, and the smaller F_{ST} values between them ($F_{ST} = 0.00$, $p > 0.05$ between preShimao_MW and XH, ZS; and $F_{ST} = 0.04$, $p = 0.03$ between preShimao_MW and MZZSGDL) (**Figures 2A,C**, **3A**; **Supplementary Table S6**). Similar with Shimao populations, the populations neighboring Shimao City (such as, MZZSGDL) indicated some slight connections with MN Qingtai supported by the same haplotype M9a1b, which only showed four-mutation differences between them in the network analysis (**Supplementary Figure S2**). Thus, these results show that the populations in and around Shimao City shared close genetic affinities.

Above all, during the LN Longshan period, the populations related to Shimao culture (containing the populations in and around Shimao City, which we call "Shimao-related

populations") showed close affinities with each other, revealing the extensive connections among the populations not only in but also around Shimao City in northern Shaanxi Province. All Shimao-related populations were shown to have a mostly local genetic origin from the preceding MN Yangshao populations in northern Shaanxi Province.

The Maternal Affinity Between the Populations Related to Shimao and the Contemporaneous Taosi Culture in the Middle Yellow River

Given the close genetic relationship between the MN Yangshao (4,836–4,530 cal BP) and LN Longshan period (4,231–3,253 cal BP) in northern Shaanxi Province, we then focused on the population interactions between the LN Shimao-related populations and the ancient humans outside northern Shaanxi Province. Previous archaeological studies had shown that stone carvings excavated from Shimao City shared cultural characteristics with the Shang Dynasty (3,500–2,900 BP) in the Central Plain of the Middle YR. Meanwhile, research on pottery from Shimao City found that the Shimao culture was closely associated with the contemporaneous Taosi culture in the southern Shanxi Province of the Middle YR (Shao, 2020; Sun and Shao, 2020). To explore the genetic relationship between Shimao-related populations and the populations in different regions of the YR, we sequenced 40 new individuals from TSZJZ related to the Taosi culture in the southern Shanxi Province of the Middle YR and collected 198 previously published ancient individuals from different regions of the YR.

We found that the Shimao-related populations shared more affinities with those NEA populations in the YR basin. Among those YR populations, the early Bronze Age (EBA) individuals in the Upper YR (GQMajiaY_EBA) and the LN individuals in the Middle (TSZJZ) and Lower YR (SD_LN), which date in and after the LN Longshan period (after 4,500 BP), carried higher proportions of haplogroups common in NEAs, such as haplogroups A (6.90–10.00%), C (6.00–7.50%), D (22.50–44.83%), and G (4.00–24.14%). Among these haplogroups, haplogroup D showed the highest proportion in these three YR populations (22.50% in TSZJZ, 26.00% in SD_LN, and 44.83% in GQMajiaY_EBA). In addition, they all have the haplogroups B (2.50–22.00%, B4′5), F (6.00–15.00%), and M (3.45–12.50%) (**Figure 3B**; **Supplementary Table S4**). These haplogroups were all found in the Shimao-related populations (**Figures 2A**, **3B**). In the DAPC plot, among the Middle YR populations, the Shimao-related populations were closer to the contemporaneous LN Longshan populations (after 4,500 BP) related to the Taosi culture (TSZJZ, 4,150–3,696 cal BP) than to the MN Yangshao populations (~5,500–5,000 BP) from the Qingtai site (QT_MN, ~5,500–5,000 BP) (**Figure 3A**). Among the Lower YR populations, the Shimao-related populations were closer to the LN Longshan populations (SD_LN, after 4,500 BP) than to the EN (SD_EN, ~9,600–7,700 BP) and MN (SD_MN, ~5,500–4,600 BP) individuals (**Figure 3A**). The DAPC suggests that the Shimao-related populations were closer to the contemporaneous LN (after 4,500 BP), rather than the EN and MN (before 4,500 BP) populations outside northern Shaanxi Province in the YR basin. The F_{ST} results also showed smaller genetic differentiation between Shimao-related individuals and the contemporaneous LN Longshan populations (F_{ST} = 0.00, p > 0.05 with TSZJZ; F_{ST} < 0.02, p > 0.02 between most Shimao-related populations and SD_LN; F_{ST} = 0.05, p = 0.00 between MZZSGDL and SD_LN) than with those earlier populations in the YR region (F_{ST} = 0.04–0.07, p < 0.05 with QT_MN; F_{ST} = 0.04–0.80, p < 0.05 with SD_MN; and F_{ST} = 0.12–0.49, p > 0.05 with SD_EN) (**Supplementary Table S6**). The haplogroup sharing analysis also indicated that the YR populations during the LN shared higher proportions of haplotypes with Shimao-related populations than those in EN and MN (9.80% in TSZJZ, whereas 9.27% in QT_MN of Middle YR region; 9.05% in SD_LN, whereas 0.98% in SD_EN and 2.79% in SD_MN of Lower YR region) (**Figure 3C**; **Supplementary Table S7**). These results also suggest that the Shimao-related populations were closer to LN populations (after 4,500 BP) than the earlier populations (before 4,500 BP) found elsewhere in the YR.

Among the LN populations across the YR, the populations related to the Shimao culture were closest to those related to the Taosi culture (F_{ST} = 0.01, p > 0.05 in TSZJZ) (**Supplementary Table S6**). Similarly, the DAPC results showed Shimao-related populations clearly clustered with TSZJZ to the exclusion of other LN and BA populations (**Figure 3A**). In addition, the haplogroup sharing analysis also indicated that the Taosi-related individuals shared slightly higher proportions (9.80% in TSZJZ; 9.05% in SD_LN) of haplotypes with the Shimao-related populations (**Figure 3C**; **Supplementary Table S7**). The network results further showed that the haplotypes carried by TSZJZ, such as A17, C4a1a2, C4a2a1, D4b2b, D4e1a, F1a1c, and F2g, were also found in Shimao-related populations and that they shared the same branches (**Figures 3D,E**; **Supplementary Figure S2**). These results indicate that, among the populations in other regions of the YR basin, the Shimao-related populations had the closest genetic affinity with the Taosi culture-related populations in southern Shanxi Province.

In all, the ancient individuals related to the Shimao culture in the LN Longshan period from northern Shaanxi Province shared more maternal relationships with the contemporaneous (but not earlier) populations in the YR region outside northern Shaanxi Province. Among these LN Longshan populations, those related to Shimao culture shared the closest relationship with those related to Taosi culture in the Middle YR. These results demonstrate the strong and extensive population interactions during the LN Longshan period, not only within the northern Shaanxi Province but also between northern Shaanxi and southern Shanxi Provinces.

The Genetic Relationship Between Shimao-Related Populations and Present-Day Humans

To explore the genetic relationships between Shimao-related populations and present-day humans, we compared their genetic affinities including the ethnic minorities (such as Daur, Mongolia, Dai, Miao, etc.), Han populations ("NChina_Han" and

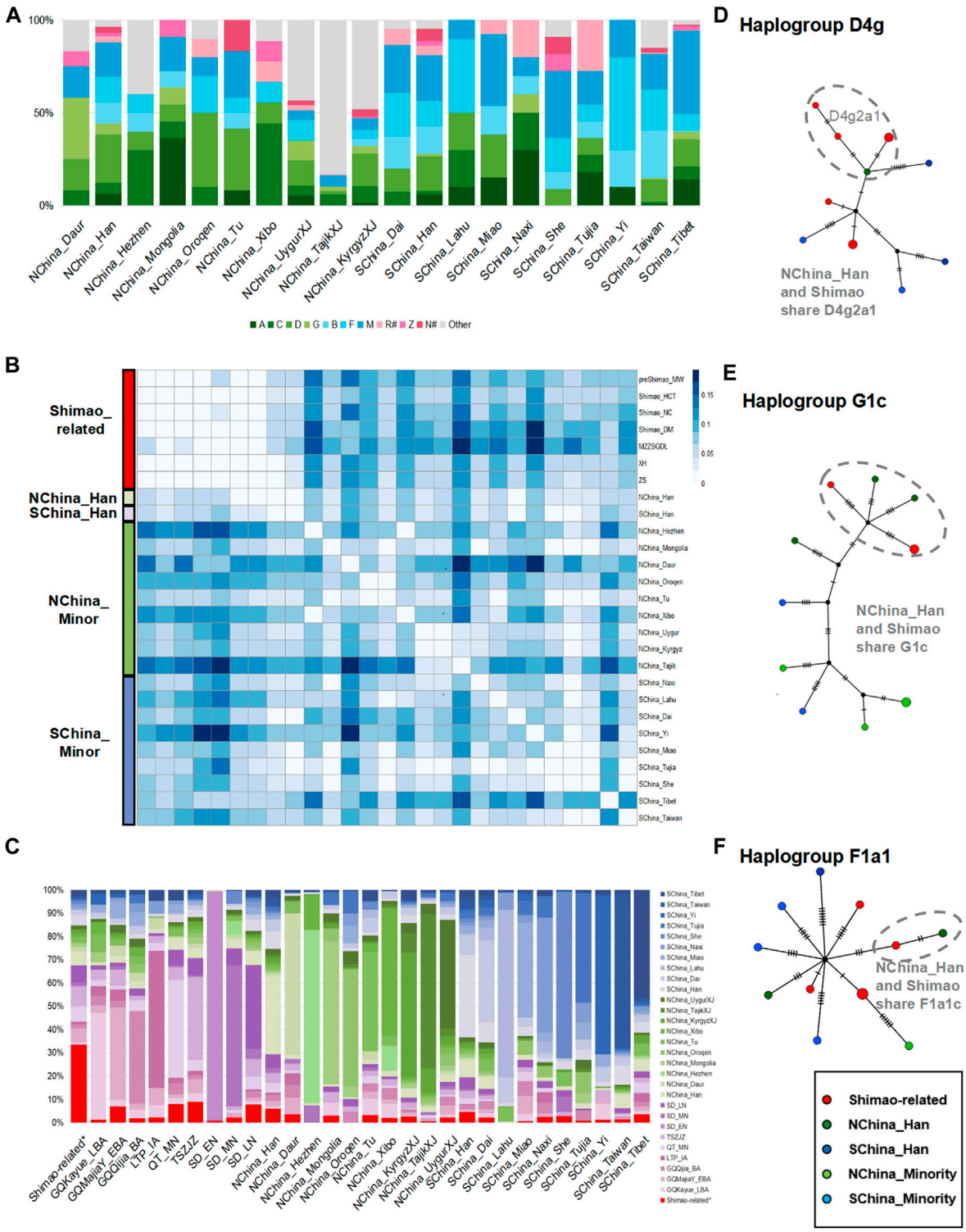


FIGURE 4 | The genetic analysis between Shimao-related populations and present-day Chinese populations. **(A)** Haplogroup frequencies. The haplogroups with green are those common in NEAs, and those with blue are common in SEAs. The haplogroups absent from Shimao-related populations are grouped in "Other." The haplogroups $R^{\#}$ and $N^{\#}$ represent the haplotypes found in East Eurasians (such as haplotype R+16189, sub-haplogroups R11 and N9, which were also observed in the Shimao-related populations). The other sub-haplogroups R and N found in West Eurasians were assigned to "Other." **(B)** Genetic distance (F_{ST}) heatmap of Shimao-related populations and present-day populations. NChina_Minor: the minorities in northern China; SChina_Minor: the minorities in southern China, SChina_Tibet: the populations in Tibet; SChina_Taiwan: the populations in Taiwan. **(C)** Haplogroup sharing analysis. Green and blue are used to represent populations in northern and southern China. Purple represents the ancient populations in YR. Median-joining networks of haplotypes D4g2a1 **(D)**, G1c **(E)**, and F1a1c **(F)** shared between Shimao-related populations and northern Han populations. Different colors of groups correspond to the heatmap in Figure 4B.

"SChina_Han"), and the populations in Tibet and Taiwan of China.

Among these present-day populations in China, the Han populations carried both the NEA dominating haplogroups A (5.95–6.46%), C (1.79–5.94%), D (19.05–25.84%), and G (1.19–5.94%) and SEA dominating haplogroups B (11.37–14.88%, B4′5), F (13.69–13.95%), and M (18.60–24.40%), and showed the highest proportion of haplogroup D (19.05%–25.84%), consistent with the Shimao-related populations (**Figure 4A**; **Supplementary Table S5**). The genetic distance analysis also showed that the Shimao-related populations (such as, Shimao_HCT, XH, ZS) were closer to Han populations ($F_{ST} < 0.03$, $p > 0.06$ in NChina_Han; $F_{ST} < 0.04$, $p > 0.06$ in SChina_Han) than other present-day minority populations, including those in Tibet and Taiwan (**Figure 4B**; **Supplementary Table S6**). The haplogroup sharing analysis showed that the Shimao-related populations shared higher proportions of haplotypes with the Han (NChina_Han, 6.04%; SChina_Han, 4.70%) than with the other present-day populations (0.00–3.65%) (**Figure 4C**; **Supplementary Table S8**). Additionally, the Shimao-related populations (such as, Shimao_HCT, XH, ZS) were genetically closer to northern Han populations (NChina_Han, $F_{ST} < 0.03$, $p > 0.06$) than the southern Han populations (SChina_Han, $F_{ST} > 0.03$, $p > 0.03$) (**Figure 4B**; **Supplementary Table S6**). Moreover, the haplogroup sharing analysis also demonstrated that the northern Han population shared higher proportions (NChina_Han, 6.04%) of haplotypes with Shimao-related populations than southern Han populations (SChina_Han, 4.70%) (**Figure 4C**; **Supplementary Table S8**). The network analysis also indicated that the Shimao-related populations and northern Han populations shared multiple haplotypes (D4g2a1, G1c, and F1a1) between them (**Figures 4D–F**). Thus, we conclude that the Shimao-related populations were closer to Han populations in northern China than to the minorities and Southern Han populations in China.

To explore which ancient populations had the closest genetic to the northern Han in China, we compared the affinities of our new Shimao-related populations and the ancient individuals in other regions of China to the northern Han. These include the EBA (GQMajiaY_EBA), BA (GQQijia_BA), late BA (GQKayue_LBA), and IA (LTP_IA) individuals from Gansu-Qinghai Province; the MN individuals (QT_MN) from Henan Province; the LN (TSZJZ) individuals from southern Shanxi Province; and the EN, MN, and LN individuals in Shandong Province (SD_EN, SD_MN, SD_LN). We found that the haplogroups A, C, D, G, Z, B (B4′5), F, M, and $R^{\#}$ were observed in the Shimao-related populations, LN Shandong individuals, and NChina_Han, whereas some of those haplogroups were absent from the other populations (**Figures 2A**, **3B**, **4A**). For example, TSZJZ lacked haplogroup $R^{\#}$, SD_MN lacked haplogroups C, F, and $R^{\#}$, and QT_MN lacked haplogroups C, $R^{\#}$, and Z (**Figure 3B**; **Supplementary Table S4**). Moreover, most of the Shimao-related populations (22.20–33.30%), SD_LN (26.00%), and NChina_Han (25.84%) carried the highest proportions of haplogroup D (**Supplementary Tables S4, S5**). The haplogroup sharing analysis showed that NChina_Han shared the highest proportions of haplotypes with Shimao-related populations (6.04%) compared to QT_MN (4.09%), TSZJZ (4.18%), and SD_LN (5.75%) (**Figure 4C**; **Supplementary Table S8**). The genetic distance analysis also showed that NChina_Han shared the closest genetic affinities with the Shimao-related populations ($F_{ST} = 0.02$, $p = 0.10$ in XH and ZS; $F_{ST} = 0.03$, $p = 0.06$ in Shimao_HCT) compared to QT_MN ($F_{ST} = 0.03$, $p = 0.00$); TSZJZ ($F_{ST} = 0.05$, $p = 0.00$); SD_LN ($F_{ST} = 0.06$, $p = 0.00$) (**Supplementary Table S6**). Thus, the Han populations in northern China shared more affinity with the Shimao-related populations than with the other published ancient individuals in China.

We found that the populations in and around Shimao City were closer to the northern Han Chinese populations than to the southern Han Chinese and minority populations. Compared to the other ancient individuals in China, we also found that these Han populations in northern China were closer to the Shimao-related populations.

DISCUSSION

Archeological research on Shimao City and the corresponding cultures reveal its importance, particularly in the Longshan period (~4,500–3,800 BP), as a crucially political and religious center in northern Shaanxi Province of the Middle YR (Sun et al., 2020a). However, the genetic origins of the Shimao people rested on the relationships between the populations related to the Shimao culture and the local populations in the preceding MN Yangshao period, along with the other ancient populations (especially those contemporaneous) in the YR basin outside northern Shaanxi Province. In the current study, we presented a large-scale dataset of ancient mitochondrial genomes from the ancient populations in northern Shaanxi and southern Shanxi Province of the Middle YR, especially those related to the Shimao and contemporaneous Taosi cultures. Through our new study, we have characterized the genetic structure and population dynamics of the Shimao-related populations and how populations changed from the MN Yangshao to the LN Longshan period through the present day.

First, previous genomic research on the populations in the Henan Province of the Middle YR found distinct genetic composition changes from the MN Yangshao period (~5,550–5,050 BP) to the LN Longshan period (~4,275–3,844 BP), with the latter having more ancestry from the south (Ning et al., 2020). Similarly, the ancient populations in the Shandong region of the Lower YR also exhibited changes in their mitochondrial genomes from the Yangshao (~5,500–4,600 BP) to Longshan (after 4,500 BP) period (Liu et al., 2021). In contrast, our results indicate that the populations in the LN Longshan period related to the Shimao culture (4,231–3,253 cal BP) in northern Shaanxi Province of Middle YR showed the closest genetic connections with their earlier populations in the MN Yangshao period (Wuzhuangguoliang and Miaoliang sites, 4,836–4,530 cal BP). This is consistent with archaeological studies showing that the relics excavated in the Longshan period maintained the features

of those in the Yangshao culture in this region, differentiating them from the Central Plain (Xing et al., 2002).

The origin of Shimao City was also uncertain (Sun et al., 2020a). Some researchers believed it was developed by the local populations and influenced by surrounding archaeological cultures (Gong, 2018), while some hypothesized that populations migrated from the southern region of the Middle YR and built Shimao City (Zhang, 2004; Duan and Dong, 2018). Interestingly, the close genetic connections between Shimao populations in the LN Longshan period and the preceding populations in the MN Yangshao period (Wuzhuangguoliang and Miaoliang sites, 4,836–4,530 cal BP) in our study support a mostly local origin for Shimao and being primarily developed by the Yangshao populations before 4,500 BP in northern Shaanxi Province. However, we also find evidence for some connections between the populations in (MN preShimao_MW and LN Shimao) and outside (Qingtai in Central Plain) of northern Shaanxi Province. Thus, despite the close connection between Shimao and its MN predecessors in northern Shaanxi Province, we cannot exclude the possibility of additional genetic contribution from the middle YR region. Future studies will clarify these relationships.

Second, the archaeological studies on potteries, tiles, and other relics unearthed from Shimao City revealed the populations in Huangchengtai was the core palaces with high hierarchy, while the populations in the Dongmen of Waicheng were mainly from sacrificial pits, which were the sacrificed victims with low social status (Sun and Shao, 2020). However, our findings revealed that these populations in different regions of Shimao City showed close genetic affinities, though they had different levels of inferred social status. Those populations in Xinhua, Muzhuzhuliang, Shengedaliang, and Zhaishan sites related to Shimao culture had varying levels of social hierarchy and also showed a close relationship with the ancestors of the Shimao people, indicating the extensive connections among the Shimao-related populations. Thus, although the LN Longshan populations in and around Shimao showed varying levels of social hierarchy, they also shared close genetic connections with each other, consistent with their proposed cultural connections (Sun et al., 2020a).

Moreover, the physical anthropology of the individuals in Dongmen suggested that their skulls were similar to those related to the Lower Xiajiadian culture (~4,150–3,590 BP) in Inner Mongolia (Chen et al., 2016). Published mitochondrial haplogroups from individuals at the Erdaojingzi site (WLR_LN in Ning et al., 2020), related to the Lower Xiajiadian culture, included B5b1a, A22, N9a1, which were all absent from the individuals at the Dongmen of Shimao City in our study. Stone carvings found in Shimao City (such as, in the center palace of Huangchengtai and Dongmen) were also observed in the Xinglongwa (~8,200–7,400 BP) and Zhaobaogou (~7,350–6,420 BP) culture related sites in Inner Mongolia (Lv, 1960; Wang, 1993; Sun and Shao, 2020). We didn't find a closely maternal affinity between the populations in Dongmen and Inner Mongolia; however, more ancient genomes from Inner Mongolia are needed to confirm the genetic affinity between them.

We then focused on the genetic affinity between Shimao-related populations and the ancient populations in other regions of the YR. Our results found that the Shimao-related populations in the LN Longshan period shared more genetic relationships with the contemporaneous LN populations in the YR region, rather than those from earlier periods. Among these YR populations in the LN Longshan period, the populations related to Shimao culture showed the closest affinity with the populations related to the Taosi culture in the Middle YR's southern Shanxi Province. Similarly, Xu (2014) conducted comparative studies of jade burial, painting and violence excavated from Shimao and Taosi cities, and showed that there were similarities in culture, and interaction in the economy, culture, and populations in these two vital cities (Xu, 2014), which were in accordance with our new findings.

Finally, we found that the LN Longshan Shimao-related populations in northern Shaanxi were closer to the present-day Han Chinese (especially the northern Han Chinese population) than to the minorities in China. A previous genetic study found that the MN Qingtai population also contributed to the Han in northern China. We found that northern Han Chinese populations shared more affinity with our Shimao-related populations than with the ancient Qingtai individuals. According to the recent census reports of the Chinese population and related historical studies, most of the humans in present-day northern Shaanxi were Han Chinese (Zhou, 2015; National Bureau of Statistics, 2021). Meanwhile, the genetic research on Han Chinese populations showed that the Han in northern China were identical and that the north-south genetic divergence was the major difference among Han populations (Li et al., 2019; He et al., 2022). Thus, our results revealed the close genetic affinities between the MN Yangshao (4,830–4,530 cal BP) and LN Longshan (4,231–3,253 cal BP) populations in northern Shaanxi Province and that they contributed to the present-day northern Han populations to a certain extent.

In summary, we conclude that the Shimao-related populations in the LN Longshan period are closely related to their predecessors in the MN Yangshao period in northern Shaanxi Province, revealing a mostly local origin for Shimao City. In addition, compared to other LN populations, the Shimao-related populations were genetically closer to the contemporaneous Taosi population in southern Shanxi Province, reflecting strong interactions in both regions of the Middle YR in the LN. As for their relationship with present-day people, the Shimao-related populations were genetically closer to northern Han population. Our study might provide a perspective for understanding the genetic affinities and population dynamics of Shimao-related populations in the Middle YR basin during the Neolithic period. Further studies with ancient genomic data will aim to test for more complex patterns of admixture and social organization in this region.

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: Genome Warehouse in

National Genomics Data Center (National Genomics Data Center Members and Partners (CNCB-NGDC), 2020; Chen et al., 2021), Beijing Institute of Genomics (China National Center for Bioinformation), Chinese Academy of Sciences, accession number PRJCA009290 (https://bigd.big.ac.cn/gwh). Other datasets for this study can be found in the **Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

QF designed and supervised the research project. QF, WW managed the project. ZS, LC, XG, ND, XP, XW, GZ, CC, LZ, and NH assembled archaeological materials and dating. QF, PC, RY, FL, XF, QD, and WP, performed or supervised wet laboratory work. QF, XF did the data processing and quality control. JX, WW and GZ, did the primary data analysis for this manuscript. JX, WW, JS, XD, JG, and QF. wrote the manuscript. All authors discussed, critically revised, and approved the final version of the manuscript.

FUNDING

This work was supported by the Chinese Academy of Sciences (CAS) and the Ministry of Finance of the People Republic of China (YSBR-019 and XDB26000000), National Natural Science Foundation of China (41925009, T2192950), National Social Science Fund of China (19ZDA232), the "Research on the Roots of Chinese Civilization" program of Zhengzhou University (XKZDJC202006), the Howard Hughes Medical Institute (grant no. 55008731), the Tencent Foundation (through the XPLORER PRIZE). XW was supported by the Key National Social Science Foundation of China (no. 16ZDA144).

ACKNOWLEDGMENTS

We would like to thank the archaeological teams from Shaanxi, Archaeology Institute of National Museum of China, and Archaeology Institute of Chinese Academy of Social Sciences.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.909267/full#supplementary-material

REFERENCES

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

Archeology Institute of America (2021). Top 10 Discoveries of the Decade. Available at: https://www.archaeology.org/issues/406-2101/features/9325-china-shimao-neolithic-city. (Accessed March 22, 2022).

Bandelt, H. J., Forster, P., and Röhl, A. (1999). Median-joining Networks for Inferring Intraspecific Phylogenies. *Mol. Biol. Evol.* 16 (1), 37–48. doi:10.1093/oxfordjournals.molbev.a026036

Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prüfer, K., et al. (2007). Patterns of Damage in Genomic DNA Sequences from a Neandertal. *Proc. Natl. Acad. Sci. U.S.A.* 104 (37), 14616–14621. doi:10.1073/pnas.0704665104

Chang, H. Y. (2005). Preliminary Study on the Process of Social Complication from Longshan Period to Early Erlitou Period. master's thesis. Chengdu (Sichuan): Sichuan University.

Chen, L., Sun, Z. Y., and Shao, J. (2017). Research on Neolithic Skeletons from Houyangwan Site of Shimao City in Shenmu County, Shaanxi Province. *Xi Bu Kao Gu.* 12 (03), 263–273.

Chen, L., Xiong, J. X., Shao, J., and Sun, Z. Y. (2016). The Research of the Forensic Analysis of Skulls from the Sacrifical Pits Located in Shimao Stie, Shaanxi Province. *Archaeol. Cult. Relics.* 37 (04), 134–142.

Chen, M. L., Ma, Y. K., Wu, S., Zheng, X. C., Kang, H. G., Sang, J., et al. (2021). Genome Warehouse: A Public Repository Housing Genome-Scale Data. *Genom. Proteom. Bioinform.* 19 (4), 584–589. doi:10.1016/j.gpb.2021.04.001

Dabney, J., and Meyer, M. (2012). Length and Gc-Biases during Sequencing Library Amplification: a Comparison of Various Polymerase-Buffer Systems with Ancient and Modern Dna Sequencing Libraries. *BioTechniques* 52 (2), 87–94. doi:10.2144/000113809

Dai, X. M. (1998). Evolution of Neolithic Cultural Pattern in the Yellow River Basin. *Acta Archaeologica Sinica.* 63 (04), 389–418.

Dong, G. H., Liu, F. W., Yang, Y. S., Wang, L., and Chen, F. H. (2016). Cultural Expansion and its Influencing Factors during Neolithic Period in the Yellow River Valley, Northern China. *Chin. J. Nat.* 38 (04), 248–252. doi:10.3969/j.issn.0253-9608.2016.04.003

Duan, T. J., and Dong, X. L. (2018). On the Nature of Remains in Shimao Site in Northern Shaanxi and Corresponding in the Perspective of *Li* Spectrum. *Res. China's Front. Archaeol.* 17 (02), 218–236.

Duan, X. Q. (2019). Study on Painted Pottery of Gansu and the Development of Prehistoric Painted Potter. *J. Northwest Univ. Natl.* 41 (06), 6–13. doi:10.14084/j.cnki.cn62-1185/c.2019.06.002

Edgar, R. C. (2004). MUSCLE: a Multiple Sequence Alignment Method with Reduced Time and Space Complexity. *BMC Bioinforma.* 5, 113. doi:10.1186/1471-2105-5-113

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

Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., et al. (2016). The Genetic History of Ice Age Europe. *Nature* 534 (7606), 200–205. doi:10.1038/nature17993

Fu, Q., Hajdinjak, M., Moldovan, O. T., Constantin, S., Mallick, S., Skoglund, P., et al. (2015). An Early Modern Human from Romania with a Recent Neanderthal Ancestor. *Nature* 524, 216–219. doi:10.1038/nature14558

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. U.S.A.* 110 (6), 2223–2227. doi:10.1073/pnas.1221359110

Gilbert, M. T. P., Bandelt, H.-J., Hofreiter, M., and Barnes, I. (2005). Assessing Ancient DNA Studies. *Trends Ecol. Evol.* 20 (10), 541–544. doi:10.1016/j.tree.2005.07.005

Gong, Q. M. (2018). The Important Harvest of Shaanxi Prehistoric Archaeology in the New Century. *Wenbo.* 35 (05), 31–50.

Guo, W. (2013). "The Communication between Northern China and Eurasian Steppe During the Longshan Period From the Stone Figures at Shimao Site.

Available at: https://kaogu.cn/html/cn/xueshuyanjiu/yanjiuxinlun/juluoyuchengshikaog/2013/1025/33681.html (Accessed March 22, 2022).

Guo, X. N., Wang, W. L., Kang, N. W., Qu, F. M., and Chen, L. (2016). Archaeology Survey of Excavation at the Shengedaliang Site, Shenmu County. *Archaeol. Cult. Relics* 37 (04), 34. doi:10.3969/j.issn.1000-7830.2016.04.004

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. doi:10.1038/nature14317

Hall, T. A. (1999). Bioedit: a User-Friendly Biological Sequence Alignment Editor and Analysis. *Nuclc Acids Symp.* 41 (41), 95–98.

He, G. L., Wang, M. G., Li, Y. X., Zou, X., Yeh, H. Y., Tang, R. K., et al. (2022). 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, N. (2004). Review on the Research History of Taosi Culture. *Anc. Civiliz.* 3 (00), 54–86.

Hou, G. L., Xu, C. J., Lv, C. Q., Chen, Q., and Lan, C. Z. (2019). Environmental Background of Yangshao Culture Expansion in Holocene. *Geogr. Res.* 38 (02), 437–444. doi:10.11821/dlyj020171214

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant Analysis of Principal Components: a New Method for the Analysis of Genetically Structured Populations. *BMC Genet.* 11, 94. doi:10.1186/1471-2156-11-94

Kang, L., Zheng, H.-X., Zhang, M., Yan, S., Li, L., Liu, L., et al. (2016). MtDNA Analysis Reveals Enriched Pathogenic Mutations in Tibetan Highlanders. *Sci. Rep.* 6, 31083. doi:10.1038/srep31083

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double Indexing Overcomes Inaccuracies in Multiplex Sequencing on the Illumina Platform. *Nucleic Acids Res.* 40 (1), e3. doi:10.1093/nar/gkr771

Ko, A. M.-S., Chen, C.-Y., Fu, Q., Delfin, F., Li, M., Chiu, H.-L., et al. (2014). Early Austronesians: into and Out of Taiwan. *Am. J. Hum. Genet.* 94 (3), 426–436. doi:10.1016/j.ajhg.2014.02.003

Leigh, J. W., and Bryant, D. (2015). Popart : Full-feature Software for Haplotype Network Construction. *Methods Ecol. Evol.* 6, 1110–1116. doi:10.1111/2041-210X.12410

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25 (14), 1754–1760. doi:10.1093/bioinformatics/btp324

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

Liu, J., Zeng, W., Sun, B., Mao, X., Zhao, Y., Wang, F., et al. (2021). Maternal Genetic Structure in Ancient Shandong between 9500 and 1800 Years Ago. *Sci. Bull.* 66 (11), 1129–1135. doi:10.1016/J.SCIB.2021.01.029

Lv, Z. E. (1960). Archaeological Survey of Linxi Site, Inner Mongolia. *Acta Archaeol. Sin.* 25 (1), 9–23.

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., et al. (2012). A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* 338 (6104), 222–226. doi:10.1126/science.1224344

Miao, B., Liu, Y., Gu, W., Wei, Q., Wu, Q., Wang, W., et al. (2021). Maternal Genetic Structure of a Neolithic Population of the Yangshao Culture. *J. Genet. Genomics* 48 (08), 746–750. doi:10.1016/j.jgg.2021.04.005

National Bureau of Statistics (2021). *Major Figures on 2020 Population Census of China*. Beijing: China Statistical Press.

National Genomics Data Center Members and Partners (CNCB-NGDC) (2020). Database Resources of the National Genomics Data Center in 2020. *Nucleic Acids Res.* 48 (D1), D24–D33. doi:10.1093/nar/gkz913

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

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155 (2), 945–959. doi:10.1093/genetics/155.2.945

Rawson, J. (2017). Shimao and Erlitou: New Perspectives on the Origins of the Bronze Industry in Central China. *Antiquity* 91 (355), E5. doi:10.15184/aqy.2016.234

Renaud, G., Stenzel, U., and Kelso, J. (2014). leeHom: Adaptor Trimming and Merging for Illumina Sequencing Reads. *Nucleic Acids Res.* 42 (18), e141. doi:10.1093/nar/gku699

Rohland, N., Harney, E., Mallick, S., Nordenfelt, S., and Reich, D. (2015). Partial Uracil-DNA-Glycosylase Treatment for Screening of Ancient DNA. *Phil. Trans. R. Soc. B* 370 (1660), 20130624. doi:10.1098/rstb.2013.0624

Shao, J. (2020). A Comparative Study of Shimao Site and Taosi Site. *Archaeology.* 66 (05), 65.

Shao, J., Pei, X. S., Di, N., Yuan, Y., Zhao, K., and He, C. L. (2021). Preliminary Report on the Excavation of the Residential Area at the Miaoyan Locality of the Zhaishan Site in Fugu County, Shaanxi Province. *Wenbo.* 38 (5), 15–28+2+113. doi:10.3969/j.issn.1000-7954.2021.05.003

Sun, Z. Y. (2005). Review of Xinhua Culture. *Archaeol. Cult. Relics* 26 (03), 40. doi:10.3969/j.issn.1000-7830.2005.03.006

Sun, Z. Y., Shao, J., Shao, A. D., Zhao, X. H., Yang, G. Q., Kang, N. W., et al. (2016). Archaeology Survey of Excavations to Hanjiagedan Locus of the Shimao Site in Shenmu County, Shaanxi Province. *Archaeol. Cult. Relics.* 37 (04), 14. doi:10.3969/j.issn.1000-7830.2016.04.002

Sun, Z. Y., and Shao, J. (2015). Archaeology Survey of Trial Excavations to Houyangwan and Hujiawa Localities of the Shimao Site in Shenmu County, Shaanxi Province. *Archeology.* 61 (05), 60.

Sun, Z. Y., Shao, J., and Di, N. (2020a). Archaeological Discovery and Research Synthesis of Shimao Site. *Cult. Relics Central China* 44 (01), 39–62.

Sun, Z. Y., Shao, J., and Di, N. (2020b). Nomenclature, Range and Age of Shimao Culture. *Archaeology.* 66 (8), 101–108.

Sun, Z. Y., and Shao, J. (2020). Research on the Unearthed Stone Sculptures from the the Large Foundation of Huangchengtai at Shimao Site. *Archaeol. Cult. Relics.* 41 (04), 40. doi:10.3969/j.issn.1000-7830.2020.04.005

Sun, Z. Y., Shao, J., Shao, A. D., Kang, N. W., Qu, F. M., and Liu, X. M. (2013). Shimao Site in Shenmu County, Shaanxi. *Archaeology.* 59 (07), 15.

Sun, Z. Y. (2016). Study on the Formation Process of Social Complexity in Northern China in the Third Millennium BC. *Archaeol. Cult. Relics.* 37 (04), 70–79. doi:10.3969/j.issn.1000-7830.2016.04.008

Tian, W., and Dai, X. M. (2018). Archaeology Survey on the Excavation of Zhoujiazhuang Site, Jiangxian County, Shanxi Province in 2013. *Archaeology.* 64 (01), 28.

van Oven, M. (2015). PhyloTree Build 17: Growing the Human Mitochondrial DNA Tree. *Forensic Sci. Int. Genet. Suppl. Ser.* 5, e392–e394. doi:10.1016/j.fsigss.2015.09.155

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer.

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, F. (1989). On the Position of the Yellow River Basin in the Origin of Chinese Civilization - Also on the Cause of the Establishment of Xia Dynasty. *J. Renmin Univ. China* 3 (03), 95–102.

Wang, G. (1993). Stone Carvings with Portrait in Xinglongwa Culture. *China Cult. Relics.* 9 (47), 3.

Wang, J. H. (2005). A Study on the Prehistoric Population in the Middle and Lower Reaches of the Yellow River. dissertation/ph.D's thesis. Jinan (Shandong): Shandong University.

Wang, W. L., Guo, X. N., Kang, N. W., Liu, X. M., and HuChen, K. L. (2015). Archaeology Survey on the Muzhuzhuliang Site in Shenmu County, Shaanxi. *Archaeol. Cult. Relics.* 36 (05), 3. doi:10.3969/j.issn.1000-7830.2015.05.001

Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.-J., et al. (2016). HaploGrep 2: Mitochondrial Haplogroup Classification in the Era of High-Throughput Sequencing. *Nucleic Acids Res.* 44 (W1), W58–W63. doi:10.1093/nar/gkw233

Xing, F. L., Li, M., and Sun, Z. Y. (2002). Brief Excavation Report of Xinhua Site, Shenmu, Shaanxi in 1999. *Archaeol. Cult. Relics.* 23 (01), 3. doi:10.3969/j.issn.1000-7830.2002.01.001

Xu, F. (2014). Preliminary Comparison of Shimao and Taosi Archaeology Discovery. *Wenbo.* 31 (01), 18. doi:10.3969/j.issn.1000-7954.2014.01.003

Xu, Y. J. (2004). The YangShao Culture Survivals' Pedigrees in Later Period in the Loess Plateau. dissertation/ph.D's thesis. Changchun (Jilin): Jilin University.

Yan, Z. B., and He, N. (2005). Archaeology Survey on the Taosi Site in Xiangfen County, Shanxi. *Acta Archaeol. Sin.* 70 (03), 307.

Yang, Y. Y. (2016). A Study of the Cultural Development and Landscape Distribution of the Majiayao Culture in Gansu and Qinghai Provinces. dissertation/master's thesis. Lanzhou (Gansu): Lanzhou University.

Zhang, X. (2015). The Research of Dawenkou Culture. dissertation/ph.D's thesis. Changchun (Jilin): Jilin University.

Zhang, Z. P. (2004). The Pottery *Li* with Side-Attached Double Lugs of Xinghua Culture. *Palace Mus. J.* 26 (04), 6. doi:10.16319/j.cnki.0452-7402.2004.04.001

Zhang, Z. P. (1996). Yangshao Age - the Prosperity of Prehistoric Society and the Transition to Civilization. *Palace Mus. J.* 18 (01), 1–44.

Zhao, J., Liu, F.-E., Lin, S., Liu, Z.-Z., Sun, Z.-Y., Wu, X.-M., et al. (2017). Investigation on Maternal Lineage of a Neolithic Group from Northern Shaanxi Based on Ancient DNA. *Mitochondrial DNA Part A* 28 (5), 732–739. doi:10.1080/24701394.2016.1177039

Zhou, S. C. (2017). The Origin of Chinese Civilization Yellow River Basin Center Talked about the Theoretical Evolution of Pluralistic Unity Theory - a New Investigation from the Perspective of Academic History. *West. Historiogr.* 1 (00), 11–33.

Zhou, W. Z. (2015). The Ethnic and Fusion of Northern Shaanxi in Historical Period. *Northwest Ethnol. Ser.* 14 (02), 14.

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 | Frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 27 June 2022
doi: 10.3389/fevo.2022.901295

Uniparental Genetic Analyses Reveal Multi-Ethnic Background of Dunhuang Foyemiaowan Population (220–907 CE) With Typical Han Chinese Archaeological Culture

Jianxue Xiong[1,2,3†], Yichen Tao[1†], Minxi Ben[1†], Yishi Yang[4], Panxin Du[1,5], Edward Allen[2], Hui Wang[2,3,6], Yiran Xu[3,6], Yao Yu[2], Hailiang Meng[1], Haoquan Bao[1], Boyan Zhou[7], Guoke Chen[4], Hui Li[1]* and Shaoqing Wen[2,3,6]**

[1] Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences & MOE Laboratory for National Development and Intelligent Governance, Fudan University, Shanghai, China, [2] Department of Cultural Heritage and Museology, Fudan University, Shanghai, China, [3] Institute of Archaeological Science, Fudan University, Shanghai, China, [4] Institute of Cultural Relics and Archaeology in Gansu Province, Lanzhou, China, [5] State Key Laboratory of Genetic Engineering, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Human Phenome Institute, Fudan University, Shanghai, China, [6] Center for the Belt and Road Archaeology and Ancient Civilizations, Fudan University, Shanghai, China, [7] Division of Biostatistics, Department of Population Health, School of Medicine, New York University, New York, NY, United States

OPEN ACCESS

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

Reviewed by:
Hoh Boon-Peng, UCSI University, Malaysia
Wibhu Kutanan, Khon Kaen University, Thailand

***Correspondence:**
Guoke Chen
chenguoke1980@sina.com
Hui Li
lhca@fudan.edu.cn
Shaoqing Wen
wenshaoqing1982@gmail.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 Ecology and Evolution

Received: *21 March 2022*
Accepted: *14 April 2022*
Published: *27 June 2022*

Citation:
Xiong J, Tao Y, Ben M, Yang Y, Du P, Allen E, Wang H, Xu Y, Yu Y, Meng H, Bao H, Zhou B, Chen G, Li H and Wen S (2022) Uniparental Genetic Analyses Reveal Multi-Ethnic Background of Dunhuang Foyemiaowan Population (220–907 CE) With Typical Han Chinese Archaeological Culture. Front. Ecol. Evol. 10:901295. doi: 10.3389/fevo.2022.901295

The relationship between archeological culture and ethnicity is invariably complex. This is especially the case for periods of national division and rapid inter-ethnic exchange, such as China's Sixteen Kingdoms (304–439 CE) and Northern and Southern Dynasties (420–589 CE). Going by tomb shape and grave goods, the Foyemiaowan cemetery at Dunhuang exhibits a typical third–tenth century Han style. Despite this, the ethnic makeup of the Foyemiaowan population has remained unclear. We therefore analyzed 485 Y-chromosomal SNPs and entire mitochondrial genomes of 34 Foyemiaowan samples. Our study yielded the following discoveries: (1) principal component analysis revealed that the Foyemiaowan population was closely clustered with Tibeto-Burman populations on the paternal side and close to Mongolic-speaking populations on the maternal side; (2) lineage comparisons at the individual level showed that the Foyemiaowan population consisted of primarily Tibeto-Burman and Han Chinese related lineages (Oα-M117, 25%;Oβ-F46, 18.75%), partially Altaic speaking North Eurasian lineages (N-F1206, 18.75%) and a slight admixture of southern East Asian lineages (O1b1a2-Page59, 6.25%; O1b1a1-PK4, 3.13%). Similarly, the maternal gene pool of Foyemiaowan contained northern East Asian (A, 4.17%; CZ, 16.67%; D, 20.83%; G, 4.17%; M9, 4.17%), southern East Asian (B, 12.51%; F, 20.83%) and western Eurasian (H, 4.17%; J, 4.17%) related lineages; (3) we discovered a relatively high genetic diversity among the Foyemiaowan population (0.891) in our ancient reference populations, indicating a complex history of population admixture. Archeological findings, stable isotope analysis

and historical documents further corroborated our results. Although in this period China's central government had relinquished control of the Hexi Corridor and regional non-Han regimes became the dominant regional power, Foyemiaowan's inhabitants remained strongly influenced by Han culture.

Keywords: Dunhuang, dynastic transitions, archeological culture, genetic diversity, multi-ethnicity

INTRODUCTION

The relationship between archeological culture and ethnic group has been examined extensively and deeply and remains a central issue within archeological research. Some studies suggest a connection between ethnicity and archeological finds, such as diagnostic pottery, tools, metallic objects like brooches, or in some cases residential areas. Yet any approach that simply equates the two has been thoroughly criticized in recent years (Upton, 1996; Jones, 1997; Renfrew and Bahn, 2007).

Cultural and ethnic homogeneity and heterogeneity are especially relevant for archeological contexts involving large-scale migration. For example, the Eurasian Bronze Age (circa. 3000–1000 BC) is widely acknowledged as a period of major cultural change—change intertwined with large-scale population migration (Allentoft et al., 2015). Multiple lines of evidence show that male-driven migration introduced Steppe ancestry to almost all Corded Ware populations in Central Europe (Caramelli et al., 2021), precipitating major shifts in burial practice (Anthony, 2007; Furholt, 2019). In a separate context, archeological analysis has demonstrated that the Quarto Cappello del Prete necropolis, in the heart of the globalized Roman Empire, contained individuals with highly heterogeneous geographical origins. Genomic data also supported these results and proved the applicability of genomic study for drawing out the ethnic contours of such multi-ancestral societies (De Angelis et al., 2021). However, ethnic groups rarely reflect the sum total of similarities and differences in "objective" cultural traits (Jones, 1997). Archeological culture and ethnic background may be distinct, as in the case of Taojiazhai (∼1700–1900 BP). This site was Han Chinese in style, though DNA evidence indicated a strong degree of Tibeto-Burman ancestry, identifying the Taojiazhai population as descendants of Di-Qiang groups (Zhao et al., 2011).

The Hexi Corridor of northwestern China was an important channel for cultural exchange and human migration between the ancient Central Plains and Western Regions (centered around modern Xinjiang). The region may also be considered an important component of the "Northwest National Corridor," one of three national corridors put forward by Fei Xiaotong in the early 1980s (Fei, 1982). During the Warring States period (476–221 BCE), the Hexi Corridor was inhabited by various nomadic peoples, known historically as the Yuezhi, Wusun and Xiongnu (Si, 2002a). The migration of population as a means of strengthening control over newly-acquired territory was a critical strategy of China's historical conquests during this and the subsequent Han period. For example, in the "Biography of Chulizi" in Sima Qian's *Historical Records* (史记. 楮里子传), we learn of Quwo (Lingbao city, Henan, China) occupied by the Qin army in 330 BC, its locals evicted from their hometown and replaced by Qin migrants (Si, 2002b). The Han dynasty government continued this gradual population-based domination of the Hexi Corridor. A series of policies was adopted to isolate Qiang people located in the Hehuang valley and Xiongnu in the Mongolian Steppe, as well as establish contact with the Western Regions. Counties were set up at Wuwei County (武威郡), Zhangye County (张掖郡), Jiuquan County (酒泉郡) and Dunhuang County (敦煌郡) running from east to west along the Hexi Corridor. According to historical documents and unearthed bamboo slips, a considerable population was relocated the Middle and Lower Yellow River watershed (i.e., Hongnong county, He'nei county, Runan county, Julu county, Yingchuan county, among others) (Liu, 2012). A recent study of the middle Hexi Corridor population at Heishuiguo has provided corroborating evidence of this transformation through Y chromosome and mtDNA analysis of individuals from that site, suggesting most Heishuiguo males had migrated from Middle and Lower Yellow River regions, while females were largely natives (Xiong et al., 2022). Meanwhile, localized archeological cultural traits also suggest that the Heishuiguo population inherited a Central Plains tradition. Mass migration of individuals not only impacted the genetic structure of the Hexi population, but also resulted in changes in local subsistence strategy (Chen et al., 2019).

Following the collapse of the Han dynasty in the third century CE, ancient China entered the Wei, Jin and Northern and Southern Dynasties period. Governments occupying the traditional heart of state power in the Central Plains largely abandoned the Hexi corridor; burgeoning regional political forces became the dominant driver of regional developments. The region experienced several dynastic transitions, from the Former Liang (314–376 CE), to Western Qin (385–400 CE and 409–431 CE), Latter Liang (386–403 CE), Northern Liang (397–439 CE), Southern Liang (397–414 CE) and Western Liang (400–421 CE) dynasties. Minority groups also established regimes under the Western Qin, Southern Liang, founded by Xianbei peoples, and the Latter Liang and Northern Liang established by Di peoples and Lushuihu (卢水胡), respectively. Archeological and historical materials, however, both suggest Han Chinese from the Central Plains had become the dominant Hexi Corridor population by the Jin Dynasty—acquiring this position through the continuing exertions of central government, causing Han and Hu (胡, general term for northern non-Han groups) to coexist in local communities (Lv, 2017). Regime changes and warfare were commonplace following the collapse of the Jin Dynasty, particularly as Chinese dynasties and nomadic polities frequently clashed. As such, the region continued to experience recurrent population inflow, through both political or military

immigration, with incomers kidnapped by regional political regimes, or refugees fleeing wars and natural disasters (Lv, 2017).

Dunhuang County, at the west end of the Hexi Corridor, lies adjacent to the Western Regions. In the wake of the Han Dynasty, Dunhuang retained its position as an important military town and frontier for central government administration over the Western Regions, successively controlled by the Latter Liang and Western Liang polities. With the population exposed to migration from the Western Regions, Mongolian Steppe, Qinghai-Tibetan Plateau and Central Plain, complex population interaction occurred here and in the entire west of the Hexi Corridor over these centuries. At Foyemiaowan, our first hypothesis anticipated a similarly diverse population.

However, study of the archeological culture at Foyemiaowan supports the Han Chinese origin hypothesis. The Foyemiaowan cemetery, in Wudong county in Dunhuang (**Figure 1A**), lies north of the Mogao Grottoes, a World Heritage Site renowned for its Buddhist mural paintings. The cemetery dates back as early as the Cao-Wei period (220–280 CE), and remained in use for approximately 600 years. This affords Foyemiaowan almost perfect coverage of the Cao-Wei to Sui/Tang periods. Tomb shape and grave goods suggest a cultural core at Foyemiaowan site that continued to adhere to a Han Chinese tradition (Chen et al., 2022). For example, numerous examples of daily-use potteries (**Figure 1D**) and bronze mirrors (see **Figures 1B,C**) in typical Han Chinese style were found in Foyemiaowan burials (Chen et al., 2022).

Were the inhabitants of Foyemiaowan Han or diverse origin? This would have been unanswerable previously, but recent years have seen the widespread study of uniparentally inherited markers in an effort to understand the population history, origin, and migration of human populations (Pamjav et al., 2017). This work offers a suite of methods that can be applied to similar questions at Foyemiaowan. This study aims to update our knowledge of the genetic history of Hexi Corridor populations based on the 485 SNPs Y-chromosome and whole mitochondrial genomes from Foyemiaowan. Such data will allow us to test the Han and diverse origin hypothesis, explore the relationship between archeological culture and ethnic group, as well consider what factors could have affected the genetic profile at Foyemiaowan.

MATERIALS AND METHODS

Materials

Foyemiaowan site contained thousands of tombs, which were excavated by the Institute of Cultural Relics and Archaeology of Gansu Province in 2015. The cemetery was divided into eight natural areas. The excavators of Foyemiaowan have argued for a four-phase division of the cemetery: Phase 1: Cao-Wei period (220–265 CE); Phase 2: Western Jin Dynasty (265–316 CE); Phase 3: Former Liang to Northern Liang Dynasty (314–439 CE); Phase 4: Sui Dynasty (581–618 CE) and Tang Dynasty (618–907 CE) (Chen et al., 2022). In this study, most samples can be assigned to the Sixteen Kingdoms period (Phase 3). Few samples were preserved from the remaining phases. Sixteen Kingdoms individuals thus form the bedrock of this study and belong to a unique period of chronic dynastic transition. Our survey population was made of up 34 ancient samples. Sex was determined by pelvic (Klales et al., 2012) and skull morphology (Buikstra, 1994), including 32 males and 2 females.

Methods

DNA Extraction and Library Preparation

For our 34 samples, we extracted DNA from the temporal bones, teeth and limb bones (**Table 1** and **Supplementary Table 1A**), using a dedicated aDNA facility in Fudan University, Shanghai. Molecular methods used for aDNA extraction and construction of Illumina libraries have been described previously (Zhu et al., 2021; Xiong et al., 2022). One extraction (no sample powder used) and one PCR blank (extract supplemented by water) were set up as negative controls for each batch of samples. Libraries were sequenced on Illumina HiSeq X10 instrument at the Annoroad Company (Beijing, China) using the 150 bp paired-end sequencing design.

Mitochondrial Capture and Sequencing

MtDNA enrichment was carried out on 13 samples that yielded a lower endogenous rate in the initial shotgun screening. Target enrichment of the mitogenome was performed using a MyGenostics Human Mitochondria Capture Kit (MyGenostics Company, Beijing, China) (Sun et al., 2021). Post-enrichment product was then quantified via qPCR and sequencing was performed using a NovaSeq 6000 platform at Mingma Technologies Company (Shanghai, China). 150 bp paired-end reads were generated according to the manufacturer's instructions.

Multiplex PCR Targeted Amplification and Sequencing for Y Chromosome

We opted for multiplex PCR targeting enrichment with short amplicons based on the NGS (Next Generation Sequencing) platform. In this study, a sensitive short amplifier primer system including 485 Y-SNPs (Wen et al., 2019; Xiong et al., 2022) was conducted to test Y-lineages for each male sample from the Foyemiaowan site. The system covered the common East Asian lineages.

Sequence Data Processing and Ancient DNA Authentication

For shotgun and mtDNA captured data, we clipped sequencing adapters and merged these using sequences by ClipAndMerge v1.7.8 (Peltzer et al., 2016), then mapped merged reads to the human reference genome (hs37d5; GRCh37 with decoy sequences) using BWA v0.7.17 (Li and Durbin, 2010). We used Dedup v0.12.3 (Peltzer et al., 2016) to remove PCR duplicates. Utilizing trimBam implemented in BamUtil v1.0.14,[1] we clipped four bases from both ends of each read to avoid an excess of remaining C- > T and G- > A transitions at the ends of DNA sequences.

The authenticity of the ancient genome sequence was mainly determined by the combination of two observations of the

[1]https://github.com/statgen/bamUtil

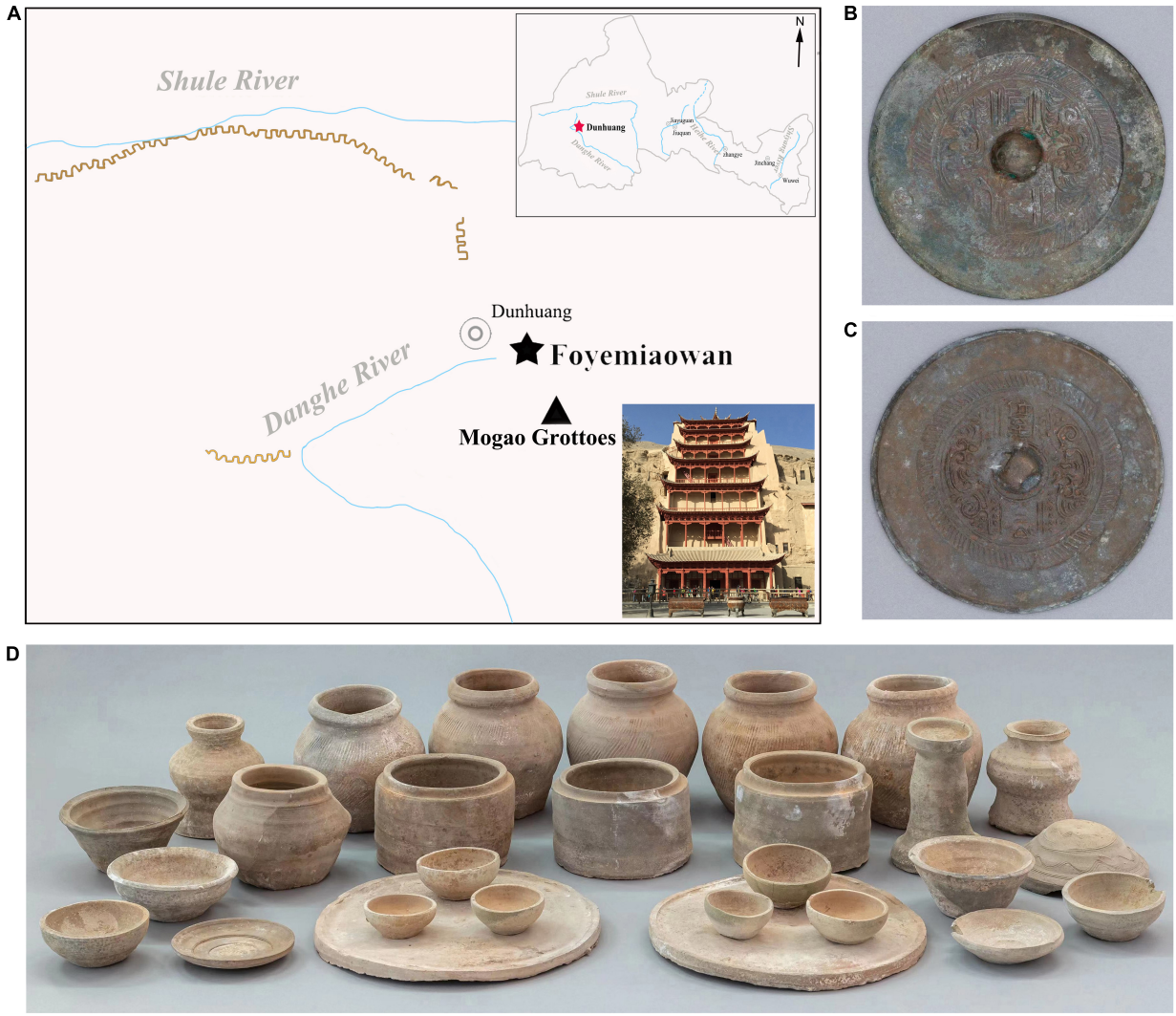


FIGURE 1 | (A) The geographical location of the Foyemiaowan site. **(B)** IIIM11:1 Bronze Mirror. **(C)** IIIM20:1 Bronze mirror. **(D)** IVM24: Daily-use potteries of Foyemiaowan cemetery.

same specimen. Firstly, we checked DNA damage pattern (**Supplementary Figure 1**) and estimated the 5′ C > T and 3′ G > A misincorporation rate using mapDamage v 2.0.61 (Jónsson et al., 2013). We then used the Schmutzi program to test mitochondrial contamination rates for all individuals (Renaud et al., 2015).

Uniparental Haplogroup Assignment

Y chromosome haplogroups were examined by aligning a set of positions in the ISOGG (International Society of Genetic Genealogy)[2] and Y-full[3] databases. Haplogroup determination was performed with the script Yleaf.py in Yleaf software (Ralf et al., 2018), which provides outputs for allele counts of ancestral and derived SNPs along a path of branches of the Y-chromosome tree (**Supplementary Table 1B**). Finally, we re-checked the SNPs by visual inspection with IGV software (Helga et al., 2013).

In order to call mtDNA consensus sequences, we employed a log2fasta program implemented in Schmutzi (Renaud et al., 2015). Mutations that appeared when checked against rCRS were also re-checked in BAM (Binary Alignment Map) files through visual inspection using the IGV software (Helga et al., 2013). Lastly, we used HaploGrep 2 (Weissensteiner et al., 2016) to assign the haplogroups (**Supplementary Table 1C**).

Principal Component Analysis and Haplogroup Diversity Calculation

Haplogroup diversity (H) was calculated using Nei's formula (Nei and Tajima, 1981). All the above analyses were performed in R 3.6.3. Reference populations are listed in **Supplementary Table 2A**. Principal component analysis (PCA) was performed using a prcomp () function of R. Visualization of PCA results

[2]http://isogg.org/

[3]https://www.yfull.com/tree/

TABLE 1 | Ancient individuals sampled in this study.

Sample ID	Archeological ID	Periods	Skeletal element	5 C-T%	Sex	Contamination rate	MT_depth	MT_Haplogroup	Identified SNPs	Y-SNP
G32712	IIM19 south	Cao-Wei	Tooth	12	Male	0.035	27.3768	F2a*	479	O*-CTS10738
G30411	IVM24 middle	Cao-Wei	Temporal bone	24	Male	0.027	37.8246	F1a1c*	480	N*-F710
F11313	IM9 later ②	Cao-Wei	Limb bone	/	Male	/	/	/	140	O-F48
G10912	VIM16 north1	Cao-Wei	Temporal bone	14	Female	0.088	14.8457	B5a2a1 + 16129*	/	/
G40812	VM6	Western Jin	Tooth	20	Male	0.181	5.3922	D5c1	94	O-F141
G40819	VIM3 south	Western Jin	Tooth	/	Male	/	/	/	479	O*-F48
F91109	IIIM50 middle	Western Jin	Metacarpal	10	Male	0.031	35.0509	F2a*	478	O-M188
G40808	VM16 middle	Western Jin	Tooth	14	Male	0.013	80.6472	Z3 + 709*	68	O-F8
G32702	IM7	Sixteen Kingdoms	Tooth	11	Male	0.018	54.1195	B4b1a2*	479	O*-F444
G32714	IM4	Sixteen Kingdoms	Tooth	/	Male	/	/	/	470	O*-F444
FA0218	IIM4	Sixteen Kingdoms	Temporal bone	/	Male	/	/	/	476	Q*-F4759
F91105	IIIM21	Sixteen Kingdoms	Tooth	20	Male	0.003	290.682	D4o*	478	N*-F2584
G30414	IIIM37 north	Sixteen Kingdoms	Temporal bone	19	Male	0.005	183.858	D4h1 + 7181T + 7673G*	479	Q*-M120
G40815	IIIM56	Sixteen Kingdoms	Tooth	15	Male	0.003	281.637	J1b1a1 + 146*	479	N*-F2569
F11306	IM2 south	Sixteen Kingdoms	Tooth	7	Male	0.103	26.1692	M11b1*	478	N*-F710
G32715	IM13 south ②	Sixteen Kingdoms	Limb bone	10	Male	0.034	68.0916	M33 + 16362*	446	C*-F8465
G10901	IIM3 south	Sixteen Kingdoms	Tooth	/	Male	/	/	/	129	O-F8
G32711	IIIM11 north	Sixteen Kingdoms	Tooth	20	Male	0.002	346.7353	F1a1c + 15314*	479	O-F8
G30412	IIM12 middle	Sixteen Kingdoms	Temporal bone	23	Male	0.007	143.8907	M9a1a*	480	O*-F1759
G32708	IIM23 south	Sixteen Kingdoms	Tooth	16	Male	0.021	49.4835	D4*	473	N*-F2130/F3361
G10115	IIM14 ①	Sixteen Kingdoms	Temporal bone	19	Male	0.017	24.4397	F1a1	479	N*-F2584
G40806	VM3 north	Sixteen Kingdoms	Tooth	21	Male	0.008	115.8781	B4a1*	424	O-F46
F91114	IIIM6	Sixteen Kingdoms	Tooth	/	Male	/	/	/	135	O-F201
G40816	IIIM46	Sixteen Kingdoms	Tooth	19	Male	0.001	624.201	D4b2b*	479	C*-F3967
F10722	IIIM27R2	Sixteen Kingdoms	Temporal bone	/	Male	/	/	/	348	O-F46
G10907	IIIM26 south	Sixteen Kingdoms	Temporal bone	15	Female	0.022	50.1714	A6*	/	/
G10903	VIM7 west	Sui	Temporal bone	15	Male	0.021	49.0852	Z3*	478	O*-M188
G30406	IIIM30	Tang	Temporal bone	18	Male	0.005	174.8135	G1a2'3*	479	O*-F4370
F91104	VM11 north	Tang	Temporal bone	8	Male	0.134	7.6485	C7a1c	470	O*-CTS1451
G40803	VM13	Tang	Tooth	9	Male	0.008	124.0274	H7b*	479	O*-CTS10738
F91101	IM15	Undetermined	Tooth	6	Male	0.037	68.0667	Z4a*	479	O*-F46
G32701	IIM9	Undetermined	Tooth	/	Male	/	/	/	472	N*-F2584
G40809	IVM26	Undetermined	Tooth	/	Male	/	/	/	479	O*-F46
F91102	IVM9 north	Undetermined	Tooth	/	Male	/	/	/	448	O*-F310

were conducted using the "ggplot2" package, as well as a pie chart. Reference populations of principal component analysis are listed in **Supplementary Table 3**. The Map of China was drawn using "mapchina" and "sf" packages.

Phylogenetic Tree Construction

For comparison with the newly generated 17 ancient mitogenomes from Foyemiaowan Site, previously published data was assembled from the Mitomap database[4] (**Supplementary Table 4**). In total, 417 mitogenomes were employed to construct the maximum-parsimony (MP) phylogenetic trees for each haplogroup (**Supplementary Table 5**), using mtPhyl v5.003 software.[5]

Molecular Dating

Coalescence time estimates were also computed using Bayesian MCMC approach implemented in the BEAST v2.4.7 software package (Drummond et al., 2012). We constructed Bayesian trees using both modern and ancient samples, with the latter dates used as tip dates for molecular clock calibration (Fu et al., 2013; Rieux et al., 2014).

For mitochondrial time estimation, we first aligned all sequences to the rCRS using the MAFFT (Katoh et al., 2017) version 7 program using the iterative refinement method: E-INS-I. Then, we partitioned the alignments into four parts with a scheme adapted in accordance with a published molecular clock calibration for human mtDNA (Rieux et al., 2014): (1) First and second nucleotides in codons of protein coding genes (PC1 + PC2), (2) third nucleotides in codons of protein coding genes (PC3), (3) rRNAs + tRNAs (RNA region) and (4) HVRI + HVRII (control region). Indels were removed from the alignments to avoid potential biases due to possible misalignments and incorrect consensus calling around these regions. Partitioning was performed using the Python script written Margaryan et al. (2017), available through Github account.[6] After partitioning, we tested the best-supported substitution models (**Supplementary Table 6**) using jModelTest v.2.1.10 (Darriba et al., 2012) with the Bayesian Information Criterion (BIC). When running BEAST, we used unlinked strict clock rates and substitution models for the four partitions, but linked the coalescent Bayesian skyline tree model. Each partition was assigned an independent mutation rate prior according to Rieux et al. (2014). We ran each MCMC for 10E8 states, sampling every 10E4 states, and designating the first

[4]https://www.mitomap.org

[5]http://eltsov.org

[6]https://github.com/GrantHov/My_Python_codes

10% as burn-in. Three independent runs were combined using LogCombiner v1.8.0 (Drummond et al., 2012). Convergence to the stationary distribution and sufficient sampling and mixing were checked by inspection of posterior samples (effective sample size > 200). Parameter estimation was based on samples combined from different chains. The best supported tree was estimated from the combined samples using the maximum clade credibility (MCC) method implemented in TreeAnnotator v1.8.0 (Drummond et al., 2012).

RESULTS

Y-Chromosome and mtDNA Haplogroup Profile

A total of 20 different Y chromosome sub-haplogroups in 32 Foyemiaowan individuals were determined according to the ISOGG's Y-DNA Haplogroup Tree 2019 (**Table 1**, **Figure 2**, and **Supplementary Figure 2**). The distribution of paternal haplogroups was revealed as: Oα-M117 (25%), Oβ-F46 (18.75%), N-F1206 (18.75%), O-M188 (6.25%), O-F444 + , F46- (6.25%), O1b1a2-page59 (6.25%), Q-M120 (6.25%), C-F1144 (6.25%), N-F2930 (3.13%) and O1b1a1-PK4 (3.13%). Haplogroup Oα-M117 exhibits high frequency in Tibeto-Burman populations (Xue et al., 2006; Gayden et al., 2007), including Deng (63.33%), Shannan Tibetan (42.31%), Luoba (30.77) and Danba Qiang (22.22%) (Kang et al., 2012; Qi et al., 2013; Wang and Li, 2013; Wang et al., 2014) and present-day Han Chinese (Yan et al., 2011) such as Shandong (24.2%), Beijing (18.5%) and Heilongjiang (16.6%) (Lang et al., 2019). Oβ-F46 is observed at high frequency with ∼11% in Han Chinese (Yan et al., 2011) and moderate to low frequency for Tibeto-Burman (Xinlong Tibetan, 6.52%; Yajing Tibetan, 6.38%; Danba Qiang, 5.56%; Nyingchi Tibetan, 4.27%) groups. However, the haplogroup Oγ-IMS-JST002611 occupying about 14% in Han Chinese, was not detected in the Foyemiaowan population, which suggests that the Foyemiaowan paternal gene pool was somewhat different from Han Chinese. N-F1206, achieving the second high frequency, is widely distributed across Northern Eurasia (Hu et al., 2015) and can be further divided into N-TAT and N-F710. N-TAT exhibited a very high frequency in Altaic (Yakuts, 70.8–91.5%; Evenks, 50.9%; Buryats, 41.4%), Uralic (Udmurts, 66.7%; Finns, 53.8%) and Indo-European (Latvians, 43.0%; Lithuanians, 40.5%) speaking populations (Ilumäe et al., 2016). N-F710 is considered as migrating from Northeast Asia to the Yellow River basin no later than 2.7 kya (Ma et al., 2021). Moreover, the haplogroups O1b1a2-Page59 (6.25%) and O1b1a1-PK4 (3.13%) reflect a minor degree of southern East Asian origins (Yan et al., 2011; Xia et al., 2019).

Library construction of 10 samples failed because of poor sample quality. As a result, we only obtained 24 valid mitochondrial gene data from 34 Foyemiaowan samples. A total

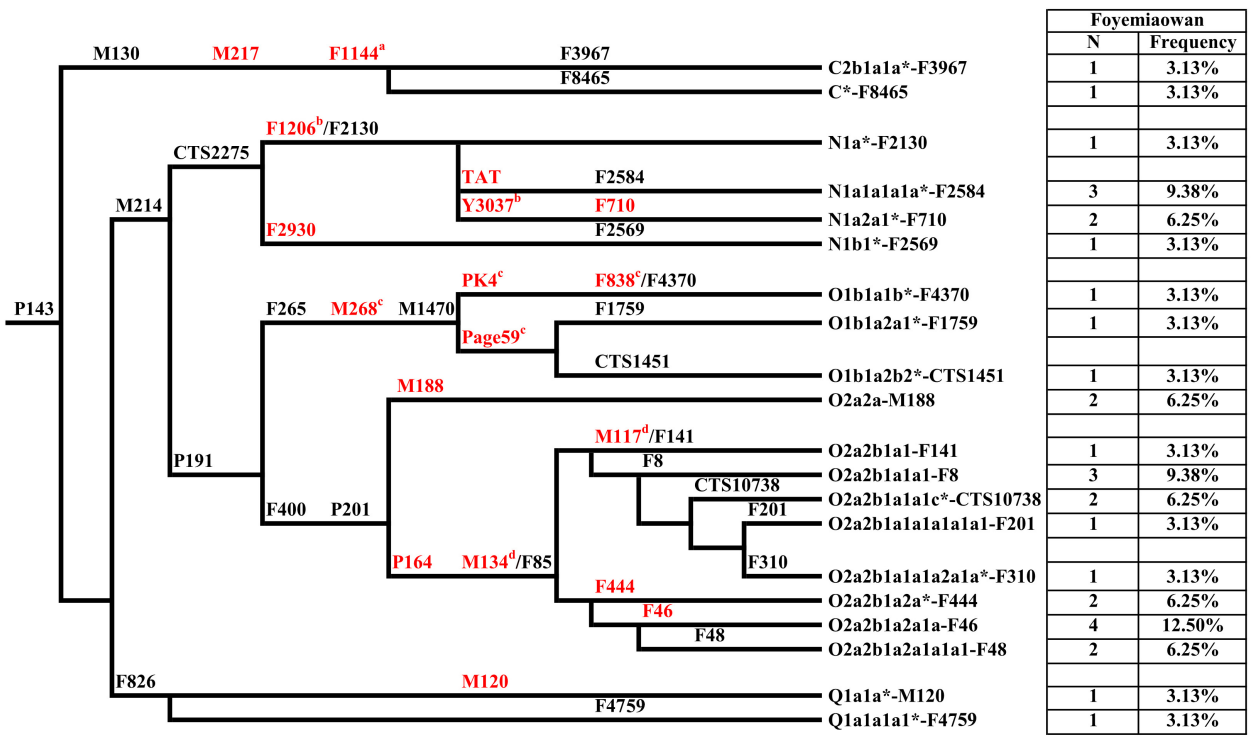
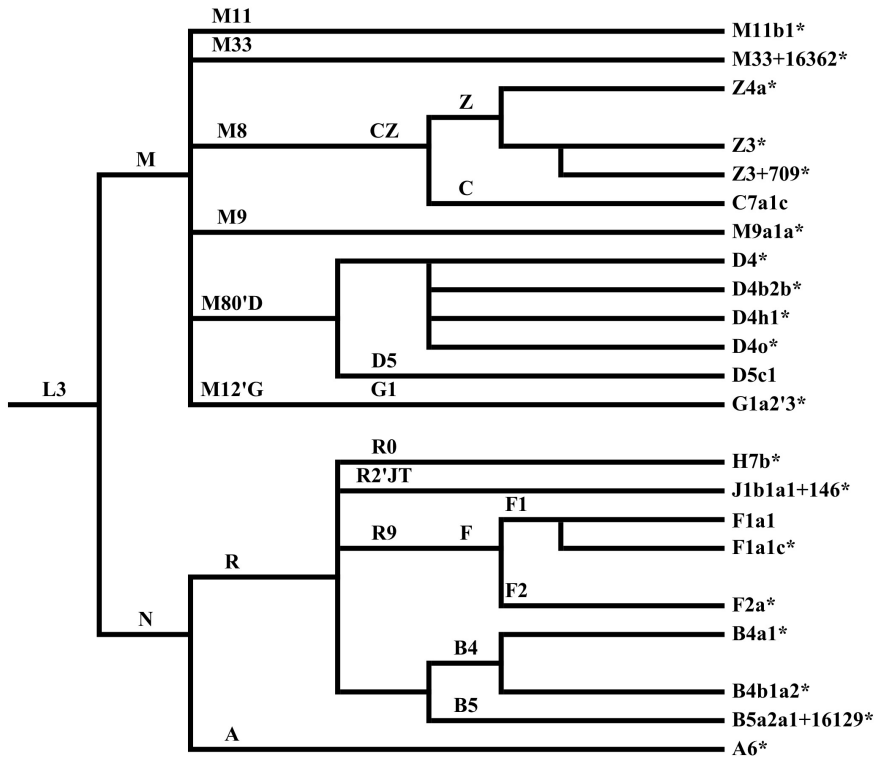


FIGURE 2 | Phylogenetic relationship of Y-chromosome haplogroups in this study, with frequencies are on the right side of the figure. Marker names are shown along the branches and haplogroup names are shown to the right. All Y-chromosome haplogroups are based on the ISOGG-2019. Potentially paraphyletic undefined subgroups are distinguished from recognized haplogroups by the asterisk symbol. Markers in red are key Y-SNPs. a–d: these markers are not designed in the Y-SNP panel but are very common in Y phylogenetic trees.



Foyemiaowan	
N	Frequency
1	4.17%
1	4.17%
1	4.17%
1	4.17%
1	4.17%
1	4.17%
1	4.17%
1	4.17%
1	4.17%
1	4.17%
1	4.17%
1	4.17%
1	4.17%
1	4.17%
1	4.17%
1	4.17%
2	8.33%
2	8.33%
1	4.17%
1	4.17%
1	4.17%
1	4.17%

FIGURE 3 | The phylogenetic relationship of mtDNA haplogroups surveyed in this study and their haplogroup-based frequencies in the sampled populations, with haplogroup names shown on the right. Asterisks distinguish potentially paraphyletic undefined subgroups from recognized haplogroups.

of 22 different mtDNA sub-haplogroups were found in 24 individuals (**Table 1** and **Supplementary Table 1C**). Haplogroup frequencies for Cao Wei to Western Jin, Sixteen Kingdoms, Sui to Tang and the entire Foyemiaowan mtDNA data are presented in **Supplementary Figure 3** and **Figure 3**. The entire population of Foyemiaowan is mainly characterized by the most typical eastern Eurasian mtDNA haplogroups (A, B, CZ, D, F, G and M9). Here, Foyemiaowan shows pronounced frequencies (>5%) of haplogroups D (20.83%), F (20.83%), CZ (16.67%), and B (12.5%) and low frequencies (< 5%) of haplogroups A (4.17%), G (4.17%), and M9 (4.17%). Notably, the haplogroup D4 occurs with the highest frequency (16.67%) in the Foyemiaowan population and could be further divided into sub-clades D4*, D4b2b* and D4h1*. D4 haplogroup is observed with high frequency in populations from the Baikal Region (30.9%), Far East (26.6%), Eastern Asia (22.5%), Altai Region (17%), northern Asia (16.7%) and western Siberia (16.4%) (Derenko et al., 2010). Additionally, the haplogroup D4 reaches a high frequency in Han populations from northern China, including Inner Mongolia (22.73%), Liaoning (21.83%) and Heilongjiang (21.51%) (Li et al., 2019). Haplogroup Z (accounting for 12.5% in Foyemiaowan) is most frequent in Northeastern Asia and Siberia populations (Schurr et al., 1999) as well as the Volga-Ural region, such as Udmurts-7.1%, Maris-2.7%, Komis-1.8% (Tambets et al., 2004; Ingman and Gyllensten, 2007). As for haplogroups B and F, these are relatively common in southern East Asia (Wen et al., 2005; Ko et al., 2014; Kutanan et al., 2017; Li et al., 2019). In the meantime, a small number of western Eurasian origin haplogroups H (4.17%) and J (4.17%) were detected. Of these, haplogroup H displays a broadly southeast-northwest distribution pattern in western Eurasia (Pereira et al., 2005) and dominates the mtDNA gene pool of Europeans (∼40–45% on average), decreasing in the Near East and the Caucasus region (∼20–30%). Haplogroup J shows a moderate frequency in Europe (6–15%) (Pliss et al., 2006; Vidrová et al., 2008), and among Norwegians (12.6%), Estonians (10.3%), Germans (8.4%), and Russians (8.0%) (Pliss et al., 2006). In summary, the dominant D4, Z, B, and F mtDNA haplogroups in Foyemiaowan are more frequent in northern and southern East Asian populations, respectively, while haplogroups J and H reveal genetic flow from western Eurasia.

Possible Origins and mtDNA Haplogroup Diversity

In order to compare mtDNA haplogroup diversity and infer possible origins of the Foyemiaowan population across a broad geographical context, we collated 564 individuals from 28 ancient populations, ranging from circa 2000 BCE to 600 CE, and covering Northern China, Mongolia, southern Siberia and

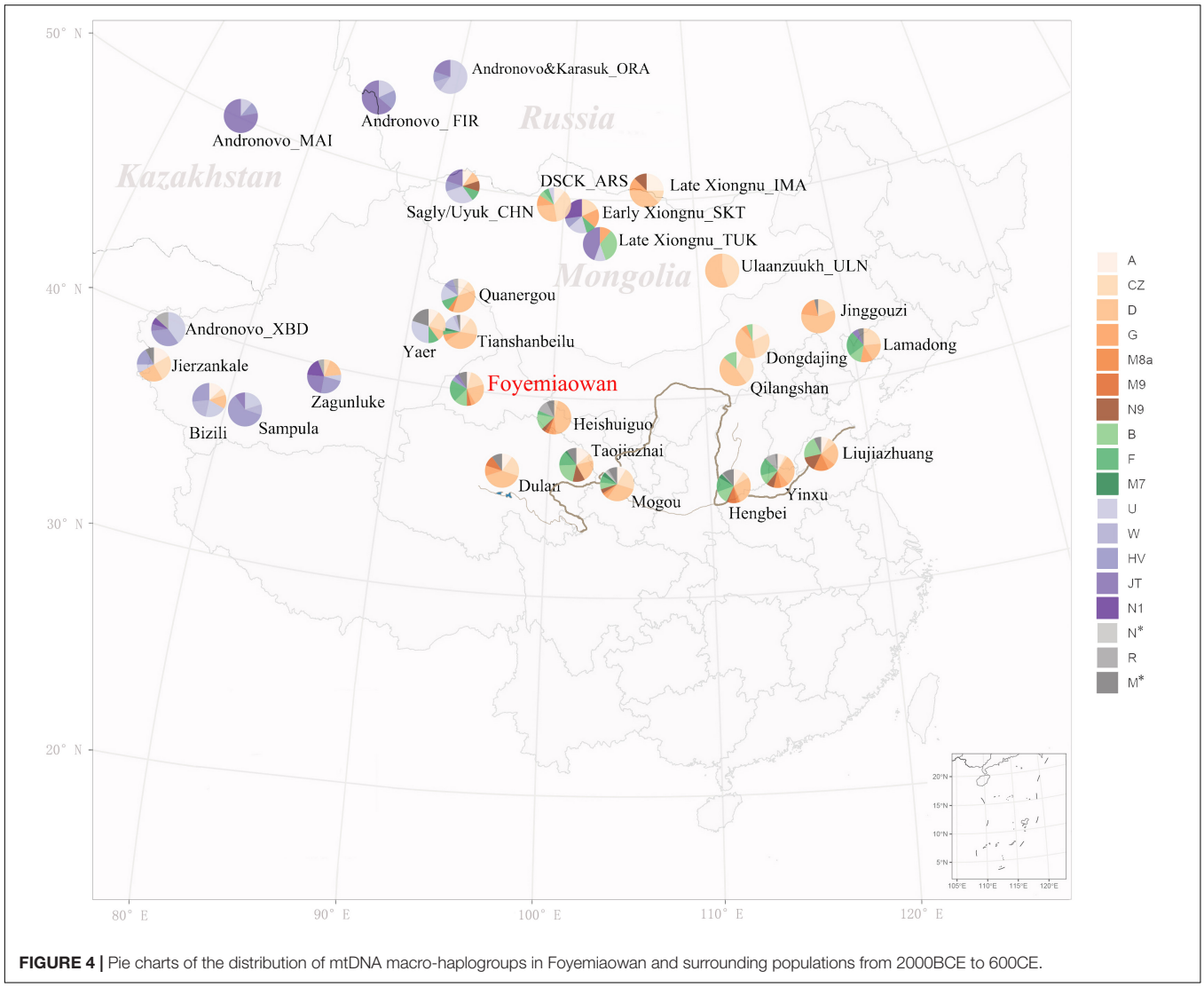


FIGURE 4 | Pie charts of the distribution of mtDNA macro-haplogroups in Foyemiaowan and surrounding populations from 2000BCE to 600CE.

eastern Kazakhstan (**Supplementary Table 2A**). As an initial step, only those ancient populations with a sample size of at least eight and timescales falling mostly within 300 years were selected for further analysis. Secondly, mtDNA haplogroups were classified into 18 macro-haplogroups, including haplogroups A, B, CZ, D, F, G, M7, M8a, M9, N9, U, W, HV, JT, N1, M*, N*, and R. Thirdly, going by distinct geographic origins, these haplogroups were further divided into four groups: northern East Asian (i.e., A, CZ, D, G, M8a, M9, and N9), southern East Asian (i.e., B, F, and M7), western Eurasian (i.e., U, W, HV, JT, and N1) and Undetermined (i.e., M*, N*, and R). Finally, haplogroup diversity in these ancient populations was computed at the level of the 18 macro-haplogroups mentioned above.

The distribution of mtDNA macro-haplogroups (**Figure 4**) revealed a differential contribution of haplogroups with three distinct geographic origins. As expected, northern East Asian (NEA) haplogroups, with a notable incidence of D, C, G and A, are more frequent in mid-eastern Mongolia and eastern Inner Mongolia regions. These haplogroups reach the highest proportion (∼100%) in Ulaanzuukh Culture and late Xiongnu Culture (central east Mongolia) populations. In the west, the occurrence of western Eurasian (WEu) haplogroups is higher and tends to decrease to the east, occurring with the highest frequency (∼100%) in western Xinjiang, eastern Kazakhstan and Southern Siberia. In the south, the presence of southern East Asian (SEA) haplogroups is notable, especially in Yellow River Valley populations such as Taojiazhai (37.2%) and Foyemiaowan (33.3%). Here, the Foyemiaowan population consisted of 50% NEA, 33.3% SEA, 8.3% WEu and 8.4% Undetermined haplogroups (**Supplementary Table 2B**), indicating the possibility of three main sources of ancestry in the above-mentioned geographical areas. Finally, examining haplogroup diversity value (**Table 2**), our highest figure (0.879–0.911) was observed at areas lying at the geographic crossroads between Western and Eastern

Steppe, including central west Mongolia and Xinjiang, centers of population migration and mixture over the past four to five millennia (Allentoft et al., 2015; Unterländer et al., 2017; Damgaard et al., 2018). Foyemiaowan exhibited the third highest (0.891) level of diversity, revealing the multi-ancestral admixture history associated with its strategic position along the Silk Road.

TABLE 2 | Haplogroup diversity value for population mitochondrial genomes.

Population	Location	Size	No. haplogroup	Haplogroup diversity
Foyemiaowan	Gansu, China	24	10	0.891304348
Foyemiaowan_Sixteen Kingdoms	Gansu, China	13	7	0.884615384
Heishuiguo	Gansu, China	27	9	0.780626781
Bizili	Xinjiang, China	15	6	0.866666667
Jierzankale	Xinjiang, China	12	6	0.878787879
Quanergou	Xinjiang, China	20	8	0.852631579
Tianshanbeilu	Xinjiang, China	29	9	0.815270936
Yaer	Xinjiang, China	10	6	0.888888889
Zagunluke	Xinjiang, China	17	7	0.867647059
Sampula	Xinjiang, China	10	4	0.644444444
Andronovo_XBD	Xinjiang, China	15	5	0.752380953
Dulan	Qinghai, China	10	6	0.844444444
Taojiazhai	Qinghai, China	43	8	0.866002214
Mogou	Gansu,China	47	12	0.854764108
Liujiazhuang	Shandong, China	14	8	0.912087913
Hengbei	Shanxi, China	72	12	0.870892019
Yinxu	Henan, China	41	10	0.864634146
Jinggouzi	Inner Mongolia, China	26	4	0.615384616
Qilangshan	Inner Mongolia, China	15	4	0.695238095
Dongdajing	Inner Mongolia, China	17	5	0.75
Lamadong	Inner Mongolia, China	17	7	0.875
DSCK_ARS	Khuvsgul, Mongolia	19	7	0.807017543
Late Xiongnu_TUK	Arkhangai, Mongolia	9	4	0.750000001
Sagly/Uyuk_CHN	Uvs, Mongolia	10	7	0.911111111
Late Xiongnu_IMA	Buryatia, Russia	8	5	0.857142857
Early Xiongnu_SKT	Khuvsgul, Mongolia	11	6	0.89090909
Ulaanzuukh_ULN	Sukhbaatar, Mongolia	9	2	0.555555555
Andronovo_ FIR	Altai Krai, Russia	22	3	0.554112554
Andronovo&Karasuk_ORA	Krasnoyarsk Krai, Russia	10	4	0.644444444
Andronovo_MAI	Almaty Oblysy, Kazakhstan	9	3	0.416666666

Haplogroup diversity was calculated based on haplogroups A, CZ, D, G, M8a, M9, N9, B, F, M7, U, W, HV, JT, N1, N, R, M*.*

Relationship of Foyemiaowan to Reference Populations

To visualize the relationships between the Foyemiaowan population and reference populations, a PCA plot was constructed according to haplogroup frequencies (**Figure 5** and **Supplementary Figure 4**). Firstly, the Foyemiaowan population was organized into two main temporal groups: a "Sixteen Kingdoms" and an "Overall" group. From the Y-chromosome perspective, the reference populations included 11,940 samples from 172 populations (**Supplementary Table 3A**). When plotted (**Figure 5A**), the overall Foyemiaowan group was projected on the Northern Han and Tibeto-Burman cline and clustered closer around Tibeto-Burman populations, especially with Qiangic people from Daofu County, Sichuan and Ü-Tsang Tibetans from Shigatse on the Tibetan Plateau. Specifically, when compared with the "Overall" Foyemiaowan group, the "Sixteen Kingdoms" group was more closely related to Tibetan populations, particularly Ü-Tsang Tibetans from Shigatse and Nyingchi. Traditionally, Ü-Tsang is perceived as the cultural heartland of the Tibetan people, comprising Nagqu, Lhasa, Shannan and Shigatse. A greater degree of genetic flow from Tibetans from the core Tibet region toward Foyemiaowan may have occurred during the Sixteen Kingdoms period. As for the mtDNA perspective, we collected 613 samples from 30 ancient populations (**Supplementary Table 3C**) and 32,777 samples from 139 modern populations (**Supplementary Table 3B**). From the mtDNA PCA plot (**Figure 5B**), we were able to observe both "Overall" and "Sixteen Kingdoms" Foyemiaowan populations clustering closely with each-other as well as Mongolic-speaking populations, especially Dongxiang and Baoan from Linxia in Gansu, Mongolia from Baotou of Inner Mongolia, Daur from northeastern China, Mongolia from Khovd aimag in Mongolia and Mongolia from Dornogovi aimag in Mongolia. Thus, the Foyemiaowan population shows close genetic ties with Mongolic-speaking populations in the north and east of modern Gansu.

In summary, interpopulation comparison shows close affinity between the Foyemiaowan population and Tibeto-Burman populations in terms of paternal structure and Mongolic populations in terms of maternal structure. This might be a consequence of long-term population migration and mixture along the ancient Silk Road.

The Phylogeography of mtDNA Haplogroups

To discern the fine structure of this Foyemiaowan population, we performed phylogenetic analyses for 14 mtDNA haplogroups (**Supplementary Table 5** and **Supplementary Figures 5–17**) using published mitogenome data, including A6, D4*, D4b2b*, D4h1c, D4o, G1a2'3, M9a1a*, Z3*, Z3 + 709, Z4a, F1a1c, F2a, H7b*, and J1b1a1 + 146. Based on the geographic origins and phylogenetic trees (**Supplementary Table 5**), we were able to group these 14 haplogroups into three groups: northern East Asian (i.e., A6, D4*, D4b2b*, D4h1c, D4o, G1a2'3, M9a1a*, Z3*, Z3 + 709, and Z4a), southern East Asian (i.e., F1a1c and F2a) and western Eurasian (i.e., H7b* and J1b1a1 + 146). Following this,

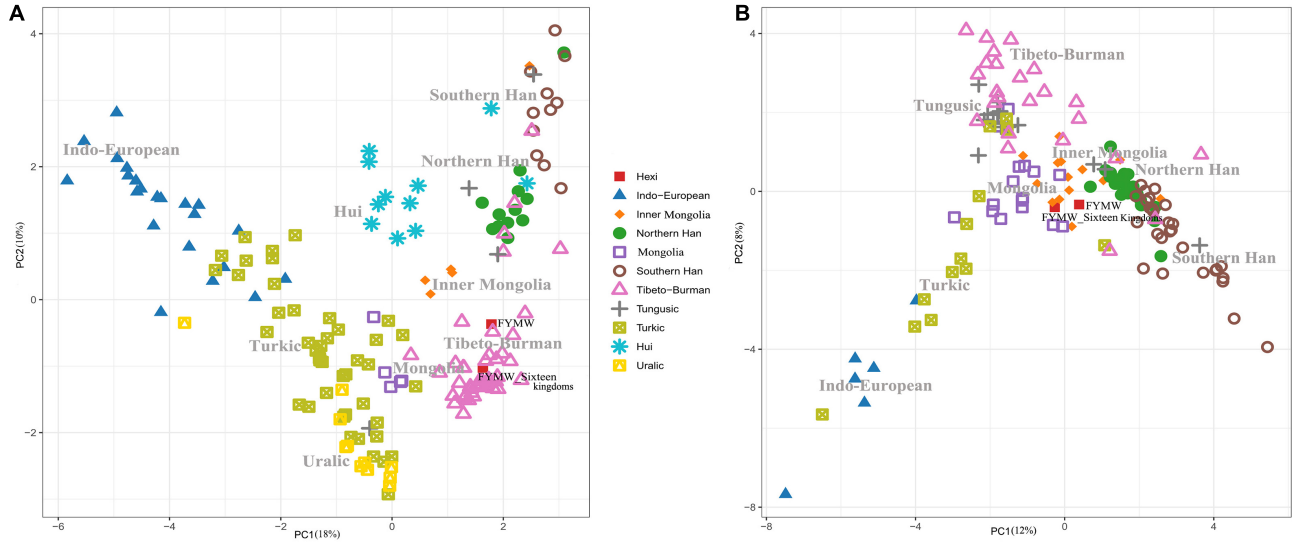


FIGURE 5 | (A) PCA plot based on Y-haplogroups frequency of Foyemiaowan population and reference populations. **(B)** PCA plot based on mtDNA haplogroups frequency of Foyemiaowan population and reference populations.

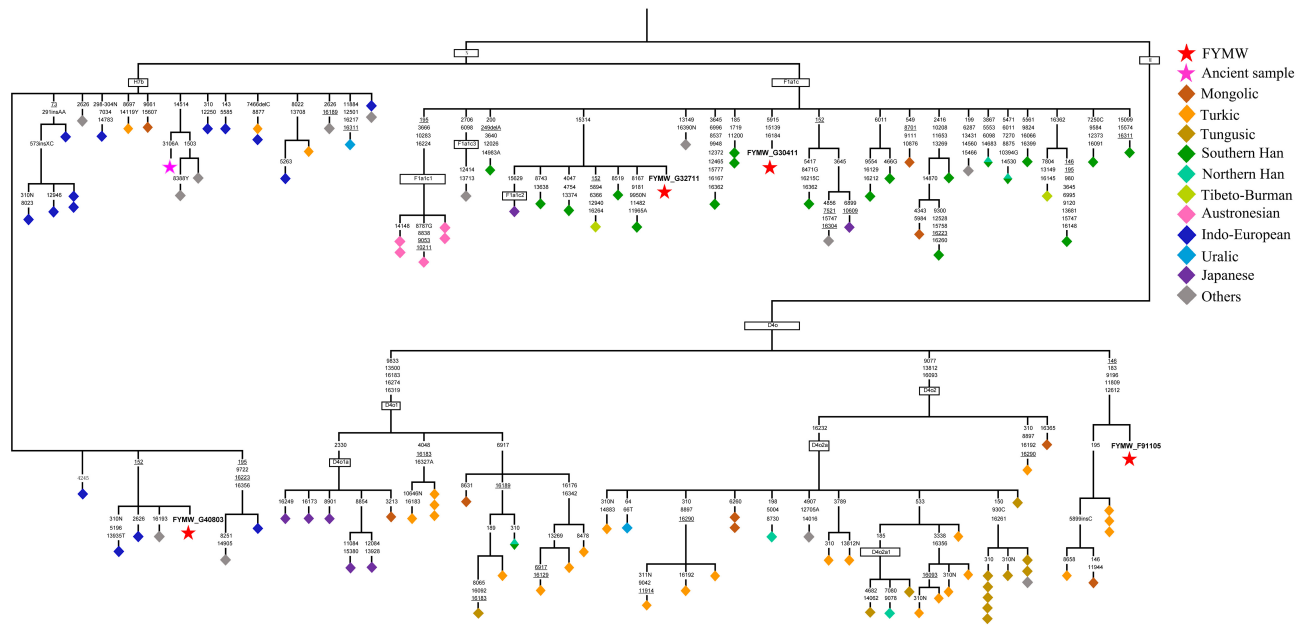


FIGURE 6 | Maximum-parsimony phylogenetic trees of complete mtDNA sequences belonging to haplogroup Haplogroups D4o, F1a1c, and H7b*. Transversions are further specified; ins and del denote insertions and deletions of nucleotides, respectively, back mutations are underlined.

three representative haplogroups (i.e., D4o, F1a1c, and H7b*) were selected from these three groups for further study.

The phylogeny of D4o sequences is illustrated in **Figure 6**. The sequence divergence of the 55 D4o complete genomes corresponds to a coalescence time estimate of 14.19 (8.79–20.65) kya (**Supplementary Table 7**). The D4o tree shows an initial split into two sister subclades, D4o1 and D4o2, encompassing predominantly Turkic, Tungusic, Mongolic and Japonic northern Asian language groups. In particular, three Teleut from the Altai region of southern Siberia, one Buryat from southern Siberia, one Uyghur from the Turpan region of eastern Xinjiang and one from Foyemiaowan clustered into a new branch, harboring the diagnostic motif 146 (back mutation)-183-9196-11809-12612, which implies tight genetic ties between southern Siberian and Foyemiaowan populations. The phylogenetic tree of F1a1c, with the coalescence time estimated as 12.84 (9.01–18.14) kya, could be further subdivided into F1a1c1, F1a1c2, and F1a1c3. F1a1c mostly manifests in Chinese and Austronesian speaking populations in southern China and Southeast Asia. Other subclades included three southern Han from south of the

Yangtze River region in China, one southern Han from Zhejiang in southern China, one Naxi from southwestern China, one Japanese and one from Foyemiaowan, falling into one distinct subclade characterized by one coding region mutation at np 15629. As for the basal branch H7b*, it was sporadically found in Indo-European speaking populations in northeastern Europe and central Asia, which shows a relatively younger coalescence time at 4.48 (2.52–7.56) kya. This haplogroup tree contained one Danish, one Russian from Vladimir near to Moscow, one individual with undetermined geographic origin and one from Foyemiaowan, forming a specific branch with one mutation at np 152 within control region. This may be indicative of western Euraisa influx into gene pool of populations in Hexi Corridor along the ancient Silk Road.

DISCUSSION

Our study supports the hypothesis that Foyemiaowan population have multiple potential ancestral sources. In our study, we analyzed the Y-chromosome and mtDNA of 34 Cao-Wei to Sui-Tang period individuals from the Foyemiaowan site, located in the west of the Hexi Corridor. After interpopulation comparison using PCA plots, we observed that the Foyemiaowan population was closely clustered with Tibeto-Burman populations on the paternal side, but intimately associated with Mongolic-speaking populations from the maternal aspect. Furthermore, we also observed the fine structure of Foyemiaowan population via lineage analysis. Y chromosome profiles of Foyemiaowan population revealed mainly Yellow River Valley origins related to Tibeto-Burman and Han Chinese populations, partially North Eurasian origins associated with Altaic speaking population and a small degree of southern East Asian origins. Similar to paternal structure, the maternal gene pool consisted of ancestries from northern East Asian (including haplogroups A, CZ, D, G, and M9), southern East Asian (including haplogroups B and F) and western Eurasian (including haplogroups H and J) groups. We can conclude that the genetic diversity of Foyemiaowan population was relatively high, and that western Eurasia lineages indicate the eastward migration of Hu people from ancient western Regions along the Silk Road.

Previous archeological and genetic studies argue that the central Hexi Corridor was densely populated by immigrants from Middle and Yellow River during the Han Dynasty (Chen et al., 2019; Xiong et al., 2022). This Han cultural element continued to play a predominant role in the Hexi Corridor even after the collapse of the dynasty. We see use of daily-use potteries, tomb guardian vases (镇墓瓶) and bronze mirrors at Foyemiaowan, items exhibiting a marked Han Chinese style (Chen et al., 2022). This study, contrary to what might be expected, found that population composition gradually trended toward diversification in the west of Hexi Corridor. An elevated value of δ15N and greater use of also C3 foods suggest that the dietary structure of Foyemiaowan population differed strikingly from farming populations consisting of migrants from the cultural core of China, such as the Heishuiguo population (Li, 2021), and veered closer to paleodietary patterns observed for nomadic populations (Wang et al., 2016; Zhang et al., 2016). Archeological findings such as painted murals unearthed in the Hexi Corridor and dating to the Cao-Wei and Sixteen Kingdoms periods support our results by showing the prevalence of non-Han customs (Gansu Province Cultural Relics Team, and Gansu Province Museum, 1985; Ma, 2000; E et al., 2009a,b). Other recorded customs also reflected the influence of a non-Han (Hu) population (**Supplementary Figure 18**). Hair style, attire, customs and facial features suggest that this population belonged to different ancient ethnic groups, such as Di peoples, Qiang peoples, Xianbei peoples, and other non-Han groups (Li, 2010).

Historical records further illustrate the multiethnic nature of the Hexi corridor during the Cao-Wei. At this time, the area included Han, Di, Qiang, Hexi Xianbei (i.e., Tufa Xianbei [秃发鲜卑], Yifu Xianbei [乙弗鲜卑], Yiyun Xianbei [意云鲜卑] etc.), Lushuihu (卢水胡), Western Region Hu (西域胡) and Tuge Hu (屠各胡) (Gao et al., 2018). The Han, Qiang and Di had settled in Hexi Corridor during the Han Dynasty (Zhao et al., 2011; Gao et al., 2018; Xiong et al., 2022). The Tufa Xianbei would later emerge as one of most powerful forces among the Hexi Xianbei. Migrating to the Hexi Corridor from the Yin Mountains in 219–256 CE (Zhou, 1987), they established the Southern Liang under Tufawugu (秃发乌孤) in 397CE. At its peak, the Tufa Xianbei commanded most of northwestern China. The Lushuihu settled along the Hexi Corridor during the Han and Wei Dynasties, founding the Northern Liang in 397 CE. This ethnic group was very complex and may have Xiongnu, Xiao Yuezhi (小月氏), Yiqu (义渠), Zahu (杂胡), Western Rong (西戎) or Zilu (赀虏) connections (Gao et al., 2018). Hu peoples from Western Region also can be found along Hexi corridor during Three Kingdoms periods (220–280 CE). The Dunhuang manuscripts (敦煌文书) record intermarriage between Han Chinese and Hu peoples (Lu, 1996). Sogdian correspondence (311CE) unearthed from Dunhuang indicates that a substantial population of Sogdian merchants migrated from Samarkand in this period (Bai, 2011). The discovery of western Eurasian haplogroups in the Foyemiaowan maternal gene pool may provide corroboration of these documents. These ethnic groups lived and intermarried at Dunhuang, strongly affecting the population history of the Hexi Corridor.

CONCLUSION

In conclusion, we have found that despite the similarity between archeological cultures at Foyemiaowan and Central Plains sites, the genetic structure and dietary structure of Foyemiaowan population differed strikingly from those of Han Chinese. This suggests that archeological culture is not consistent with ethnicity at Foyemiaowan. Yet although in this period dynastic governments had gradually relinquished control of the Hexi Corridor, and regional political forces established by various non-Han peoples became the dominant factors of Hexi history, Han cultural factors still strongly affected the Foyemiaowan peoples.

DATA AVAILABILITY STATEMENT

The FASTA files of mtDNA reported in this article have been deposited in the Genome Warehouse in National Genomics

Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation, under accession number: GWHBHOV01000000-GWHBHPS01000000 that is publicly accessible at https://ngdc.cncb.ac.cn/gwh.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Fudan University of Life Sciences. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

GC, HL, and SW designed the study. HL and SW supervised the study. GC, YiY, and HW provided the materials and resources. HM collected the samples. GC and YiY performed the archeological data analysis. JX and YX performed the genetic laboratory work. YT, MB, YaY, and BZ performed the genetic data analysis. JX, HB, YT, and PD integrated the genetic data. JX, PD, SW, YT, and EA wrote and edited the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by the National Key R&D Program of China (2020YFE0201600 and 2020YFC1521607), the National Social Science Fund of China (19VJX074), the National Natural Science Foundation of China (32070576), B&R Joint Laboratory of Eurasian Anthropology (18490750300), the Major Research Program of National Natural Science Foundation of China (91731303), the Major Project of National Social Science Foundation of China granted to Shaoqing Wen (20&ZD212), and the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), the 111 Project (B13016).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.901295/full#supplementary-material

Supplementary Figure 1 | Pattern of DNA degradation.

Supplementary Figure 2 | Phylogenetic relationship of Y-chromosome haplogroups in this study and their frequencies in Cao-Wei and Western Jin, Sixteen Kingdoms, Sui and Tang Dynasties, Undetermined.

Supplementary Figure 3 | Phylogenetic relationship of mtDNA haplogroups in this study and their frequencies in Cao-Wei and Western Jin, Sixteen Kingdoms, Sui and Tang Dynasties, Undetermined.

Supplementary Figure 4 | PCA plot based on mtDNA haplogroups frequency of Foyemiaowan populations and ancient reference populations.

Supplementary Figure 5 | Bayesian maximum clade credibility (MCC) tree of complete mtDNA sequences belonging to haplogroup A6.

Supplementary Figure 6 | Bayesian maximum clade credibility (MCC) tree of complete mtDNA sequences belonging to haplogroup D4.

Supplementary Figure 7 | Bayesian maximum clade credibility (MCC) tree of complete mtDNA sequences belonging to haplogroup D4b2b.

Supplementary Figure 8 | Bayesian maximum clade credibility (MCC) tree of complete mtDNA sequences belonging to haplogroup D4h1c.

Supplementary Figure 9 | Bayesian maximum clade credibility (MCC) tree of complete mtDNA sequences belonging to haplogroup D4o.

Supplementary Figure 10 | Bayesian maximum clade credibility (MCC) tree of complete mtDNA sequences belonging to haplogroup F1a1c.

Supplementary Figure 11 | Bayesian maximum clade credibility (MCC) tree of complete mtDNA sequences belonging to haplogroup F2a.

Supplementary Figure 12 | Bayesian maximum clade credibility (MCC) tree of complete mtDNA sequences belonging to haplogroup G1a2'3.

Supplementary Figure 13 | Bayesian maximum clade credibility (MCC) tree of complete mtDNA sequences belonging to haplogroup H7b.

Supplementary Figure 14 | Bayesian maximum clade credibility (MCC) tree of complete mtDNA sequences belonging to haplogroup J1b1a1 + 146.

Supplementary Figure 15 | Bayesian maximum clade credibility (MCC) tree of complete mtDNA sequences belonging to haplogroup M9a1a.

Supplementary Figure 16 | Bayesian maximum clade credibility (MCC) tree of complete mtDNA sequences belonging to haplogroup Z3.

Supplementary Figure 17 | Bayesian maximum clade credibility (MCC) tree of complete mtDNA sequences belonging to haplogroup Z3 + 709.

Supplementary Figure 18 | (A) Xincheng cemetery (Wei to Jin Dynasties), Burial 6, showing two people picking mulberry (E et al., 2009a). **(B)** Xigou cemetery (Wei to Jin Dynasties), containing two individuals, a male was riding a horse on the left side, a woman with long hair standing on his right side (Ma, 2000). **(C)** Xincheng mural tomb (Wei to Jin Dynasties), Burial 3, depicts two people living inside the yurt, one sleeping, the other cooking (Gansu Province Cultural Relics Team, and Gansu Province Museum, 1985). **(D)** Dingjiazha cemetery (Sixteen Kingdoms), Burial5, showing a man plowing (E et al., 2009b).

REFERENCES

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, 167–172. doi: 10.1038/nature14507

Anthony, D. (2007). *The Horse, the Wheel and Language. How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World*. Princeton, NJ: Princeton University Press.

Bai, X. (2011). The study of ethnic structure in the Hexi Corridor during Wei and Jin Dynasties. *Soc. Sci.* 169, 31–34.

Buikstra, J. E. (1994). *Standards for Data Collection from Human Skeletal Remains, Research Series no. 44*. Fayetteville: Arkansas Archeological Survey.

Caramelli, D., Posth, C., and Rickards, O. (2021). Reconstruction of the human peopling of Europe: a genetic insight. *Ann. Hum. Biol.* 48, 175–178. doi: 10.1080/03014460.2021.1955472

Chen, G. K., Ma, H. L., and Wang, Y. A. (2022). *Dunhuang: The Excavation Report of Foyemiaowan-Xindiantai Cemetery in 2015*. Lanzhou: Gansu Education Publishing House.

Chen, G. K., Yang, Y., and Liu, B. (2019). *Ganzhou, Zhangye: The Excavation Report of Han Dynasty Cemetery in Heishuiguo Site (Volume II)*. Lanzhou: Gansu Education Publishing House.

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. doi: 10.1038/s41586-018-0094-2

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9:772. doi: 10.1038/nmeth.2109

De Angelis, F., Veltre, V., Romboni, M., Di Corcia, T., Scano, G., Martínez-Labarga, C., et al. (2021). Ancient genomes from a rural site in Imperial Rome (1st-3rd cent. CE): a genetic junction in the Roman Empire. *Ann. Hum. Biol.* 48, 234–246. doi: 10.1080/03014460.2021.1944313

Derenko, M., Malyarchuk, B., Grzybowski, T., Denisova, G., Rogalla, U., Perkova, M., et al. (2010). Origin and post-glacial dispersal of mitochondrial DNA haplogroups C and D in Northern Asia. *PLoS One* 5:e15214. doi: 10.1371/journal.pone.0015214

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* 29, 1969–1973. doi: 10.1093/molbev/mss075

E, J., Zheng, B. L., and Gao, G. X. (2009a). *Murals Unearthed from Wei, Jin and Tang Tombs in Gansu Province (Volume I)*. Lanzhou: Gansu University Press.

E, J., Zheng, B. L., and Gao, G. X. (2009b). *Murals Unearthed from Wei, Jin and Tang Tombs in Gansu Province (Volume III)*. Lanzhou: Gansu University Press.

Fei, X. T. (1982). On the question of deepening the development of ethnographic surveys. *J. South Cent. Univ. Natl.* 3, 2–6. doi: 10.19898/j.cnki.42-1704/c.1982.03.001

Fu, Q., Mittnik, A., Johnson, P., Bos, K., Lari, M., Bollongino, R., et al. (2013). A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* 23, 553–559. doi: 10.1016/j.cub.2013.02.044

Furholt, M. (2019). Re-integrating archaeology: a contribution to aDNA studies and the migration discourse on the 3rd millennium BC in Europe. *Proc. Prehistor. Soc.* 85, 115–129. doi: 10.1017/ppr.2019.4

Gansu Province Cultural Relics Team, and Gansu Province Museum (1985). *Jiayuguan Mural Cemetery Excavation Report*. Beijing: Cultural Relics Publishing House.

Gao, R., Jia, X., and Pu, Z. (2018). *Sinification and Minoritization: The National Fusion of Hexi Population in the Han and Tang Dynasties*. Beijing: China Social Sciences Press.

Gayden, T., Cadenas, A. M., Regueiro, M., Singh, N. B., Zhivotovsky, L. A., Underhill, P. A., et al. (2007). The Himalayas as a directional barrier to gene flow. *Am. J. Hum. Genet.* 80, 884–894. doi: 10.1086/516757

Helga, T., James, T. R., and Jill, P. M. (2013). Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–192. doi: 10.1093/bib/bbs017

Hu, K., Yan, S., Liu, K., Ning, C., Wei, L., Li, S., et al. (2015). The dichotomy structure of Y chromosome Haplogroup N. *arXiv [Preprint]* doi: 10.48550/arXiv.1504.06463

Ilumäe, A. M., Reidla, M., Chukhryaeva, M., Järve, M., Post, H., Karmin, M., et al. (2016). Human Y chromosome Haplogroup N: a non-trivial time-resolved phylogeography that cuts across language families. *Am. J. Hum. Genet.* 99, 163–173. doi: 10.1016/j.ajhg.2016.05.025

Ingman, M., and Gyllensten, U. (2007). A recent genetic link between Sami and the Volga-Ural region of Russia. *Eur. J. Hum. Genet.* 15, 115–120. doi: 10.1038/sj.ejhg.5201712

Jones, S. (1997). *The Archaeology of Ethnicity: Constructing Identities in the Past and Present*, 1st Edn. Abingdon: Routledge, 108.

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29, 1682–1684. doi: 10.1093/bioinformatics/btt193

Kang, L., Lu, Y., Wang, C., Hu, K., Chen, F., Liu, K., et al. (2012). Y-chromosome O3 Haplogroup diversity in Sino-Tibetan populations reveals two migration routes into the eastern Himalayas. *Ann. Hum. Genet.* 76, 92–99. doi: 10.1111/j.1469-1809.2011.00690.x

Katoh, K., Rozewicki, J., and Yamada, K. (2017). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* 20, 1160–1166. doi: 10.1093/bib/bbx108

Klales, A. R., Ousley, S. D., and Vollner, J. M. (2012). A revised method of sexing the human innominate using Phenice's nonmetric traits and statistical methods. *Am. J. Phys. Anthropol.* 149, 104–114. doi: 10.1002/ajpa.22102

Ko, A. M. S., Chen, C. Y., Fu, Q., Delfin, F., Li, M., Chiu, H. L., et al. (2014). Early Austronesians: into and out of Taiwan. *Am. J. Hum. Genet.* 94, 426–436. doi: 10.1016/j.ajhg.2014.02.003

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

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

Li, S. H. (2010). Discussing the image of minority ethnic-groups in mural cemetery at Wei and Jin dynasties. *Huaxia Archaeol.* 4, 122–125. doi: 10.16143/j.cnki.1001-9928.2010.04.015

Li, X. (2021). *Human Diets and its Influencing Factors During Han and Jin Periods in the Hexi Corridor and its Adjacent Areas*. Doctoral Dissertation. Lanzhou: Lanzhou University.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinform.* 26, 589–595. doi: 10.1093/bioinformatics/btp698

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

Liu, Y. J. (2012). *Research on Migrants in Hexi During the Han Dynasty*. Doctoral Dissertation. Lznhou: Northwest Normal University.

Lu, Q. F. (1996). Sinification of sogdians at dunhuang from tang to song dynasties. *Hist. Res.* 6, 25–34.

Lv, X. D. (2017). The history of Hexi immigrant during the Wuliang period from the view of a minority culture corridor. *J. Hexi Univ.* 33, 61–65.

Ma, J. H. (2000). *Painted Bricks of Wei and Jin Tombs in Xigou, Jiuquan, Gansu Province*. Chongqing: Chongqing Publishing Group.

Ma, P., Yang, X., Yan, S., Li, C., Gao, S., Han, B., et al. (2021). Ancient Y-DNA with reconstructed phylogeny provides insights into the demographic history of paternal Haplogroup N1a2-F1360. *J. Genet. Genomics* 48, 1130–1133. doi: 10.1016/j.jgg.2021.07.018

Margaryan, A., Derenko, M., Hovhannisyan, H., Malyarchuk, B., Heller, R., Khachatryan, Z., et al. (2017). Eight millennia of matrilineal genetic continuity in the South Caucasus. *Curr. Biol.* 27, 2023–2028. doi: 10.1016/j.cub.2017.05.087

Nei, M., and Tajima, F. (1981). DNA polymorphism detectable by restriction endonucleases. *Genetics* 97, 145–163. doi: 10.1093/genetics/97.1.145

Pamjav, H., Fóthi, Á., Fehér, T., and Fóthi, E. (2017). A study of the Bodrogköz population in north-eastern Hungary by Y chromosomal haplotypes and haplogroups. *Mol. Genet. Genomics* 292, 883–894. doi: 10.1007/s00438-017-1319-z

Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., et al. (2016). EAGER: efficient ancient genome reconstruction. *Genome Biol.* 17:60. doi: 10.1186/s13059-016-0918-z

Pereira, L., Richards, M., Goios, A., Alonso, A., Albarrán, C., Garcia, O., et al. (2005). High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. *Genome Res.* 15, 19–24. doi: 10.1101/gr.3182305

Pliss, L., Tambets, K., Loogväli, E. L., Pronina, N., Lazdins, M., Krumina, A., et al. (2006). Mitochondrial DNA portrait of Latvians: towards the understanding of the genetic structure of Baltic-speaking populations. *Ann. Hum. Genet.* 70(Pt 4), 439–458. doi: 10.1111/j.1469-1809.2005.00238.x

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

Ralf, A., Montiel González, D., Zhong, K., and Kayser, M. (2018). Yleaf: software for human Y-chromosomal haplogroup inference from next-generation sequencing data. *Mol. Biol. Evol.* 35, 1291–1294. doi: 10.1093/molbev/msy032

Renaud, G., Slon, V., Duggan, A. T., and Kelso, J. (2015). Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.* 16:224. doi: 10.1186/s13059-015-0776-0

Renfrew, C., and Bahn, P. G. (2007). *Archaeology Essentials: Theories, Methods, and Practice*. London: Thames & Hudson College.

Rieux, A., Eriksson, A., Li, M., Sobkowiak, B., Weinert, L. A., Warmuth, V., et al. (2014). Improved calibration of the human mitochondrial clock using ancient genomes. *Mol. Biol. Evol.* 31, 2780–2792. doi: 10.1093/molbev/msu222

Schurr, T. G., Sukernik, R. I., Starikovskaya, Y. B., and Wallace, D. C. (1999). Mitochondrial DNA variation in Koryaks and Itel'men: population replacement in the Okhotsk Sea-Bering Sea region during the Neolithic. *Am. J. Phys. Anthropol.* 108, 1–39. doi: 10.1002/(SICI)1096-8644(199901)108:1<1::AID-AJPA1>3.0.CO;2-1

Si, M. Q. Han Dynasty. (2002a). *Shiji*. Changsha: Yuelu Publishing House, 697–698.

Si, M. Q. Han Dynasty. (2002b). *Shiji. Biography of Chulizi and Gan Mao*. Changsha: Yuelu Publishing House, 436.

Sun, X. F., Wen, S. Q., Lu, C. Q., Zhou, B. Y., Curnoe, D., Lu, H. Y., et al. (2021). Ancient DNA and multimethod dating confirm the late arrival of anatomically modern humans in southern China. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2019158118. doi: 10.1073/pnas.2019158118

Tambets, K., Rootsi, S., Kivisild, T., Help, H., Serk, P., Loogväli, E. L., et al. (2004). The western and eastern roots of the Saami–the story of genetic "outliers" told by mitochondrial DNA and Y chromosomes. *Am. J. Hum. Genet.* 74, 661–682. doi: 10.1086/383203

Unterländer, M., Palstra, F., Lazaridis, I., Pilipenko, A., Hofmanová, Z., Groß, M., et al. (2017). Ancestry and demography and descendants of Iron Age nomads of the Eurasian Steppe. *Nat. Commun.* 8:14615. doi: 10.1038/ncomms14615

Upton, D. (1996). Ethnicity, authenticity, and invented traditions. *Hist. Archaeol.* 30, 1–7. doi: 10.1007/bf03373584

Vidrová, V., Tesarová, M., Trefilova, E., Honzík, T., Magner, M., and Zeman, J. (2008). Mitochondrial DNA haplogroups in the Czech population compared to other European countries. *Hum. Biol.* 80, 669–674. doi: 10.3378/1534-6617-80.6.669

Wang, C. C., and Li, H. (2013). Inferring human history in East Asia from Y chromosomes. *Investig. Genet.* 4:11. doi: 10.1186/2041-2223-4-11

Wang, C. C., Wang, L. X., Shrestha, R., Zhang, M., Huang, X. Y., Hu, K., et al. (2014). Genetic structure of qiangic populations residing in the Western Sichuan corridor. *PLoS One* 9:e103772. doi: 10.1371/journal.pone.0103772

Wang, T. T., Fuller, B. T., Wei, D., Chang, X. E., and Hu, Y. W. (2016). Investigating dietary patterns with stable isotope ratios of collagen and starch grain analysis of dental calculus at the Iron Age cemetery site of Heigouliang, Xinjiang, China. *Int. J. Osteoarchaeol.* 26, 693–704. doi: 10.1002/oa.2467

Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H. J., et al. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* 44, W58–W63. doi: 10.1093/nar/gkw233

Wen, B., Li, H., Gao, S., Mao, X., Gao, Y., Li, F., et al. (2005). Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. *Mol. Biol. Evol.* 22, 725–734. doi: 10.1093/molbev/msi055

Wen, S. Q., Bao, R. Y., Zhou, B. Y., Du, P. X., Sun, C., Chen, L., et al. (2019). China National DNA Martyry: a beacon of hope for the martyrs' coming home. *J. Hum. Genet.* 64, 1045–1047. doi: 10.1038/s10038-019-0649-6

Xia, Z., Yan, S., Wang, C., Zheng, H., Zhang, F., Liu, Y., et al. (2019). Inland-coastal bifurcation of southern East Asians revealed by Hmong-Mien genomic history. *bioRxiv [Preprint]* doi: 10.1101/730903

Xiong, J. X., Du, P. X., Chen, G. K., Tao, Y. C., Zhou, B. Y., and Yang, Y. S. (2022). Sex-biased population admixture mediated subsistence strategy transition of Heishuiguo people in Han dynasty Hexi corridor. *Front. Genet.* 13:827277. doi: 10.3389/fgene.2022.827277

Xue, Y., Zerjal, T., Bao, W., Zhu, S., Shu, Q., Xu, J., et al. (2006). Male demography in East Asia: a north-south contrast in human population expansion times. *Genetics* 172, 2431–2439. doi: 10.1534/genetics.105.054270

Yan, S., Wang, C. C., Wang, C. C., Li, H., Li, S. L., and Jin, L. (2011). An updated tree of Y-chromosome haplogroup O and revised phylogenetic positions of mutations P164 and PK4. *Eur. J. Hum. Genet.* 19, 1013–1015. doi: 10.1038/ejhg.2011.64

Zhang, X., Wei, D., Wu, Y., Nie, Y., and Hu, Y. (2016). Carbon and nitrogen stable isotope ratio analysis of Bronze Age humans from the Xiabandi cemetery, Xinjiang, China: implications for cultural interactions between the East and West. *Chin. Sci. Bull.* 61, 3509–3519. doi: 10.1360/N972016-00514

Zhao, Y. B., Li, H. J., Li, S. N., Yu, C. C., Gao, S. Z., Xu, Z., et al. (2011). Ancient DNA evidence supports the contribution of Di-Qiang people to the Han Chinese gene pool. *Am. J. Phys. Anthropol.* 144, 258–268. doi: 10.1002/ajpa.21399

Zhou, W. Z. (1987). *Southern Liang and Western Qin, Xi'an*. Shaanxi: People's publishing house, 8.

Zhu, K., Du, P., Xiong, J., Ren, X., Sun, C., Tao, Y., et al. (2021). Comparative performance of the MGISEQ-2000 and Illumina X-ten sequencing platforms for paleogenomics. *Front. Genet.* 12:745508. doi: 10.3389/fgene.2021.745508

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 | Frontiers in Genetics

BRIEF RESEARCH REPORT
published: 11 July 2022
doi: 10.3389/fgene.2022.902804

Whole-Genome Sequencing and Genomic Variant Analysis of Kazakh Individuals

Ulykbek Kairov[1]†, Askhat Molkenov[1†], Aigul Sharip[1†], Saule Rakhimova[2], Madina Seidualy[1], Arang Rhie[3], Ulan Kozhamkulov[2], Maxat Zhabagin[2], Jong-Il Kim[3], Joseph H. Lee[4], Joseph D. Terwilliger[5], Jeong-Sun Seo[3], Zhaxybay Zhumadilov[2,6] and Ainur Akilzhanova[2]**

[1]Laboratory of Bioinformatics and Systems Biology, Center for Life Sciences, National Laboratory Astana, Nazarbayev University, Nur-Sultan, Kazakhstan, [2]Laboratory of Genomic and Personalized Medicine, National Laboratory Astana, Nazarbayev University, Nur-Sultan, Kazakhstan, [3]Ilchun Genomic Medicine Institute, Seoul National University, Seoul, Korea, [4]Sergievsky Center, Departments of Neurology and Epidemiology, Taub Institute, Columbia University, New York City, NY, United States, [5]Departments of Genetics and Development and Psychiatry, Sergievsky Center, Columbia University, New York City, NY, United States, [6]School of Medicine, Nazarbayev University, Nur-Sultan, Kazakhstan

OPEN ACCESS

Edited by:
Shaoqing Wen,
Fudan University, China

Reviewed by:
Sumi John,
Dasman Diabetes Institute, Kuwait
Naoki Osada,
Hokkaido University, Japan
Massimo Mezzavilla,
Institute for Maternal and Child Health
Burlo Garofolo (IRCCS), Italy

***Correspondence:**
Ulykbek Kairov
ulykbek.kairov@nu.edu.kz
Ainur Akilzhanova
akilzhanova@nu.edu.kz

†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 March 2022
Accepted: 06 June 2022
Published: 11 July 2022

Citation:
Kairov U, Molkenov A, Sharip A, Rakhimova S, Seidualy M, Rhie A, Kozhamkulov U, Zhabagin M, Kim J-I, Lee JH, Terwilliger JD, Seo J-S, Zhumadilov Z and Akilzhanova A (2022) Whole-Genome Sequencing and Genomic Variant Analysis of Kazakh Individuals. Front. Genet. 13:902804. doi: 10.3389/fgene.2022.902804

Kazakhstan, the ninth-largest country in the world, is located along the Great Silk Road and connects Europe with Asia. Historically, its territory has been inhabited by nomadic tribes, and modern-day Kazakhstan is a multiethnic country with a dominant Kazakh population. We sequenced and analyzed the genomes of five ethnic Kazakhs at high coverage using the Illumina HiSeq2000 next-generation sequencing platform. The five Kazakhs yielded a total number of base pairs ranging from 87,308,581,400 to 107,526,741,301. On average, 99.06% were properly mapped. Based on the Het/Hom and Ti/Tv ratios, the quality of the genomic data ranged from 1.35 to 1.49 and from 2.07 to 2.08, respectively. Genetic variants were identified and annotated. Functional analysis of the genetic variants identified several variants that were associated with higher risks of metabolic and neurogenerative diseases. The present study showed high levels of genetic admixture of Kazakhs that were comparable to those of other Central Asians. These whole-genome sequence data of healthy Kazakhs could contribute significantly to biomedical studies of common diseases as their findings could allow better insight into the genotype–phenotype relations at the population level.

Keywords: Kazakh whole genomes, next-generation sequencing, genome analysis, human genetics, Kazakhstan, whole-genome sequence (WGS), whole-genome sequence analysis

1 INTRODUCTION

Kazakhstan is the ninth-largest country in the world situated at the intersection of Europe and Asia. Historically, the territory of Kazakhstan has been inhabited by nomadic tribes, and at the present time, ethnic Kazakhs represent the majority in this multiethnic country.

Advances in next-generation sequencing (NGS) technologies allow sequencing of the entire genomes with a reasonable price, permitting a study of genetic variants in populations of interest. With the implementation of various efficient methods for the analysis of whole-genome sequencing (WGS), more comprehensive characterization of the genome is now possible. Prior to NGS advances, the knowledge of genetic variants relied primarily on the microarray data.

However, shortage of predesigned microarray probes limited the discovery of novel variants (Wong et al., 2014).

Large-scale genomic studies can now address the important issue of the evolutionary history of human populations and comparative genomics (Xing et al., 2010). Unlike the advances in next-generation technologies, the collection of human DNA samples can be slow and unevenly distributed across different populations. Therefore, combined human genetic diversity datasets—the Human Genome Diversity Project (HGDP), HapMap, and 1000G projects—represent only a small subset of the global populations (Xing et al., 2010). Heyer et al. have provided detailed analysis of mitochondrial and Y-chromosome markers that highlighted the differences between different populations in Central Asia that are consistent with the anthropological data (Heyer at al., 2009). Unfortunately, the Kazakh genome data are limited in the aforementioned human genome databases. Only in 2020, Seidualy et al. published the results from genetic variant and admixture analysis using the WGS data of one healthy Kazakh female (Seidualy et al., 2020). However, Narasimhan et al. did report detailed analyses of hundreds of ancient individuals from the Central and South Asian regions, shedding light on genetic exchanges in Eurasia (Narasimhan et al., 2019).

Our study aims to perform the WGS of five healthy Kazakhs to provide insight into the genetic structure and diversity of Kazakhs in Kazakhstan. Specifically, we report WGS analysis of ethnic Kazakhs of both sexes using 30-fold coverage (26x–33x) generated by the Illumina HiSeq2000 NGS platform. The results of genomic analysis provide a valuable contribution to a better understanding of the genetic diversity and the landscape of the Central Asia region. Furthermore, sharing of these WGS data with the scientific community can serve as a valuable resource for comparative population studies and for biomedical studies for investigating disease associations.

2 MATERIALS AND METHODS

2.1 Ethical Consideration, Sample Information, and DNA Extraction

The study protocol has been approved by the Institutional Review Board of the Center for Life Sciences IREC, Nazarbayev University (protocol #3, 4/04/2012). Study participants have agreed to share their genome data for the current and future studies and agreed to release their data to the public databases.

2.2 DNA Extraction, DNA Library Construction, and Whole-Genome Sequencing

Genomic DNA was extracted from peripheral blood using a Qiagen QI Amp mini kit. The concentration and quality of the isolated DNA were quantified using a NanoDrop spectrophotometer (Thermo Fisher Scientific, United States) and Qubit Fluorimeter (Thermo Fisher Scientific, United States). One μg of genomic DNA was used for paired-end DNA library preparation using a Illumina TruSeq DNA Preparation kit following the manufacturer's recommendations (Illumina, United States). DNA libraries have been evaluated *via* Bioanalyzer 2,100 (Agilent Technologies, United States). DNA fragments were hybridized to the flow cell surface using the HiSeq paired-end cluster kit and later were amplified for the formation of clusters using Illumina cBot. The samples were sequenced using the Illumina HiSeq2000 NGS platform.

2.3 Bioinformatics Analysis

2.3.1 Raw Data Preprocessing

Raw data files obtained from Illumina sequencing platforms in binary base call (bcl) format were converted to the fastq file format using the bcl2fastq v.2.20 tool. The quality of the generated sequences has been evaluated using FastQC v.0.11.7 (Andrews, 2010).

2.3.2 Mapping of Sequencing Reads

Reads were aligned to the human reference genome (NCBI GRCh37, hg19) and reference mitochondrial DNA rCRS (NC_012,920) using Burrows–Wheeler Aligner v.0.7.12 (Li and Durbin, 2010) with default options and paired-end mode. Alignments corresponding to specific samples were combined to a single BAM file, and duplicates were marked using Picard tools v.1.130.

2.3.3 Identification of Genomic Variants

Non-duplicate reads have been selected for downstream analysis, and the obtained BAM files were adjusted using base quality score recalibration (BQSR) and variant quality score recalibration (VQSR) procedure with default parameters. BQSR is a data preprocessing step that identifies systematic errors generated by a sequencing machine. VQSR is a complex filtering procedure that helps remove artifacts using machine learning techniques (Augoff et al., 2014). Genome Analysis Toolkit (GATK) v.3.7 and haplotype caller procedure have been used for genomic variant calling (Mckenna et al., 2010).

2.3.4 Genomic Variants Analysis and Functional Annotation

Identified genomic variants were further annotated using ANNOVAR v. 2016Feb01 (Wang et al., 2010). The functional impact of the SNPs was then evaluated using SIFT (Ng and Henikoff, 2003) and PolyPhen (Adzhubei et al., 2010). Nonsynonymous SNPs were considered damaging if SIFT yielded a score ≤0.05 and PolyPhen-2 yielded a HVAR score ≥0.95. For gene-based annotation, three annotation databases (hg19_ALL.sites. 2015_08. txt; hg19-1000g2015_all; and dbSNP v.138/150) have been used. WebGestalt has been used for functional enrichment analysis of identified genetic variants (Wang et al., 2017).

2.3.5 Principal Component Analysis and Admixture

Datasets. We used three datasets—namely, the Human Genome Diversity Project (HGDP), genotype dataset from Jorde lab, and 1000 Genomes Project dataset—for the comparative population analysis of Kazakh samples in relation to worldwide populations (Cavalli-Sforza, 2005; Xing et al., 2010; The 1000 Genomes

TABLE 1 | Alignment statistics of whole-genome sequencing of Kazakh individuals.

Parameter	KAZ_WG2	KAZ_WG4	KAZ_WG5	KAZ_WG6	KAZ_WG7
In total reads	1,064,621,201	987,933,678	899,102,501	864,441,400	927,234,150
Duplicates	118,659,323	442,049,149	202,996,427	141,226,504	131,307,919
Mapped	1,060,168,691	976,604,161	889,481,111	855,464,242	917,897,457
Mapped (%)	99.58%	98.85%	98.93%	98.96%	98.99%
Singletons	2,363,888	8,184,471	7,569,803	6,525,956	7,505,472
Singletons (%)	0.22%	0.83%	0.84%	0.75%	0.81%
Throughput (bp)	107,526,741,301	99,781,301,478	90,809,352,601	87,308,581,400	93,650,649,150
Human Genome Fold Coverage (mapped)	33.10	30.49	27.77	26.71	28.66

Project Consortium, 2012). The Human Genome Diversity Project (HGDP) and Jorde lab dataset are the large-scale studies of human genome diversity. These datasets represent 64 different populations. The 1000 Genomes Project covers 2,500 individuals from 20 different populations worldwide. Principal Component Analysis. GCTA (Genome-wide Complex Trait Analysis) software was used to visualize the relationships between individuals from these datasets and assess the population structure. Prior to admixture analysis across different populations, PLINK v.1.07 (Purcell et al., 2007) was used to prune SNPs that are in high linkage disequilibrium. Subsequently, 113,290 SNPs and 129,777 SNPs were used for analysis. To convert VCF files to PLINK format files and compute Fixation Index (Fst), VCFTOOLS v.0.1.12b was used. Admixture Analysis. We performed the admixture analysis as implemented in ADMIXTURE v.1.23 by setting the number of ancestral population (k) to a range of 5–10 when comparing the membership of each Kazakh genome to dominant population groups (Alexander et al., 2009; Xing et al., 2010).

2.3.6 Maternal and Paternal Ancestry Analysis

Every Kazakh individual was appointed to a unique mitochondrial haplogroup based on the whole sequences of mitochondrial DNA. Mitochondrial genomic variants have been identified using SamTools v.1.2 (Li et al., 2009). Then, haplogrep v.2 has been used for mitochondrial haplogroup identification and visualization (Kloss-Brandstatter et al., 2011). Y-chromosome haplogroups were manually determined for four Kazakh males in our group using ISOOG phylogenetic tree information and Y-chromosome genomic variants were identified. In addition, Yleaf kit v.2.1 (Thermo Fisher Scientific, United States) has been used for the identification and validation of Y-STR markers (Ralf et al., 2018).

3 RESULTS

3.1 Whole-Genome Sequencing and Mapping Results

A total of five healthy Kazakh individuals, including four male and one female sample, from Kazakhstan were recruited, and their samples were sequenced with Illumina HiSeq2000 for achieving 30-fold coverage (**Table 1** , **Supplementary Table S1**). For the alignment and mapping of each sequence generated, the hg19 reference genome was used. For the five Kazakh individuals, the total number of sequenced base pairs varies from 87,308,581,400 to 107,526,741,301, and on average, 99.06% were properly mapped. Sequencing quality was high as measured by the ratio of heterozygous SNVs to homozygous SNVs (Het/Hom) and transition/transversion (Ts/Tv) ranged from 1.35 to 1.49 and from 2.07 to 2.08, respectively. See **Supplementary Table S2** for additional information on sequencing quality.

3.2 Genomic Variants and Functional Annotation

Genetic variants were identified using the Genome Analysis Toolkit (GATK, version 3.7) and haplotype caller. The annotation of genetic variants was performed by SIFT, PolyPhen2, SNPedia, and ClinVar using ANNOVAR. The number of identified SNVs, insertions, and deletions for each individual are represented in **Table 2**.

We identified a total of 15,578,079 SNVs and 2,760,989 indels (1,235,092 deletions and 1,525,897 insertions) in the five sequenced individuals. We identified novel variants that were not previously catalogued in the single-nucleotide polymorphism database dbSNP (avsnp138 and avsnp150). There are on average 247,018 deletions and 305,179 insertions among Kazakh individuals.

ANNOVAR was used to categorize SNPs into groups based on their genomic location and functional annotation (**Supplementary Figures S1, S2**). The frequency distribution of variants based on genome location shows that the majority of genetic variants were detected in intergenic (50.58%) and intronic (40.72%) regions, as expected (**Supplementary Figure S1**). The pie-chart of functional distribution (func.refGene) of all the variants and novel variants (separately) from five Kazakh individuals diagram illustrates that the intergenic (54.25%) and intronic (36.67%) variants were called with the highest frequency (**Supplementary Figure S2**).

We analyzed the length and number of indels based on their genomic locations. Indel sizes of variants were extracted from the VCF files, and their average number is shown in Supplementary Figure (**Supplementary Figure S3**). This figure illustrates log10 values of the count of genetic variants. Therefore, SNP variant number is displayed when indel size is 0. The number of mapped deletions was higher than that of insertions, and the length of deletions was greater than that of insertions.

TABLE 2 | Total genetic variants identified in five Kazakh samples.

Sample	SNV	Novel SNP	MNVs	Novel MNVs	Deletions	Novel deletions	Insertions	Novel insertions
KAZ_WG2	3,158,814	14,898	209,294	29,841	312,682	15,856	312,420	33,077
KAZ_WG4	3,035,717	13,059	193,632	26,850	297,412	14,149	293,352	27,914
KAZ_WG5	3,110,974	14,122	202,611	28,370	306,392	15,069	305,205	30,820
KAZ_WG6	3,141,190	14,385	202,819	28,390	304,443	14,383	306,471	30,764
KAZ_WG7	3,131,384	15,011	204,417	28,801	14,163	28,813	308,449	32,003
AVERAGE	3,115,615.8	14,295	202,554.6	28,450.4	247,018.4	17,654	305,179.4	30,915.6

Our study found 19,555 nonsynonymous somatic SNPs (nsSNPs) in all the five individuals (**Supplementary Table S9**). Out of a common 19,555 nsSNPs, 1,141 were homozygous nonreference (**Supplementary Table S10**) and among them 604 nsSNPs with a read depth higher than 5 were revealed as the private variants (**Supplementary Table S11**). Homozygous nonreference 1,141 nsSNPs were further scrutinized for over-representation analysis by WebGestalt tool. The results of analysis determined a significant enrichment of genes in several KEGG pathways ($p < 0.05$), such as olfactory transduction, complement and coagulation cascades, ATP-bind cassette (ABC) transporter, and ECM–receptor interaction (**Supplementary Table S3**). Only olfactory transduction remained statistically significant after calibrating for multiple testing with the False Discovery Rate (FDR) lower than 0.05.

The density of the identified nsSNPs on genomic regions is demonstrated on **Supplementary Figure S4**. Chromosomes 1 and 11 had a larger number of nsSNPs, more than 300 nsSNPs per chromosome (**Supplementary Figure S4**). Among these SNPs, 34 nsSNPs were predicted to be damaging to the protein product by SIFT, PolyPhen, and Provean (**Supplementary Table S4**, **Supplementary Figure S5**). Chromosomes 11 and 17 had the highest number of genomic variants which were predicted to be damaging. A number of olfactory genes and zinc finger protein family members were commonly found to be damaging.

3.3 Y-Chromosome and mtDNA Haplogroup Analysis

We identified the mitochondrial variants and assessed the mitochondrial haplogroups using HaploGrep (Kloss-Brandstatter et al., 2011). The samples belong to haplogroups H7a1, Z3c, J1b2, F1b1b, and T2b34. Haplogroup H7 was identified mostly in the European population, whereas other haplogroups were principally present in Asia (Wong et al., 2014).

We then analyzed the Y-chromosome haplogroups and haplotypes of Kazakh male individuals using Y-chromosome genetic variants and 17 STR-region identification approaches. Among the four Kazakh male samples examined, different haplogroups of the Y-chromosome were identified (**Supplementary Table S5**). Two Kazakh samples were allocated to haplogroup R1, which were mostly identified in Central Asia. Another two have been designated to haplogroups N1a and O2a. Y-STR region profiles of analyzed samples have been accessible in **Supplementary Table S6**.

3.4 Admixture and Principal Component Analysis

3.4.1 Principal Component Analysis

Principal component analysis (PCA) was applied to evaluate the population structure of the Kazakh samples with representatives of worldwide populations from 1000G, HGDP, and Jorde genomic project for understanding the ancestral origins of the Kazakh population (**Figure 1**). PCA analysis of the Kazakh samples with populations from 1000G databases (The 1000 Genomes Project Consortium, 2012) is demonstrated in **Figure 1A**. We have repeated the PCA analysis with populations from the HGDP project to validate our findings (**Figure 1B**). PCA analysis of the Kazakh individuals with several Central Asian and Eurasian populations selected from the HGDP and Jorde genomic projects is represented in **Figure 1C** and shows the position of the Kazakh samples on the genetic map across Eurasia.

3.4.2 ADMIXTURE Analysis

ADMIXTURE conducts the unsupervised clustering of a great number of samples and at the same time, each individual can be represented as a combination of clusters (Alexander et al., 2009). To evaluate the population admixture for each sample in the worldwide to dominant population groups, we performed admixture analysis of 3,805 samples collected from 1000G, HGDP, and Jorde genomic projects (Cavalli-Sforza, 2005; The 1000 Genomes Project Consortium, 2012) together with Kazakh samples (**Figure 2**). The number of ancestral populations was fixed to range from five to ten. Kazakh individuals are genetically diverse population at the whole-genome level and show similar ancestral patterns with populations from Central Asia. As shown in **Figures 1**, **2**, ethnic Kazakhs comprised European and Asian admixture, as expected.

3.5 Disease Association and Pathways

Variants from all the five healthy (i.e., non-diseased) individuals were further checked for the functional analyses of variants with SNPEDIA databases (Cariaso and Lennon, 2012). Following the annotation effort *via* SNPEDIA, we have extracted only those genetic variants that were common in all the five samples with a selection magnitude ≥1, a recommended subjective ratio by SNPEDIA. A total of 32 variants were found to have a negative impact on disease (**Supplementary Table S7**). Among our samples, we found three genetic variants (rs9300039,

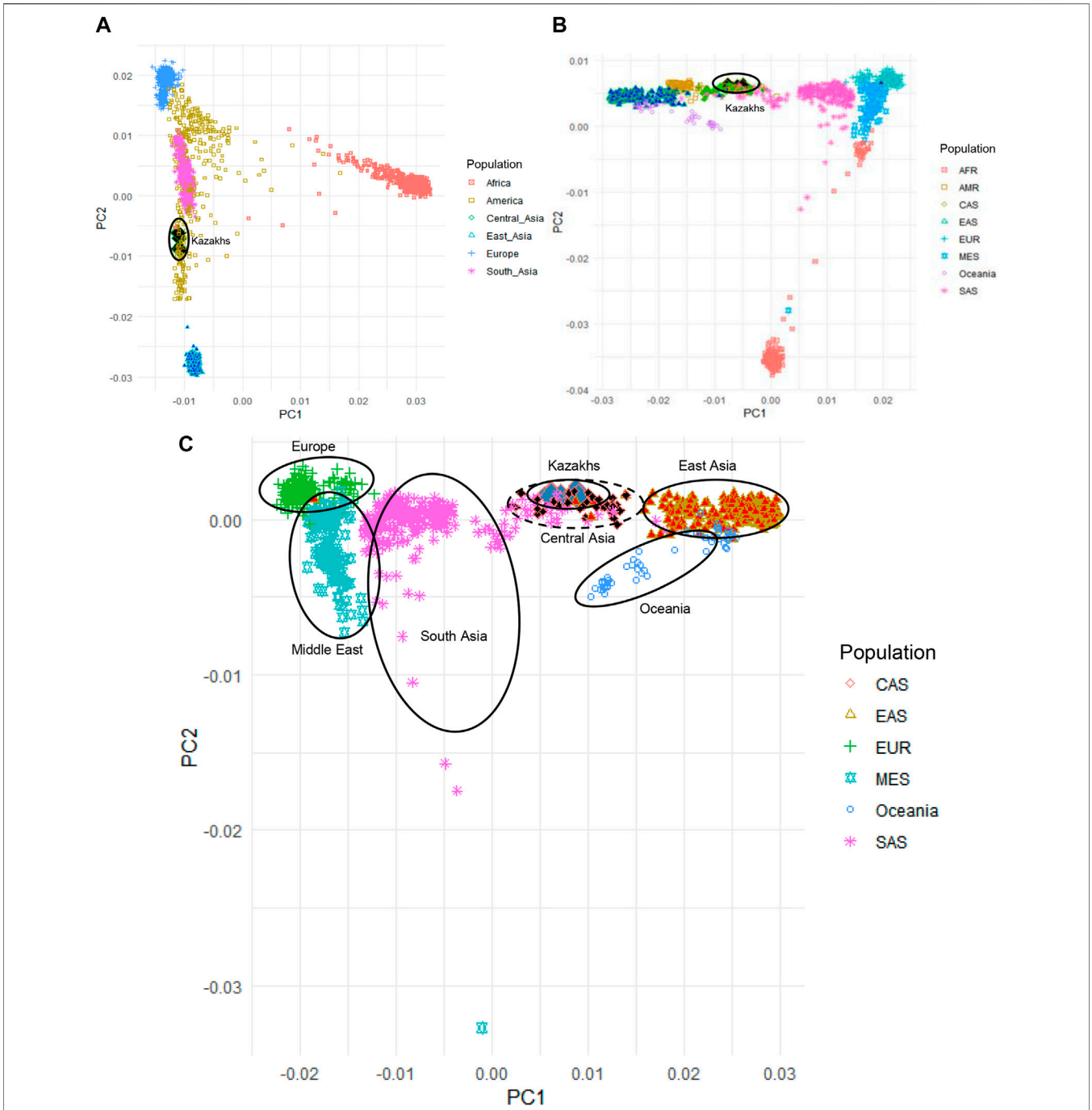


FIGURE 1 | Principal component analysis (PCA) of Kazakh samples along with samples from worldwide populations from 1000G, HGDP, and Jorde genomic projects. **(A)** PCA plot of Kazakh samples and 1000 G project; **(B)** PCA plot of Kazakh samples with the HGDP project; **(C)** PCA plot of Kazakh samples with Eurasian populations from the HGDP and Jorde genomic projects. Kazakh samples are highlighted in blue rhombus. AFR-Africa, AMR-America, CAS—Central Asia, EAS-East Asia, EUR-Europe, MES—Middle East, and SAS-South Asia.

rs11037909 in EXT2 and rs3740878 in EXT2) that potentially increase the risk for type 2 diabetes. We also identified genetic variants associated with metabolic syndrome, hypertension (rs1805762 in M6PR), and increased body weight (rs5746059 in TNFRSF1B). Moreover, several genetic variants were found to increase risk for neurodegenerative diseases, such as Alzheimer's disease (rs4938369 in BACE1) or schizophrenia (rs6932590 in TRNAV27). We note that it is difficult to assess the value of the findings since the identified genetic variants from this small set of non-diseased individuals is inconclusive since 1) they may later

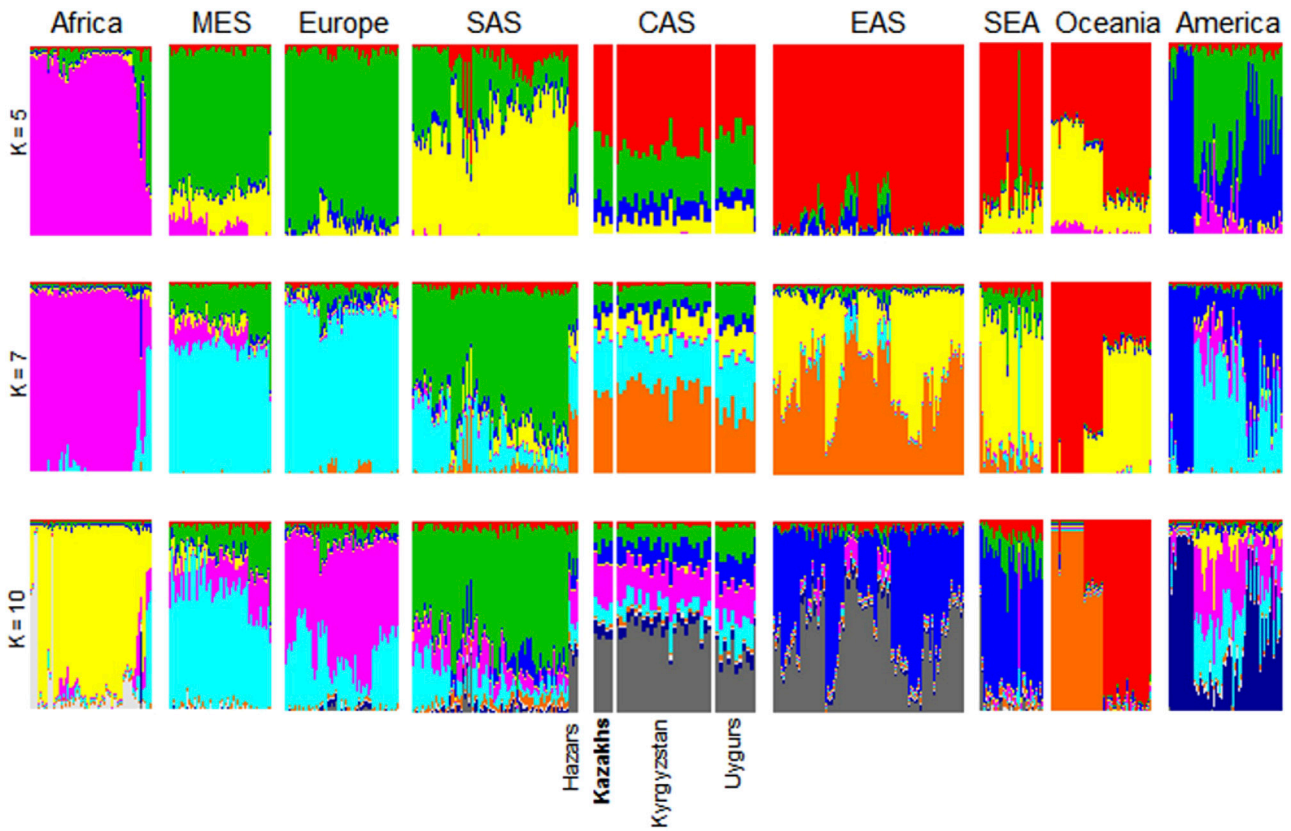


FIGURE 2 | ADMIXTURE plots for Kazakhs assuming a different number of ancestral populations. These figures were generated using genomic data from 3,805 individuals collected from the 1000G, HGDP, and Jorde genomic projects. ADMIXTURE analysis assumed ancestral populations (k) to be 5, 7, or 10. The samples were clustered according to their geographical location to represent their genetic structure. Particularly, Central Asians (i.e., Hazars, Kazakhs, Kyrgyzstan, and Uygurs) had comparable levels of genetic admixture when varying Ks were assumed. AFR-Africa, AMR-America, CAS—Central Asia, EAS-East Asia, EUR-Europe, MES—Middle East, and SAS-South Asia.

develop the disease or 2) the risk variants identified from different populations may have little impact on ethnic Kazakhs due to differing genetic and environmental backgrounds. We will address these possibilities in our future studies when we have an adequate number of clinically diagnosed cases vs. controls.

In addition, we applied the Combined Annotation Dependent Depletion (CADD, C score ≥20) and FATHMM-MLK (Damaged) databases for prediction of the potential functional impact of SNVs. A total of 189 genetic variants affecting 164 different genes met these filtering criteria. We then analyzed overrepresented diseases in these 164 genes by curating information from the DisGeNET database by the utilization of WebGestalt. **Supplementary Table S8** shows the list of overrepresented diseases with p-value <0.05 and corresponding FDR values. There are three significantly affected changes with FDR <0.05, such as muscular dystrophy and other dystrophic changes.

3.6 Data Availability

Sequencing data have been uploaded to the NCBI SRA read archive under accession number PRJNA374772 (Kairov et al., 2021). The VCF file that contains all the genomic variants described in this study is available upon request.

4 DISCUSSION

Advances in NGS technologies have allowed sequencing of the entire genome for potentially large-scale genomic projects. Once such data are generated, and they can be used on a wide range of topics from comparative genomics to other health studies that predict or estimate genetic risks or even to treatment studies. Most existing human genome databases include a limited number of populations and a limited number of individuals within, focusing on European populations. Consequently, no Kazakh genomes are represented in most databases. In this study, we present the comprehensive analysis results of WGS data of ethnic Kazakhs generated using the NGS platform. The WGS data were obtained at high coverage (29.3X on average) for four men and one woman from Kazakhstan. The alignment of the obtained genomic sequences on reference genome hg19 has shown that an average of 99.06% was mapped. Quality assessment of data based on the ratios of heterozygous SNVs to homozygous SNVs (Het/Hom) and transition/transversion (Ts/Tv) ratios demonstrated that our whole-genome sequences were acceptable to the standard practice in the field. This study provides a useful application in biomedicine and makes valuable contributions

to our understanding of the genetic landscape and the diversity of Central Asian populations.

Genetic variant analysis of five sequenced Kazakh individuals has identified a total of 15,578,079 SNVs and 2,760,989 indels with further annotations. Novel genetic variants were identified for each individual (**Table 2**). The SNPs have been categorized based on genomic location and annotation (**Supplementary Figures S1, S2**). Functional analysis of all the common nonsynonymous somatic SNPs (nsSNPs) among five Kazakhs revealed several pathways with significant enrichment of genes (**Supplementary Table S3**). Further analysis of these pathways has demonstrated that only olfactory transduction remained significant after multiple testing corrections with false discovery rate (FDR). This is consistent with an earlier study by Gudbjartsson et al. that showed significant enrichment of olfactory genes in a large-scale WGS of the Icelandics (Gudbjartsson et al., 2015). Genomic mapping of identified genetic variants across human chromosomes have shown that chromosomes 1 and 11 have larger number of nsSNPs (**Supplementary Figure S4**). Altogether, 34 nsSNPs that are predicted to be damaging to the protein product were identified using SIFT, PolyPhen, and Provean prediction methods. Specifically, a number of olfactory gene families (mostly on chromosome 11) and zinc finger proteins were identified.

We have performed PCA, admixture, and identification of haplogroups based on Y-chromosome and mitochondrial DNA. Based on the PCA that compared ethnic Kazakhs to worldwide populations from 1000 Genomes Project, HGDP, and Jorge genome project (**Figure 1**), the first principal component (PC1) distinguished African populations from all other populations, which indicates that the first PC divided populations according to their genetic heterogeneity. The second principal component (PC2) mostly divided the population based on their longitudinal location and heterogeneity (**Figure 1A**). When compared with the populations from the HGDP project, repeated PCA analysis yielded a similar distribution of samples and consistent with the findings from 1000G project, where PC1 distinguished populations based on the longitudinal distribution from west to east (PC1) and genetic diversity of population (PC2; See **Figure 1B**). In addition, we have performed PCA analysis on Kazakhs with several Central Asian and other Eurasian samples selected from the HGDP and Jorde genomic projects to further characterize the position of the Kazakh samples in the genetic maps of human populations in Eurasia (**Figure 1C**). PC1 and PC2 mainly reflected the geographic distribution of the populations, with the majority of genetic variations explained by their locations (Xing et al., 2010). The comparison of populations genetically supports the fact that the Kazakh population is located in the middle of the European and East Asian populations.

Taking the abovementioned results one step further, the assignment of mitochondrial haplogroups of Kazakh individuals demonstrated that four out of five Kazakh samples were set to haplogroups that were prevalent in Asia, and only haplogroup H7a1 was mostly found in Europe. Analysis of Y-chromosome haplogroups of Kazakh men identified four different haplogroups, two of which were assigned to R1 haplogroups, mostly found in Central Asia, and another two were defined as haplogroup N1a and O2a. The diversity of the Kazakh Y-chromosome haplogroups has been reported previously (Tarlykov et al., 2013; Balanovsky et al., 2015; Zhabagin et al., 2017), and a large number of variants were reported. A comparison of the examined male haplotypes with the national database reveals similar variants within five mutational steps for three haplotypes (WG2, WG4, and WG5) (Zhabagin et al., 2019).

Admixture analysis of the population SNP data provided a degree of admixture of populations for each sample. It showed that ethnic Kazakhs are genetically admixed and share similar ancestral patterns to other Central Asians (**Figure 2**). We observed that ethnic Kazakhs were consistently admixed between the populations of Europe and Asia, such as Hazaars, Kyrgyzstan, and Uygurs. The majority of Hazaars, a historical nomadic Turkish tribe, now reside in Pakistan. Although Hazaran samples were gathered from Pakistan, they are genetically similar to Central Asians than to Pakistanis as shown in several studies (Rosenberg et al., 2002; Li et al., 2008). Therefore, Hazarans were grouped together with Uygur and Kyrgyz populations as Central Asians (Hodoglugil and Mahley, 2012). Throughout history, the ethnic Kazakhs have been a nomadic group that has migrated in different regions of Central Asia, leading to a high degree of admixture with local populations (Hodoglugil and Mahley, 2012). Historical migration of the ethnic Kazakhs reflects the genetic structure of the current day ethnic Kazakhs. This study is helpful in identifying the place of the Kazakh population relative to the worldwide populations.

WGSs of ethnic Kazakhs also provide a valuable contribution to biomedical research and further improvement of diagnostics by finding disease associations for various genetic variants. Functional analysis of genetic variants using SNPEDIA databases has identified 32 variants that may have negative effects on various diseases. Many genetic variants were associated with symptoms of metabolic syndrome, such as hypertension (rs1805762), type 2 diabetes (rs9300039, rs11037909, and rs3740878), and increased body weight (rs5746059). Higher risk for neurodegenerative diseases has been linked with several genetic variants (rs4938369 and rs6932590). We have revealed that muscular dystrophy was overrepresented among analyzed ethnic Kazakhs.

In conclusion, this WGS study further characterized the genetic structure and diversity of five unrelated ethnic Kazakhs in comparison to world populations. We showed high genetic admixture of Kazakh genomes at the autosomal level and similar complex genetic heterogeneity of Central Asians. These whole-genome sequences of healthy Kazakh individuals provide invaluable resources for further studies of modern human origin and evolution, causal variants for Kazakh characteristic disease/traits, and personal medicine. The genomic data of a larger number of Kazakh individuals will help answer these questions in the contexts of population research and personalized medicine.

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://www.ncbi.nlm.nih.gov/, PRJNA374772.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board and permission has been acquired from the Center for Life Sciences IREC, Nazarbayev University (protocol #3, 4/04/2012). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

UK and AS wrote the manuscript. UK, AS, AM, MS, and AR contributed to data analysis and implementation of software/code. SR, UK, MZ, and AA contributed to enrolling individuals, experimental works, and validation. UK, J-IK, JL, JT, J-SS, and AA contributed to interpretation and critical revision of the manuscript. UK, JL, and AA conceptualized and supervised the research. UK, AA, and ZZ involved in funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the funding of the Ministry of Education and Science of the Republic of Kazakhstan (AP05135430, AP09563474, PTF O.0703, and AP08855353).

ACKNOWLEDGMENTS

We would like to thank Daniyar Karabayev and Asset Daniyarov for their help in input data preparation and performing multiple sequentially Markovian coalescence analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.902804/full#supplementary-material

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A Method and Server for Predicting Damaging Missense Mutations. *Nat. Methods* 7, 248–249. doi:10.1038/nmeth0410-248

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

Andrews, S. (2010). FASTQC: A Quality Control Tool for High Throughput Sequence Data. Online.

Augoff, K., Hryniewicz-Jankowska, A., Tabola, R., Czapla, L., Szelachowski, P., Wierzbicki, J., et al. (2014). Upregulated Expression and Activation of Membrane-Associated Proteases in Esophageal Squamous Cell Carcinoma. *Oncol. Rep.* 31, 2820–2826. doi:10.3892/or.2014.3162

Balanovsky, O., Zhabagin, M., Agdzhoyan, A., Chukhryaeva, M., Zaporozhchenko, V., Utevska, O., et al. (2015). Deep Phylogenetic Analysis of Haplogroup G1 Provides Estimates of SNP and STR Mutation Rates on the Human Y-Chromosome and Reveals Migrations of Iranic Speakers. *Plos One* 10, e0122968. doi:10.1371/journal.pone.0122968

Cariaso, M., and Lennon, G. (2012). SNPedia: A Wiki Supporting Personal Genome Annotation, Interpretation and Analysis. *Nucleic Acids Res.* 40 (Database issue), D1308–D1312. doi:10.1093/nar/gkr798

Cavalli-Sforza, L. L. (2005). The Human Genome Diversity Project: Past, Present and Future. *Nat. Rev. Genet.* 6, 333–340. doi:10.1038/nrg1596

Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., et al. (2015). Large-scale Whole-Genome Sequencing of the Icelandic Population. *Nat. Genet.* 47, 435–444. doi:10.1038/ng.3247

Heyer, E., Balaresque, P., Jobling, M. A., Quintana-Murci, L., Chaix, R., Segurel, L., et al. (2009). Genetic Diversity and the Emergence of Ethnic Groups in Central Asia. *BMC Genet.* 10, 49. doi:10.1186/1471-2156-10-49

Hodoglugil, U., and Mahley, R. W. (2012). Turkish Population Structure and Genetic Ancestry Reveal Relatedness Among Eurasian Populations. *Ann. Hum. Genet.* 76, 128–141. doi:10.1111/j.1469-1809.2011.00701.x

Kairov, U., Molkenov, A., Rakhimova, S., Kozhamkulov, U., Sharip, A., Karabayev, D., et al. (2021). Whole-genome Sequencing Data of Kazakh Individuals. *BMC Res. Notes* 14, 45. doi:10.1186/s13104-021-05464-4

Kloss-Brandstätter, A., Pacher, D., Schönherr, S., Weissensteiner, H., Binna, R., Specht, G., et al. (2011). HaploGrep: A Fast and Reliable Algorithm for Automatic Classification of Mitochondrial DNA Haplogroups. *Hum. Mutat.* 32, 25–32. doi:10.1002/humu.21382

Li, H., and Durbin, R. (2010). Fast and Accurate Long-Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 26, 589–595. doi:10.1093/bioinformatics/btp698

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: A MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data. *Genome Res.* 20, 1297–1303. doi:10.1101/gr.107524.110

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

Ng, P. C., and Henikoff, S. (2003). SIFT: Predicting Amino Acid Changes that Affect Protein Function. *Nucleic Acids Res.* 31, 3812–3814. doi:10.1093/nar/gkg509

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., 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

Ralf, A., González, D. M., Zhong, K., and Kayser, M. (2018). Yleaf: Software for Human Y-Chromosomal Haplogroup Inference from Next-Generation Sequencing Data. *Mol. Biol. Evol.* 35, 1820. doi:10.1093/molbev/msy080

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., et al. (2002). Genetic Structure of Human Populations. *Science* 298, 2381–2385. doi:10.1126/science.1078311

Seidualy, M., Blazyte, A., Jeon, S., Bhak, Y., Jeon, Y., Kim, J., et al. (2020). Decoding a Highly Mixed Kazakh Genome. *Hum. Genet.* 139 (5), 557–568. doi:10.1007/s00439-020-02132-8

Tarlykov, P. V., Zholdybayeva, E. V., Akilzhanova, A. R., Nurkina, Z. M., Sabitov, Z. M., Rakhypbekov, T. K., et al. (2013). Mitochondrial and Y-Chromosomal Profile of the Kazakh Population from East Kazakhstan. *Croat. Med. J.* 54, 17–24. doi:10.3325/cmj.2013.54.17

The 1000 Genomes Project Consortium (2012). An Integrated Map of Genetic Variation from 1,092 Human Genomes. *Nature* 491, 56–65. doi:10.1038/nature11632

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional Annotation of Genetic Variants from Next-Generation Sequencing Data. *Nucleic Acids Res.* 38, e164. doi:10.1093/nar/gkq603

Wang, J., Vasaikar, S., Shi, Z., Greer, M., and Zhang, B. (2017). WebGestalt 2017: A More Comprehensive, Powerful, Flexible and Interactive Gene Set Enrichment Analysis Toolkit. *Nucleic Acids Res.* 45 (W1), W130–W137. doi:10.1093/nar/gkx356

Wong, L. P., Lai, J. K., Saw, W. Y., Ong, R. T., Cheng, A. Y., Pillai, N. E., et al. (2014). Insights into the Genetic Structure and Diversity of 38 South Asian Indians from Deep Whole-Genome Sequencing. *PLoS Genet.* 10, e1004377. doi:10.1371/journal.pgen.1004377

Xing, J., Watkins, W. S., Shlien, A., Walker, E., Huff, C. D., Witherspoon, D. J., et al. (2010). Toward a More Uniform Sampling of Human Genetic Diversity: A Survey of Worldwide Populations by High-Density Genotyping. *Genomics* 96, 199–210. doi:10.1016/j.ygeno.2010.07.004

Zhabagin, M., Balanovska, E., Sabitov, Z., Kuznetsova, M., Agdzhoyan, A., Balaganskaya, O., et al. (2017). The Connection of the Genetic, Cultural and Geographic Landscapes of Transoxiana. *Sci. Rep.* 7, 3085. doi:10.1038/s41598-017-03176-z

Zhabagin, M., Sarkytbayeva, A., Tazhigulova, I., Yerezhepov, D., Li, S., Akilzhanov, R., et al. (2019). Development of the Kazakhstan Y-Chromosome Haplotype Reference Database: Analysis of 27 Y-STR in Kazakh Population. *Int. J. Leg. Med.* 133, 1029–1032. doi:10.1007/s00414-018-1859-8

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 | Frontiers in Ecology and Evolution

TYPE Original Research
PUBLISHED 10 August 2022
DOI 10.3389/fevo.2022.932004

OPEN ACCESS

EDITED BY
Gyaneshwer Chaubey,
Banaras Hindu University, India

REVIEWED BY
Sara V. Good,
University of Winnipeg, Canada
Levon Yepiskoposyan,
Armenian National Academy
of Sciences, Armenia

*CORRESPONDENCE
Pengfei Sheng
shengpengfei@fudan.edu.cn
Chuan-Chao Wang
wang@xmu.edu.cn
Shaoqing Wen
wenshaoqing1982@gmail.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 Ecology and Evolution

RECEIVED 29 April 2022
ACCEPTED 18 July 2022
PUBLISHED 10 August 2022

CITATION
Allen E, Yu Y, Yang X, Xu Y, Du P, Xiong J, Chen D, Tian X, Wu Y, Qin X, Sheng P, Wang C-C and Wen S (2022) Multidisciplinary lines of evidence reveal East/Northeast Asian origins of agriculturalist/pastoralist residents at a Han dynasty military outpost in ancient Xinjiang.
Front. Ecol. Evol. 10:932004.
doi: 10.3389/fevo.2022.932004

Multidisciplinary lines of evidence reveal East/Northeast Asian origins of agriculturalist/pastoralist residents at a Han dynasty military outpost in ancient Xinjiang

Edward Allen[1†], Yao Yu[1†], Xiaomin Yang[2†], Yiran Xu[3,4], Panxin Du[5], Jianxue Xiong[5], Dian Chen[6], Xiaohong Tian[7], Yong Wu[7], Xiaoli Qin[1], Pengfei Sheng[1,3,4*], Chuan-Chao Wang[2,8,9,10*] and Shaoqing Wen[1,3,4*]

[1]MOE Laboratory for National Development and Intelligent Governance, Department of Cultural Heritage and Museology, Fudan University, Shanghai, China, [2]Department of Anthropology and Ethnology, School of Sociology and Anthropology, Institute of Anthropology, Xiamen University, Xiamen, China, [3]Institute of Archaeological Science, Fudan University, Shanghai, China, [4]Center for the Belt and Road Archaeology and Ancient Civilizations, Fudan University, Shanghai, China, [5]Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Fudan University, Shanghai, China, [6]Department of Archaeology and Anthropology, University of Chinese Academy of Sciences, Beijing, China, [7]Xinjiang Institute of Cultural Relics and Archaeology, Urumqi, China, [8]State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen, China, [9]State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen, China, [10]Institute of Artificial Intelligence, Xiamen University, Xiamen, China

Han/non-Han interactions were engrained among the border regions of ancient Imperial China. Yet, little is known about either the genetic origins or the lifeways of these border peoples. Our study applies tools from ancient deoxyribonucleic acid (DNA) and stable isotope analysis to the study of a Han dynasty population at the Shichengzi site in modern-day Xinjiang. Isotopic analysis ($\delta^{13}C$ and $\delta^{15}N$) of human (n = 8), animal (n = 26), and crop remains (n = 23) from Shichengzi indicated that dietary patterns among site inhabitants could be split among agro-pastoral and agricultural groups based on differences in the collagen ^{15}N ratios. DNA analysis divided the four Shichengzi samples into two groups, with one group primarily harboring the ancient Northeast Asian (ANA) related ancestry, while the other showed a dominant Late Neolithic Yellow River (YR_LN) related ancestry. Both ancient DNA and stable isotope evidence point to the Northeast Asian origins of pastoralists and East Asian origins of Han agriculturalists, who, nonetheless, shared a single burial space at Shichengzi. This study thus provides clear

evidence for the multiple origins and identities of populations across the porous border represented by the Han Empire and surrounding regions and proposes a new model for the interpretation of border culture in early Imperial China.

KEYWORDS

ancient DNA, isotope analysis, Han population, lifestyle, Silk Road, Western Regions

Introduction

An increasing number of studies seek to combine ancient deoxyribonucleic acid (DNA) work with multiple forms of scientific and classic archeological analysis (Haak et al., 2008; Frei et al., 2015, 2019; Mittnik et al., 2019; Blank et al., 2021; Ingman et al., 2021; Patterson et al., 2022). Integrating DNA and carbon, nitrogen stable isotope analysis has improved our knowledge of ancient origins and human/non-human paleodietary practices (Wilson et al., 2007; Alkass et al., 2013; Ames et al., 2015). Such a multidisciplinary approach can offer fascinating new insights into ancient population history and subsistence in various contexts but has rarely been applied to the bordering regions of Chinese imperial history (e.g., Ning et al., 2019). Employing the research methods of this emerging field, our study opens new vistas for border research through a combination of DNA data with paleodietary and other archeological evidence at a Han dynasty border site.

From 2014 to 2019, Shichengzi (43°36′59.1″N, 89°45′43.2″E, 1,770 m amsl), located in Qitai County, Xinjiang, on the northern slopes of the Tianshan Mountains, was identified as a major Han dynasty garrison and was scientifically excavated (**Figure 1** and **Supplementary Figure 1**). Laid out over approximately 110,000 m^2, C-14 dates suggest a site occupation from the first century BC to the third century AD (Sheng et al., 2020). Multiple architectural features were uncovered and dated to the period of the agricultural garrison. These included the main gate, rammed mud walls, moats, buildings, a wealth of pottery, tiles, weapons, and agricultural implements (Area A, Area C, and Area D). Burials and a single Han-style kiln were excavated in Area B (Tian et al., 2018, 2020). Sheng et al. (2020) subsequently characterized Shichengzi as a "melting pot" of Han and non-Han culture, using archaeobotanical evidence to demonstrate a local adaptation to the Tianshan Mountain settings against relatively amenable climactic conditions circa 2000–1700 BP (Sheng et al., 2021).

From the first century BC to the third century, interactions between the Han and agro-pastoralist empires and peoples of Central Asia AD were considered a high point in early Silk Road cross-cultural communications (Millward, 2013; Spengler, 2019). In China's written histories, Han Emperor Wu (141–87 BC) inspired this dramatic shift circa 119 BC, seeking allies and shoring up Han presence in a quest to subjugate the feared Xiongnu confederation to his direct north (Ban, 1962; Fan, 2000). Subsequent Han emperors strengthened the Han military presence in the "Western Regions" (primarily located in modern-day Xinjiang) (Zhang and Tian, 2015; Li, 2017). Chinese scholars have historically documented and voluminously researched the forced migration of ethnic Han Chinese from the Central Plains region to the empire's western extremities during these decades (see Yang, 1991; Zhu, 2012). Against the diverse ecological backdrop of Central Asia, a stratified Han order was established in agricultural garrisons (*tuntian* 屯田; lit. "fortified agricultural fields") (Luo et al., 2018) and other settlement types. This helped stabilize the Han Empire's political and economic influence on the steppe. In practice, however, inter-migration and exchange between groups are believed to have been widespread. One scholar characterized the "cultural mediation, assimilation, rejection, or integration" between the Han center and its peripheries (Di Cosmo, 2009) around this time. With Shizhengzi, the question of interaction between Han, Xiongnu, and Western Region groups became one of acute interest.

Here, we fuse traditional archeological approaches with C and N isotopes, and ancient DNA data analysis of four individuals at Shichengzi (**Figure 1a**), previously confirmed as a Han dynasty agricultural garrison (*tuntian*) (Sheng et al., 2020). Divergent genetic and isotopic profiles, nonetheless, shared similar burial practices at this site. Considered as a combination of diverging dietary practices around a common burial space and possibly site use, we argue that this "mutualism" was negotiated against a backdrop of Han expansion/regional accommodation and long-established, flexible pastoralist subsistence strategies. Our study contributes to current knowledge by demonstrating one way, in which Han migrants and agro-pastoralists of the Western Regions interacted within the agricultural garrison context and explores the possible consequences for our understanding of Han border formation processes at the northwest borders. The methodological approach combines archeological sciences and history and will be useful in re-evaluating border formation in imperial Chinese history in short-term and long-term contexts.

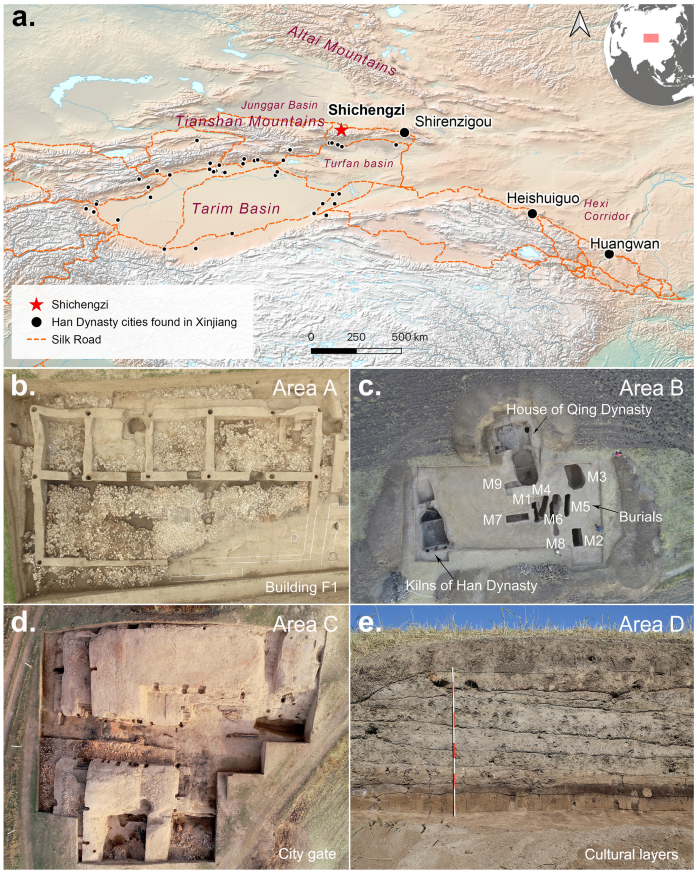


FIGURE 1
(a) Map of northwestern China. Small dots represent Han cities found in Xinjiang. **(b–e)** Main findings at Areas A–D at the Shichengzi site.

Materials and methods

Accelerator mass spectrometry radiocarbon dating

To determine the date of Shichengzi's occupation, human and animal bone, charcoal, and crop seed samples (n = 14) recovered from the site were sent to Beta Analytic Inc., Miami, Florida, United States, for radiocarbon analysis. All dates were calibrated using the IntCal 20 calibration curve (Reimer et al., 2020) and OxCal v4.3.2.[1] Detailed sample information is listed in **Supplementary Table 1**.

Burials

A total of ten burials, one sacrificial horse pit and one pottery kiln, were excavated in the west of the Shichengzi site. The burials were divided into six vertical shaft earth pit tombs, three vertical shaft side-chamber burials, and a vertical shaft burial with a second-tier ledge (*ercengtai*). All were individually prone burials with the head facing west and a trough-shaped coffin. A mortise-tenon structure was employed for the inner-coffin and outer-coffin discovered in the second-tier ledge burial.

The side-chamber burials M1 and M3 contained Han dynasty *wuzhu* coins (M3), a bronze plaque, a bronze ring, an iron knife, and various utilitarian potteries (M1). The burial at M2, the second-tier ledge burial, was entombed in a wooden outer coffin, setting this tomb dramatically apart from M1 and M3. M2 was also buried with goat astragalus bones, silk, lacquer, and beads. The burial style and choice of burial goods strongly suggest two different populations, likely a Han population (like M3) and a pastoralist population (like M2), as pointed out in Sheng et al. (2020).

Stable C and N isotope analysis

The collagen from human bone samples at Shichengzi (n = 8) was extracted using protocols outlined in Richards and Hedges (1999). Carbonized seeds (n = 2)

1 http://c14.arch.ox.ac.uk/oxcal.html

of *Setaria italica* and *Panicum miliaceum* collected from Shichengzi by flotation were also pretreated for stable C and N isotopic determination. We collated published stable carbon and nitrogen isotopic results for multiple varieties of food resources found at Shichengzi, including isotope data (n = 47) of charred wheat/naked barley grains, sheep/goat, cattle, and horse, as well as dog bones from Shichengzi. Additionally, we revisited the existing C and N isotopic evidence for the paleo diet at Shirenzigou (n = 38), dated 2200–1900 BP, and two typical Han cemeteries at Heishuiguo (n = 116) and Huangwan (n = 7), located in the Hexi Corridor during 2100–1700 BP. This would provide comparative data on subsistence strategies. All existing isotopic data are summarized in **Supplementary Tables 2**, **3**.

Deoxyribonucleic acid sampling, extraction, library preparation, and sequencing

Sampling was performed on the teeth of 4 individuals at Shichengzi. Ancient DNA extraction and Illumina double-stranded DNA sequencing library preparation were performed at Fudan University and established protocols for working with ancient human DNA were followed (Knapp et al., 2012; Sun et al., 2021). Human remains were surface-cleaned and ground to a fine powder. We used 100 mg of bone powder to extract the DNA using the robotic magnetic beads method. We prepared double-stranded libraries following Meyer's protocols (Meyer and Kircher, 2010) but with minor modifications and without conducting UDG treatment for all samples. Libraries were amplified with indexing primers in two parallel PCRs using Q5 High-Fidelity DNA Polymerase (NEB). We qualified the clean-up libraries using Qubit 2.0. Finally, we sequenced the libraries on an Illumina HiSeq X10 instrument at the Annoroad Company, China, in the 150-bp paired-end sequencing design.

Sequence data processing

We clipped the Illumina sequencing adapters using AdapterRemoval v2.2.0 (Schubert et al., 2016). We mapped the merged reads with 30 or more bases to the human reference genome (hs37d5) using BWA v0.7.17 (Li and Durbin, 2010) with parameters "-l 1024 -n.01." We removed PCR duplicates using DeDup v0.12.3 (Peltzer et al., 2016). We used the trimBam function in bamUtils v1.0.13 (Jun et al., 2015)[2] to trim the first and last two base pairs (bp) of each read to remove deamination-based 5′ C > T and 3′ G > A misincorporations and minimize the impact of postmortem DNA damage on genotyping. For the SNPs in the "1240k" panel (Mathieson et al., 2015), we randomly sampled a single high-quality base from a high-quality base (Phred-scaled base quality score 30 or higher) as a pseudo-diploid genotype, using the pileupCaller program.[3]

Ancient deoxyribonucleic acid authentication

The quality of ancient genomic material was assessed through a suite of methods. First, we tabulated patterns of post-mortem chemical modifications expected for ancient DNA using mapDamage v2.0.6 (Jónsson et al., 2013). Second, we estimated mitochondrial contamination rates from modern humans for all individuals using Schmutzi v1.5.1 (Renaud et al., 2015). Third, we measured the nuclear genome contamination rate in males based on X chromosome data as implemented in ANGSD v0.910 (Korneliussen et al., 2014). Since males have only a single copy of the X chromosome, mismatches between bases, aligned to the same polymorphic position, and beyond the level of sequencing error are considered evidence of contamination.

Genetic sexing and uniparental haplotype assigning

We assigned the biological sex of the ancient samples with the aid of the programs Rx (Mittnik et al., 2016) and Ry (Skoglund et al., 2013).

For mtDNA, we employed the log2fasta program built-in Schmutzi (Renaud et al., 2015) to call mitochondrial consensus sequences from the Schmutzi output. Variations that appeared when checked against rCRS were re-checked in BAM (Binary Alignment Map) files through visual inspection with IGV software (Helga et al., 2013). We then used Haplogrep 2 (Weissensteiner et al., 2016) to assign haplogroups. Y chromosome haplogroups were examined by aligning a set of positions in the ISOGG (International Society of Genetic Genealogy[4] and Y-full[5] databases, and analysis was performed in the case of a base and mapping quality exceeding 30. Haplogroup determination was performed with the script Yleaf.py in Yleaf software (Ralf et al., 2018), which provides outputs for allele counts of ancestral and derived SNPs along a path of branches of the Y-chromosome tree. Finally, we re-checked SNPs by visual inspection with IGV software (Helga et al., 2013).

2 https://github.com/statgen/bamUtil

3 https://github.com/stschiff/sequenceTools

4 http://isogg.org

5 https://www.yfull.com/tree/

Estimating genetic relatedness

We utilized READ (Kuhn et al., 2018) software to detect the degree of genetic kinship between ancient individuals.

Data merging

We merged our data with two sets of previously published worldwide populations' genotype datasets using the *mergeit* program from EIGENSOFT (Patterson et al., 2006), one is based on the Affymetrix Human Origins array dataset containing 597,573 SNPs ("HO" dataset) (Mallick et al., 2016; Jeong et al., 2019), while the other is based on 1240K capture dataset (1,233,013 SNPs, including all the ancient samples and shotgun-sequenced modern samples) (Haak et al., 2015; Damgaard et al., 2018a,b; Narasimhan et al., 2019; Ning et al., 2019, 2020; Wang et al., 2019, Wang et al., 2021; Jeong et al., 2020; Yang et al., 2020; Mao et al., 2021; Zhang et al., 2021).

Population structure analysis

We performed principal component analysis (PCA) as implemented in the *smartpca* v16000 (Patterson et al., 2006) on the HO dataset with options "lsqproject: YES" and "shrinkmode: YES" and projected ancient individuals onto the calculated components. We also conducted an unsupervised ADMIXTURE analysis with ADMIXTURE v1.3.0 (Alexander et al., 2009) after pruning for linkage disequilibrium by PLINK v1.90 (Chang et al., 2015) using parameters "– indep-pairwise 200 25 0.4" and generated a total of 271,578 SNPs for ADMIXTURE analysis. We separately labeled individuals with significant differences in clusters and ancestries.

f-statistics

We used outgroup-f_3 statistics (Patterson et al., 2012) to measure the genetic relationship between two populations following their divergence from the outgroup. We calculated f_4 statistics built-in ADMIXTOOLS (Patterson et al., 2012) with the "*f4mode*: YES" parameter to further reveal the genetic differences between the two studied subgroups, and their genetic relationships with other ancient and modern eastern Eurasian populations. All f statistics were based on a 1240k dataset.

Pairwise-*qpWave* homogeneous analysis

We further utilized the pairwise *qpWave* (Patterson et al., 2012; Agranat-Tamir et al., 2020) to test whether pairs of populations form genetic clade related to a set of outgroups.

Admixture modeling with *qpAdm*

We applied the *qpAdm* program (Patterson et al., 2012) based on the 1240k dataset to model ancestry proportions of our ancient Shichengzi population related to one or two different ancestral sources from a set of reference populations, along with parameters "allsnp: YES" and "details: YES." We also used a model competition approach to find the best model. We took pairs of fitted models, added the source population from one to the reference population set of the other, and assessed whether it continued to fit (Sirak et al., 2021).

Results

C-14 dating and C, N isotope analysis

All AMS radiocarbon dates for archeological materials recovered from Shichengzi are presented in **Supplementary Table 1**. Dates ranged from 40 cal BC to 420 cal AD (2σ). Pending more evidence from unearthed artifacts and Chinese historical literature, we believe that Shichengzi was occupied from the first century BC to the third century AD.

As plotted in **Figure 2**, eight sampled human individuals could be divided into two groups based on differences in δ^{15}N values. The δ^{15}N values of these human samples ranged from 9.6 to 14.3%, suggesting the consumption of different volumes of proteins. The mean δ^{15}N value of individuals in Group A (M1, M2, M4, M5, M6, and M8) was 13.8 ± 0.4%, demonstrating consumption of considerable ^{15}N-enriched foods, such as meat and milk products. In contrast, the two individuals in Group B (M3 and M9) exhibited relatively low δ^{15}N values (9.6% and 10.3%), revealing a consumption of less protein-rich foods than Group A. The δ^{13}C values from the osteological material ranged from -16 to -18.2%, suggesting a similar dietary pattern among individuals partaking in diets harboring C_3-based proteins. All raw data are laid out in detail in **Supplementary Table 4**.

As shown in **Figure 2**, most negative δ^{13}C and δ^{15}N values were derived from charred wheat and naked barley seed samples. These ranged from -25.8 to -19.8% (mean ± *SD*: -23.4 ± 1.6%) and 3.1–11.8% (mean ± *SD*: 6.9 ± 2.3%). Newly acquired stable carbon and nitrogen isotopic results for foxtail millet and common millet stood at -11.9, -11.6%, and 9.6, 10.7%, respectively. This crop data established an isotope baseline for C_3 and C_4 crop foods at Shichengzi. The δ^{13}C and δ^{15}N values of a horse (*Equus caballus*) bone recovered from the site emerged at -20.2 and 5.4%, respectively. Since this N isotope value was lower than most wheat and naked barley grains, it suggested that the horse consumed a diet heavily influenced by wild C_3 terrestrial grasses and/or shrubs. In addition, the mean δ^{13}C and δ^{15}N values of sheep/goat (n = 20) and cattle (n = 4) stood at -18.4 ± 0.5, -18.9 ± 0.8% and 8 ± 1.1, 8.3 ± 1.6%, revealing that

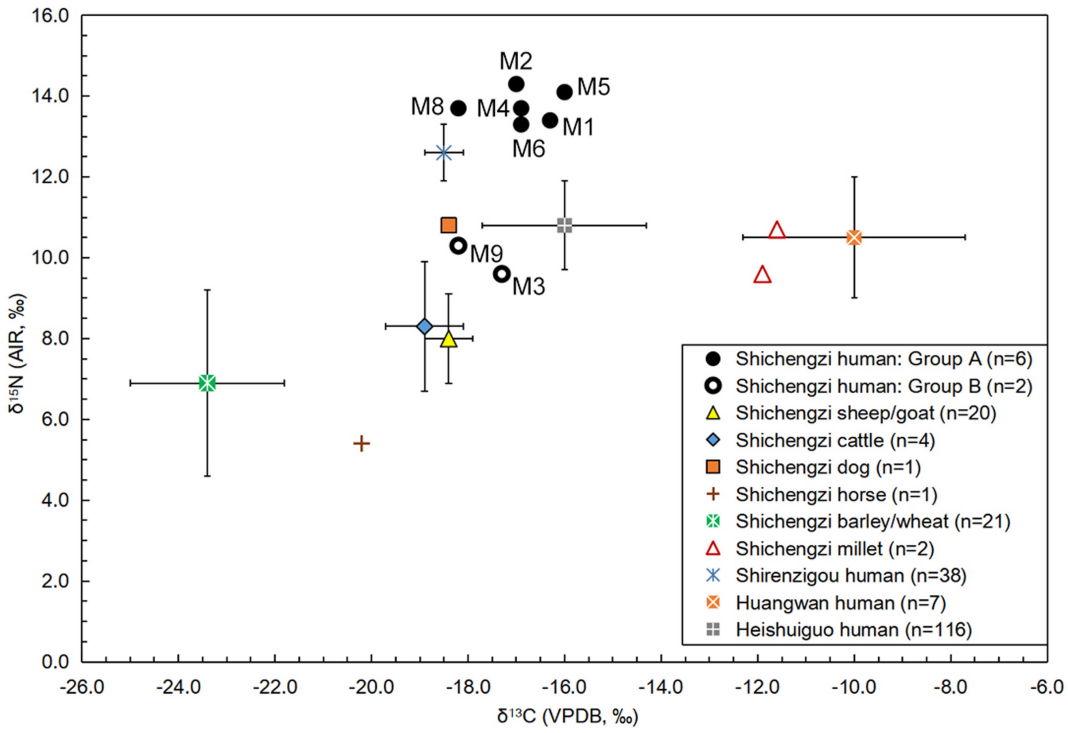


FIGURE 2
Stable C and N isotopic data of the crop, animal, and human discovered from the Shichengzi and other three relevant sites.

domestic herbivore diets were influenced by C_3-based proteins primarily based on *Triticoid* cereals and wild plants.

Previous isotopic studies have demonstrated that ^{15}N in human collagen is enriched by + 3∼5% relative to diet (Hedges and Reynard, 2007). This provided a baseline for further interpretation of the Shichengzi diet upon the above-mentioned C, N isotopic baseline for potential food resources around the site. As shown in **Figure 2**, the mean $\delta^{15}N$ value of these humans in Group A (13.8%) was 5.8, 5.5, and 6.9 (>3∼5%) higher than values for sheep/goat and cattle bones and crop grains of wheat and naked barley, respectively, suggesting consumption of high degrees of meat and milk products derived from domesticated herbivores found in the site vicinity. In contrast, the mean $\delta^{15}N$ value of the Group B population (10%) was 2, 1.7, and 3.1 (<3∼5% or around 3%) higher than the mean ^{15}N values of sheep/goat and cattle bones, as well as wheat and naked barley seeds, respectively. Moreover, both individuals in Group B exhibited a lower $\delta^{15}N$ value than a canine sample (10.8%) recovered from Shichengzi, suggesting that this group regularly dined on barley and wheat foods, with minimal consumption of meat from domestic sheep/goat and cattle. If so, it is reasonable to believe that Group A individuals represent pastoral inhabitants, and Group B populations were likely agriculturalists.

Comparative $\delta^{13}C$ and $\delta^{15}N$ data collected from three sites near Shichengzi have been summarized in **Supplementary Table 3** and plotted in **Figure 2**. The mean $\delta^{13}C$ and $\delta^{15}N$ values of agro-pastoralists found at Shirenzigou (**Figure 1a**) were −18.5 ± 0.4 and 12.6 ± 0.7%, respectively, showing that Shirenzigou diets resembled individuals from Group A at Shichengzi and were significantly influenced by ^{15}N enriched C_3-based proteins. On the contrary, mean $\delta^{13}C$ and $\delta^{15}N$ values of Han populations at the Huangwan site and Heishuiguo site in the Hexi Corridor (see **Figure 1a**) were -10 ± 2.3, 10.5 ± 1.5% and -16 ± 1.7, 10.8 ± 1.1%, respectively. Although individuals at Huangwan yielded a far enriched $\delta^{13}C$ value, both Han populations were akin to the Shichengzi B population exhibiting nearly identical and markedly low-level of $\delta^{15}N$ levels, indicating that their consumption of animal foods was rather limited and pointing to a basis of inter-group dietary similarity. Thus, despite variability within our assigned categories, when considered within overall food webs, alongside corresponding animal diets, and within the broader northwest China Han context, we feel confident in dividing the Shichengzi population into an agricultural and an agro-pastoral group.

Ancient genome data production

We initially screened four ancient individuals by shallow shotgun sequencing of one Illumina sequencing library per

individual, resulting in coverage from 0.023 to 0.064X (**Supplementary Table 5**). We verified the authenticity of the genome data through a series of methods. All samples showed typical characteristics of ancient DNA with postmodern chemical damage (**Supplementary Figure 2**). All individuals exhibited negligible contamination from modern populations and were, thus, suitable for subsequent analysis, although we note that M4 showed low mitochondrial coverage that would prove insufficient for estimating the mitochondrial contamination rate. Our kinship analysis confirmed that all pairs of individuals were unrelated. We retained all individuals for subsequent analysis. Mitochondrial haplogroups for all Shichengzi individuals belonged to D4 clades, including D4j15, D4c2b, D4j7, and D4c2c. The two males (M3 and M9) were assigned to Y chromosomal haplogroup O (i.e., O1b*-F435 and O2a2b2a1a-F4110), a typical East Asian haplogroup. Due to low coverage, the other male, M4, was only tentatively assigned to haplogroup N1a1a1a1a-CTS1077, which is prevailing in Northeast Asia, implying the dual paternal origins of the Shichengzi population (**Supplementary Table 5**).

Overall genomic structure

We observed that the four Shichengzi samples were separated into two groups in the PCA plot (**Figure 3**). The pair of individuals M1 and M4 clustered in the upper section, together with ancient Neolithic to Iron Age eastern Mongolia Plateau (including Mongolia_N, Ulaanzuukh_SlabGrave, lateXiongnu) and Lake Baikal Eneolithic Hunter-Gatherer (HG) (including Russia_Lokomotiv and Russia_Shamanka populations). The pair of individuals M3 and M9 clustered with ancient agricultural populations from the Yellow River Basin, as well as modern Han Chinese. Our ADMIXTURE analysis revealed a similar pattern (K = 4; **Figure 4**), with the four Shichengzi individuals divided into two groups with different ancestry compositions. The samples M1 and M4 mainly harbored Ancient Northeast Asian (ANA) related ancestry, which was enriched in ANA HG (represented by AR_EN), while the samples M3 and M9 exhibited a genetic profile akin to millet farmers in the Yellow River Basin from the Late Neolithic Age onward (included YR_LN, YR_LBIA) (Ning et al., 2020).

Dual origins of Shichengzi residents

Next, we utilized outgroup-f_3 analysis to further determine the genetic differences among Shichengzi individuals and explore the genetic affinity of Shichengzi to modern/ancient Eurasians on a quantitative basis. Here the high genetic heterogeneity among Shichengzi groups was vividly exemplified: M3 and M9 shared more alleles with Yellow River millet farmers (YR_LN and YR_LBIA) and modern Han groups, while M1 and M4 exhibited a close genetic relationship with Northeast Asian populations who harbored a larger amount of ANA ancestry, including ANA (Neolithic Mongolia_N_East/Russia_Lokomotiv_Eneolithic/DevilsCave_N/Russia_Shamanka_Eneolithic, Bronze Age's Ulaanzuukh_SlabGrave/WLR_BA_o) (Jeong et al., 2020; Ning et al., 2020; Wang et al., 2021) and modern Tungusic-speaking groups (**Figure 5** and **Supplementary Table 6**). We confirmed the above finding by f_4 statistics (**Supplementary Figure 3**). Pairwise *qpWave* analysis revealed a similar clustering pattern with outgroup-f_3 statistics but further demonstrated the genetic difference among Shichengzi individuals (**Figure 6**). M3 and M9 were genetically homogeneous with Yellow River millet farmers ($p > 0.05$), while M1 clustered not only with Yellow River millet farmers but also with HG from Mongolia Plateau and Baikal Lake ($p > 0.05$). Notably, M4 only clustered with HG from Mongolia Plateau and Baikal Lake ($p > 0.05$). Based on those results from PCA, ADMIXTURE, f-statistics, and pairwise-*qpWave* homogeneous test, we grouped the four individuals into Shichengzi_A (M1, M4) and Shichengzi_B (M3, M9). Shichengzi_A showed a closer genetic relationship with Neolithic HG in Mongolia Plateau and Baikal Lake and Bronze Age Ulaanzuukh_SlabGrave populations, who derived the majority of the ancestry from ANA, as opposed to Yellow River millet farmers [positive values of f_4 (Mbuti, Shichengzi_A; Yellow River, Mongolia Plateau)]. Shichengzi_A also shared more alleles with modern Tungusic and Mongolic speaking populations than Sino-Tibetan speakers as reflected in positive values of f_4 (Mbuti, Shichengzi_A; Sino-Tibetan, Tungusic/Mongolic) (**Supplementary Figures 4**, **5**). By contrast, Shichengzi_B showed a closer genetic affinity with Yellow River millet farmers than with ancient populations of Mongolia Plateau, as reflected in negative values of f_4 (Mbuti, Shichengzi_B; Yellow River, Mongolia Plateau). Shichengzi_B also shared more alleles with Sino-Tibetan populations than with Altaic speaking groups, as shown in the negative values of f_4 (Mbuti, Shichengzi_B; Sino-Tibetan, Altaic) (**Supplementary Figures 4**, **5**).

We further used *qpWave* and *qpAdm* to explore the number of ancestry sources and plausible admixture models for the Shichengzi groups (**Supplementary Table 7**). A minimum of two ancestral streams could indicate the likely origins of both Shichengzi groups. Modeling for Shichengzi_A showed a likely descent from ancient Mongolia Plateau populations, harboring mostly ANA ancestry; while one-way modeling with Ulaanzuukh_SlabGrave as a single source failed when Mongolia_N_North was included in the outgroup set (p = 0.0063), but a fitted two-way model was acquired only with the inclusion of Ami as the other source at an estimated proportion of 25.8 ± 7.9%. The models with YR_LN as a source failed to explain the genetic

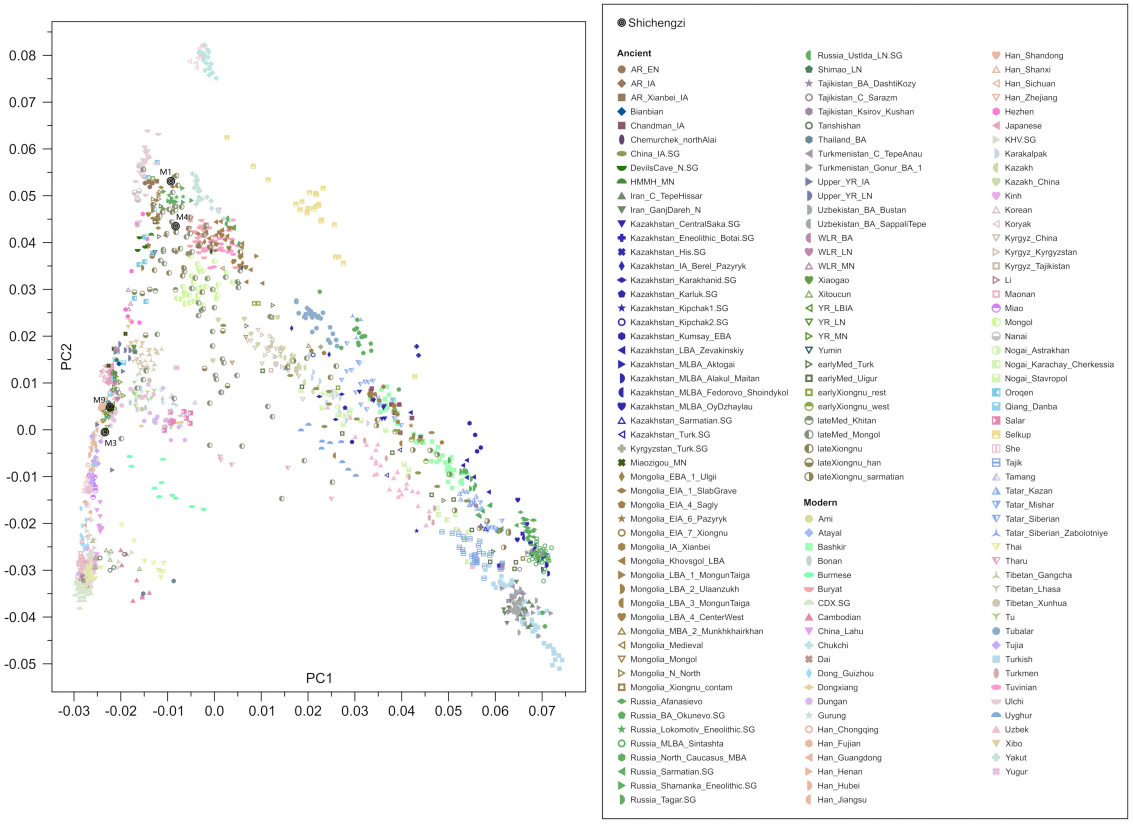


FIGURE 3
PCA of ancient samples. Projection of ancient samples onto PCA dimensions 1 and 2 defined by East Asian, Central Asian, European, Siberian, and other populations.

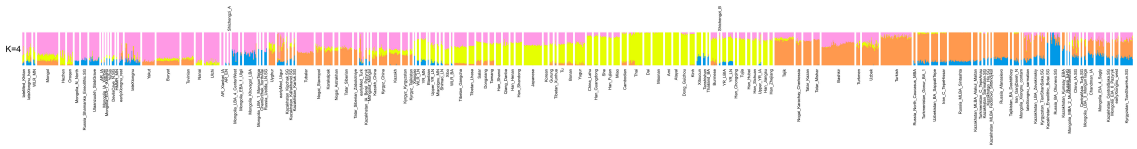


FIGURE 4
The result of ADMIXTURE analysis (K = 4) among Eurasian groups shows that two Shichengzi groups harbor disparate ancestral composition patterns. Populations mainly from southern China and Southeast Asia (yellow), Mongolia (light magenta), and West Eurasia (Orange and blue).

variation of Shichengzi_A, suggesting a very limited genetic influence from farming groups. Shichengzi_B exhibited a thoroughly different profile in its ancestry composition by deriving ancestry from YR_LN rather than ANA-related populations. When we added Mongolia_North to the outgroup set, the models involving YR_LN as a single source still provided a fit for Shichengzi_B ($0.37 < p < 0.81$). The admixture modeling results at the individual level of M3 and M9, M1, and M4 were consistent with the Shichengzi_B and Shichengzi_A results at the group level, respectively.

On Xiongnu and Wusun connections

In Han times, the Xiongnu, a pastoralist regime to the dynasty's north, flourished on the Mongolia Plateau. As stated above, a possible connection between Shichengzi inhabitants and Xiongnu is worth further exploration. In our study, the Xiongnu population was genetically classified into eastern and western groups, with the eastern group (earlyXiongnu_rest, late Xiongnu, and lateXiongnu_Han) exhibiting dominant Northeast Asian or Han-related ancestry (Jeong et al., 2020). We found that sample M4 also formed

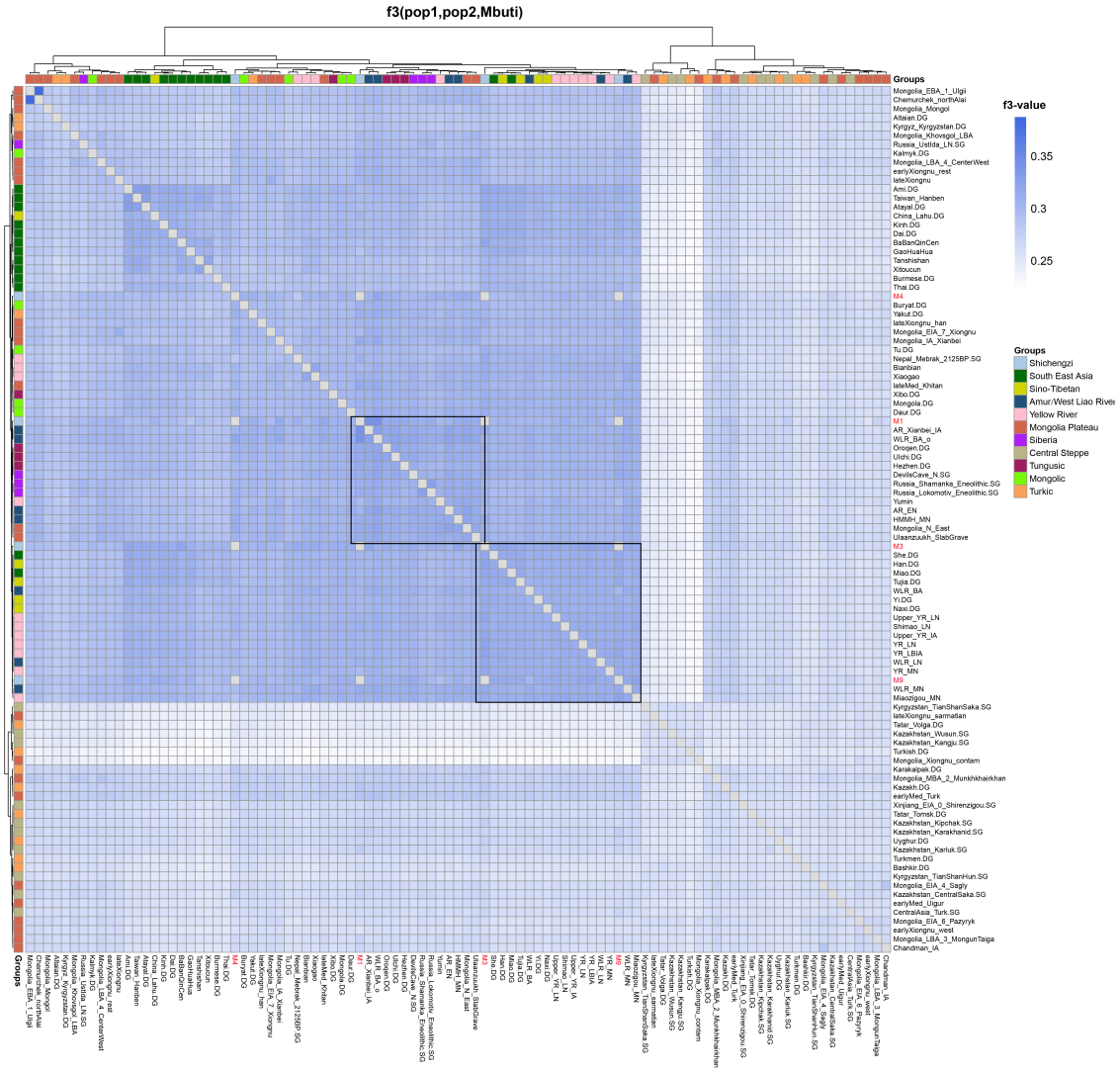


FIGURE 5
The results of f_3(X1, X2; Mbuti) based on the 1240k dataset show the genetic relationship among East Eurasians. Darker colors indicated a larger-f3 value and more shared alleles between a pair of tests.

a genetic clade with early Xiongnu rest as reflected in non-significant values of f_4 (X, Mbuti; M4, earlyXiongnu_rest) (**Supplementary Figure 3**). The pairwise-*qpWave* analysis also demonstrated that M1 and M4 formed a genetic clade with lateXiongnu_Han, and M4 also exhibited certain similarities with late Xiongnu (**Figure 6**).

During and after the reign of Emperor *Wudi*, the Han grand strategy in Western Regions was described historically as "associating with distant countries and attacking those nearby" (*yuan jiao jin gong*): allying with Western Region states to better take the fight to the feared Xiongnu. The powerful Wusun (乌孙) state was one such contemporary polity subsequently courted by the Han. With this historical information in mind, we also explored the relationship between Shichengzi and Wusun groups. Our f_4 analysis in the form of f_4 (X, Mbuti; Shichengzi, Wusun) showed a genetically significant difference between Shichengzi and Wusun/Kangju populations in Central Asia, who derived most of their ancestry from West Eurasian groups (Z < -3 when X included West Eurasians, such as Sintashta, Turkmenistan_Gonur_BA_1, and Anatolia_N) (**Supplementary Table 8**).

We also explored the genetic relationship between Shichengzi and samples from the Shirenzigou site, an early Iron Age agro-pastoralist site on the northern slopes of the eastern Tianshan Mountains, Xinjiang (Ning et al., 2019). We conducted an f_4 (X, Mbuti; Shichengzi, Shirenzigou) and observed significant differences between both Shichengzi

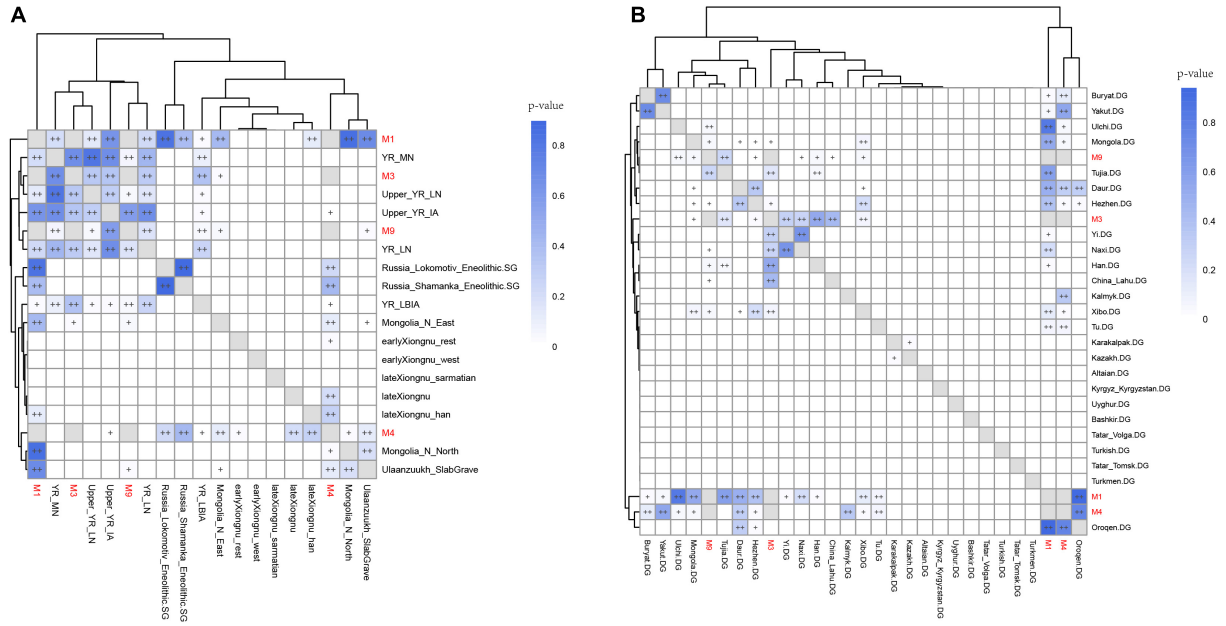


FIGURE 6
The results of qpWave analysis show the different genetic affinity patterns of four Shichengzi individuals to ancient populations in the Yellow River valley **(A)** and Mongolia Plateau **(B)**. "++" represented values greater than 0.05, and "+" represented values lower than 0.05 but larger than 0.01. $p > 0.05$ indicated the supported genetic clade between pairs. Computations were based on the outgroup set (Mbuti.DG, Onge.DG, Yana_UP.SG, Loschbour.DG, Iran_GanjDareh_N, Turkmenistan_Gonur_BA_1, Russia_Afanasievo, Anatolia_N, Kazakhstan_Eneolithic_Botai.SG, AR19K, Liangdao2, Japan_Jomon, Miaozigou_MN).

groups and Shirenzigou people: both Shichengzi groups harbored more East Eurasian ancestry than Shirenzigou ($Z > 3$ when X included East Eurasians) (**Supplementary Figure 6**). This refuted the possibility of direct migration between the two locales.

Discussion

This study has presented a combination of evidence: the paleogenetics of the Shichengzi population; C and N isotope analysis of the Shichengzi paleodiet; and Shichengzi's excavated burials. Shichengzi inhabitants are divided into two separate genetic groups and paleodietary patterns. This combination of DNA and stable isotope analysis has used methods unavailable to traditional archeology and historiography and conclusively demonstrated the high likelihood of joint agropastoralist/agriculturalist use of the Shichengzi site during the Han (202 BC–AD 8 and AD 25—220) dynasty. Use of site space and burial data (**Supplementary Figure 1** and **Figures 1b–e**), however, points to close interaction between these different groups alongside the expression of differences. Subsequently, we argue that these Shichengzi groups were contemporaries, each group involved in who were both involved in the social and political life of this Han frontier post. Combining our new C and N isotopic evidence of human dietary patterns from the Shichengzi site with isotopic data from three other related sites, Shirenzigou, Huangwan, and Heishuiguo (**Figure 1a**), strengthens the claim that these pastoralist and agricultural subsistence strategies were maintained within a shared sociopolitical environment around Shichengzi.

The Shichengzi evidence is useful for updating our understanding of lifeways and lifestyles within the Han border system. With a widened array of archeological data, we may also begin to place the archeology of Han borders in a global context. The archeology of Roman border formation processes, for example, has previously addressed the issue of agropastoral/agricultural interaction. Leaning on arguments that stressed the entanglement of plant and animal domestication processes within specific ecosystems – a relationship of "mutualism" as opposed to the artificial and rigid division of agricultural and pastoralist pathways (Langlie and Capriles, 2021; Boucher, 1985), Banning argued for an underlining mutualism in agropastoralist/agriculturalist interactions on the Roman frontier (Banning, 1986). Parker (1987) built on this argument with a critique pointing out that only a "policed" frontier could realistically contain any kind of mutualistic relationship. While Shichengzi undeniably witnessed a degree of "integration" and "accommodation" of Han and non-Han subjects (Di Cosmo, 2009) under the banner

of Imperial administrative control and "policing," we find that such "mutualism" also expressed a localized knowledge and adaptation to new sociopolitical realities. Archeologically speaking, the Han system was adapted by these farming and pastoral peoples. In terms of personal livelihoods and wealth, the impressive grave goods of the likely agro-pastoralist individual M2 may suggest that local elites bought into and indeed perhaps embraced the Han system within their own broader networks. By contrast, the bare burials of Han migrants, such as individual M3, suggesting that they occupied a considerably lowlier status in life, in line with literary evidence suggesting much of the Han population moved to these regions occupied a servile or imprisoned status (Yang, 1991). We can look forward to advances in work on urbanism in the Xinjiang region during and around this time period, work which will highlight the interconnections and regional temporal precedents of local and Han imperial formation processes.

Our DNA and paleodietary analysis has also improved our sense of lifeways and subsistence at Shichengzi and thus other Han frontier garrisons. Recent bioarcheological studies of the Roman frontier, inspired by the social archeological literature of the early 2000s (Meskell and Preucel, 2004), have begun to consider their subjects as "embodied" individuals within this complex liminal space (Gowland, 2017). At Shichengzi, diversity of subsistence pathways and genetic background coalesced into an "embodied" border identity of mutually benefiting pastoralist and agricultural groups. Agro-pastoral groups may have attached themselves to Shichengzi for economic and social benefit and become embodied parts of border reality through the long-term consequences of such decisions. In turn, we can also begin to suggest the possibility that this rapid acclimatization to the agricultural garrison system was equally facilitated by agro-pastoralist adaptations, with engagement running both ways between agricultural groups. This rapid adaptation to Han space further suggests longer-term, synergistic agro-pastoralist/agricultural relationships in the Xinjiang region. The Han system, in a sense, bought and adapted itself to this network.

The divergence in the genetic and paleodietary profile of burials at Shichengzi suggests that Han and Altaic-speaking agriculturalist/agro-pastoralist populations inhabited and used the site extensively. As the agricultural garrison system and evidence in the form of Han-style pottery kilns both suggest, this was a quintessentially imported Han space. Yet, while Han/non-Han burials may be distinguishable by the presence of *wuzhu* coins in the latter case and the vertical-shaft earth pit burials of agro-pastoralist populations across Xinjiang in this period for the latter (Tian, 2021), the observed customary use of this shared burial chamber and long-term association of both groups with Shichengzi should be considered further as evidence of more long-term, habitual accommodation. Our evidence points to the long-term adaptation of both agriculturalists and non-agriculturalist to the agricultural garrison system, in a process more deeply etched in the region and more flexible than accommodation might suggest. This offers a promising line of analysis as new archeological material for the Western Regions of the Han Empire is unearthed and curated using the methods of scientific archeology.

Conclusion

Our study offers the first ancient DNA data providing coverage of the Han dynasty (202 BC–220 AD) border regions. This is combined with isotopic and burial data and allows us to put forward a new reading of Han border formation and development. Border identity can be understood as an array of plural identities, combining agricultural and agro-pastoral lifeways in highly localized contexts, adapting to new sociopolitical realities, and finding means of mutual benefit and group individuation. Future projects will attempt to extend the study of Han border history through analysis of multiple site types, delving further into short- and long-term border complexity in northwest China, using the tools highlighted in this essay alongside additional traditional and scientific archeological analysis.

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 permission and oversight were provided by the Ethics Committee of Fudan University of Life Sciences to study their ancient genomes. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

PS, C-CW, SW, and EA conceived the idea for the study. EA, DC, XT, YW, XQ, and PS performed or supervised the archeological work. YX, PD, JX, and SW performed or supervised the wet laboratory work. YY, XY, and C-CW analyzed the data. EA, YY, XY, PS, SW, and C-CW wrote and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by the National Key R&D Program of China (2020YFE0201600 and 2020YFC1521607), the National Social Science Fund of China (19VJX074 and 21CKG022), the National Natural Science Foundation of China (32070576, 31801040, and 32111530227), the B&R Joint Laboratory of Eurasian Anthropology (18490750300), the Major Research Program of National Natural Science Foundation of China (91731303), the Major Project of National Social Science Foundation of China granted to C-CW (21&ZD285), Xiaohua Deng (20&ZD248), and SW (20&ZD212), 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), the Nanqiang Outstanding Young Talents Program of Xiamen University (X2123302), the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), the 111 Project (B13016), the Major Project of Marxist Theoretical Research and Construction Project (2021MZD014), and European Research Council (ERC) grant to Dan Xu (ERC-2019-ADG-883700-TRAM).

Acknowledgments

We are thankful to Prof. Yi Guo at Zhejiang University for his help with the pre-treatment experiment of bone samples.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.932004/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Location of the Shichengzi site and the photograph of the four excavation areas (A–D) at Shichengzi during 2014–2019.

SUPPLEMENTARY FIGURE 2
Shared genetic drift estimated via f4-statistics of the form **(A)** f4(Mbuti, Shichengzi_A; Sino-Tibetan, Altaic) and **(B)** f4(Mbuti, Shichengzi_B; Sino-Tibetan, Altaic) based on the merged 1240K dataset.

SUPPLEMENTARY FIGURE 3
Shared genetic drift estimated via f4-statistics of the form (A) f4(Mbuti, Shichengzi_A; Yellow River, Mongolia Plateau) and (B) f4(Mbuti, Shichengzi_B; Yellow River, Mongolia Plateau) based on the merged 1240K dataset.

SUPPLEMENTARY FIGURE 4
Results of f4-statistics performed in the form of f4(X, Mbuti; Shichengzi, Shirezigou) to explore the genetic similarities and differentiation between Shichengzi populations and possible ancestral source populations.

SUPPLEMENTARY FIGURE 5
Results of f4-statistics performed in the form of f4(X, Mbuti; Shichengzi_ind, EA) to explore the genetic similarities and differentiation between Shichengzi populations and possible ancestral source populations.

SUPPLEMENTARY FIGURE 6
(A) Mapdamage result shows mismatch frequency in different distance from 5′ and 3′ end of sequence read, four colors representing four individuals. **(B)** Kinship detected between pairs of individuals.

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, 1146–1157.e11. 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, 1655–1664. doi: 10.1101/gr.094052.109

Alkass, K., Saitoh, H., Buchholz, B. A., Bernard, S., Holmlund, G., Senn, D. R., et al. (2013). Analysis of radiocarbon, stable isotopes and DNA in teeth to facilitate identification of unknown decedents. *PLoS One* 8:e69597. doi: 10.1371/journal.pone.0069597

Ames, K. M., Richards, M. P., Speller, C. F., Yang, D. Y., Lyman, R. L., and Butler, V. L. (2015). Stable isotope and ancient DNA analysis of dog remains from Cathlapotle (45CL1), a contact-era site on the Lower Columbia River. *J. Archaeol. Sci.* 57, 268–282. doi: 10.1016/j.jas.2015.02.038

Ban, G. (1962). *Han Shu*. Beijing: China Publishing House.

Banning, E. B. (1986). Peasants, pastoralists, and pax romana: mutualism in the highlands of Jordan. *Bull. Am. Sch. Orient. Res.* 261, 25–50.

Boucher, D. H. (1985). "The idea of mutualism, past, and future," in *The biology of mutualism: ecology and evolution*, ed. D. H. Boucher (Oxford: Oxford University Press), 1–28.

Blank, M., Sjögren, K. G., Knipper, C., Frei, K. M., Malmström, H., Fraser, M., et al. (2021). Mobility patterns in inland southwestern Sweden during the Neolithic and Early Bronze Age. Archaeol. *Anthropol. Sci.* 13:64. doi: 10.1007/s12520-021-01294-4

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. doi: 10.1186/s13742-015-0047-8

Damgaard, P., de, B., Martiniano, R., Kamm, J., Moreno-Mayar, J. V., Kroonen, G., et al. (2018b). The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* 360:eaar7711. doi: 10.1126/science.aar7711

Damgaard, P., de, B., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., et al. (2018a). 137 ancient human genomes from across the Eurasian steppes. *Nature* 557, 369–374. doi: 10.1038/s41586-018-0094-2

Di Cosmo, N. (2009). Han frontiers: toward an integrated view. *J. Am. Orient. Soc.* 129, 199–214.

Fan, Y. (2000). *Hou Han Shu*. Beijing: China Publishing House.

Frei, K., Mannering, U., Kristiansen, K., Allentoft, M. E., Wilson, A. S., Skals, I., et al. (2015). Tracing the dynamic life story of a Bronze Age Female. *Sci. Rep.* 5:10431. doi: 10.1038/srep10431

Frei, K. M., Bergerbrant, S., Sjögren, K. G., Jørkov, M. L., Lynnerup, N., Harvig, L., et al. (2019). Mapping human mobility during the third and second millennia BC in present-day Denmark. *PLoS One* 14:e0219850. doi: 10.1371/journal.pone.0219850

Gowland, R. L. (2017). Embodied identities in Roman Britain: a bioarchaeological approach. *Britannia* 48, 175–194.

Haak, W., Brandt, G., Jong, H. N. D., Meyer, C., Ganslmeier, R., Heyd, V., et al. (2008). Ancient DNA, Strontium isotopes, and osteological analyses shed light on social and kinship organization of the Later Stone Age. *Proc. Natl. Acad. Sci. U.S.A.* 105, 18226–18231. doi: 10.1073/pnas.0807592105

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. doi: 10.1038/nature14317

Hedges, R., and Reynard, L. (2007). Nitrogen isotopes and the trophic level of humans in archaeology. *J. Archaeol. Sci.* 34, 1240–1251.

Helga, T., James, T. R., and Jill, P. M. (2013). Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinf.* 14, 178–192. doi: 10.1093/bib/bbs017

Ingman, T., Eisenmann, S., Skourtanioti, E., Akar, M., Ilgner, J., Gnecchi Ruscone, G. A., et al. (2021). Human mobility at Tell Atchana (Alalakh), Hatay, Turkey during the 2nd millennium BC: Integration of isotopic and genomic evidence. *PLoS One* 16:e0241883. doi: 10.1371/journal.pone.0241883

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. 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, 890–904.e29. doi: 10.1016/j.cell.2020.10.015

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29, 1682–1684. doi: 10.1093/bioinformatics/btt193

Jun, G., Wing, M. K., Abecasis, G. R., and Kang, H. M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res.* 25, 918–925. doi: 10.1101/gr.176552.114

Knapp, M., Clarke, A. C., Horsburgh, K. A., and Matisoo-Smith, E. A. (2012). Setting the stage – Building and working in an ancient DNA laboratory. *Ann. Anat.* 194, 3–6. doi: 10.1016/j.aanat.2011.03.008

Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* 15:356. doi: 10.1186/s12859-014-0356-4

Kuhn, J. M. M., Jakobsson, M., and Günther, T. (2018). Estimating genetic kin relationships in prehistoric populations. *PLoS One* 13:e0195491. doi: 10.1371/journal.pone.0195491

Langlie, B. S., and Capriles, J. M. (2021). Paleoethnobotanical evidence points to agricultural mutualism among early camelid pastoralists of the Andean central Altiplano. *Archaeol. Anthropol. Sci.* 13:107. doi: 10.1007/s12520-021-01343-y

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

Li, N. (2017). The organization and management of the *tuntian* system in the Western Regions during Han dynasties. *Agricult. Archaeol.* 1, 124–132.

Luo, L., Wang, X., Lasaponara, R., Xiang, B., Zhen, J., Zhu, L., et al. (2018). Auto-Extraction of linear archaeological traces of tuntian irrigation canals in miran site (China) from Gaofen-1 Satellite Imagery. *Remote Sens.* 10:718.

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.e13. doi: 10.1016/j.cell.2021.04.040

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

Meskell, L., and Preucel, R. W. (2004). *Identities. a companion to social archaeology*. Oxford: Oxford University Press.

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 2010:pdb.prot5448. doi: 10.1101/pdb.prot5448

Millward, J. A. (2013). *The silk road: a very short introduction*. Oxford: Oxford University Press.

Mittnik, A., Massy, K., Knipper, C., Wittenborn, F., and Promerová, M. B. (2019). Kinship-based social inequality in Bronze Age Europe. *Science* 366, 731–734. doi: 10.1126/science.aax6219

Mittnik, A., Wang, C.-C., Svoboda, J., and Krause, J. (2016). A molecular approach to the sexing of the triple burial at the upper paleolithic site of dolní vistonice. *PLoS One* 11:e0163019. doi: 10.1371/journal.pone.0163019

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. 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.* 11:2700. 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.e4. doi: 10.1016/j.cub.2019.06.044

Parker, S. T. (1987). Peasants, Pastoralists, and "Pax Romana": A Different View. *Bull. Am. Sch. Orient. Res.* 265, 35–51.

Patterson, N., Isakov, M., Booth, T., Büster, L., Fischer, C. E., Olalde, I., et al. (2022). Large-scale migration into Britain during the Middle to Late Bronze Age. *Nature* 601, 588–594. doi: 10.1038/s41586-021-04287-4

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 Genetics* 2:e190. doi: 10.1371/journal.pgen.0020190

Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., et al. (2016). EAGER: efficient ancient genome reconstruction. *Genome Biol.* 17:60. doi: 10.1186/s13059-016-0918-z

Ralf, A., Montiel González, D., Zhong, K., and Kayser, M. (2018). Yleaf: software for human Y-Chromosomal haplogroup inference from next-generation sequencing data. *Mol. Biol. Evol.* 35, 1291–1294. doi: 10.1093/molbev/msy032

Reimer, P., Austin, W., Bard, E., Bayliss, A., Blackwell, P. G., Ramsey, C. B., et al. (2020). The IntCal20 Northern Hemisphere radiocarbon age calibration curve (0–55 cal kBP). *Radiocarbon* 62, 725–757.

Renaud, G., Slon, V., Duggan, A. T., and Kelso, J. (2015). Schmutzi: Estimation of Contamination and Endogenous Mitochondrial Consensus Calling for Ancient DNA. *Genome Biol.* 16:224. doi: 10.1186/s13059-015-0776-0

Richards, M. P., and Hedges, R. E. M. (1999). Stable isotope evidence for similarities in the types of marine foods used by late mesolithic humans at sites along the atlantic coast of Europe. *J. Archaeol. Sci.* 26, 717–722.

Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* 9:88. doi: 10.1186/s13104-016-1900-2

Sheng, P., Liu, Y., Tian, X., Wu, Y., and Guan, Y. (2021). Paleo-environmental implications of the micro-botanical remains recovered from a military garrison of Han Dynasty in Xinjiang. *J. Archaeol. Sci.* 39:103176.

Sheng, P., Storozum, M., Tian, X., and Wu, Y. (2020). Foodways on the Han dynasty's western frontier: archeobotanical and isotopic investigations at Shichengzi, Xinjiang, China. *Holocene* 30, 1174–1185.

Sirak, K. A., Fernandes, D. M., Lipson, M., Mallick, S., Mah, M., Olalde, I., et al. (2021). Social stratification without genetic differentiation at the site of Kulubnarti in Christian Period Nubia. *Nat. Commun.* 12:7283. doi: 10.1038/s41467-021-27356-8

Skoglund, P., Storå, J., Götherström, A., and Jakobsson, M. (2013). Accurate sex identification of ancient human remains using DNA shotgun sequencing. *J. Archaeol. Sci.* 40, 4477–4482. doi: 10.1016/j.jas.2013.07.004

Spengler, R. N. III (2019). *Fruit from the Sands: The Silk Road Origins of the Foods We Eat*. Berkeley: University of California Press.

Sun, X., Wen, S., Lu, C., Zhou, B., Curnoe, D., Lu, H., et al. (2021). Ancient DNA and multimethod dating confirm the late arrival of anatomically modern humans in southern China. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2019158118. doi: 10.1073/pnas.2019158118

Tian, X. (2021). Archaeological excavation and findings from the Shichengzi site in Qitai County, Xinjiang. *Cult, Relics World* 7, 71–74.

Tian, X., Feng, Z., and Wu, Y. (2020). The 2018 Excavation of the Shichengzi Site in Qitai County. Xinjiang. *Archaeology* 12, 21–40.

Tian, X., Wu, Y., Duo, S., Zhang, S., and Chen, X. (2018). The 2016 Excavation of the Shichengzi Site in Qitai County, Xinjiang. *Cult. Relics* 5, 4–25.

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*. 10:590. 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, 413–419. doi: 10.1038/s41586-021-03336-2

Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.-J., et al. (2016). HaploGrep 2: Mitochondrial Haplogroup Classification in the Era of High-Throughput Sequencing. *Nucleic Acids Res.* 44, W58–W63. doi: 10.1093/nar/gkw233

Wilson, A. S., Taylor, T., Ceruti, M. C., Chavez, J. A., Reinhard, J., Grimes, V., et al. (2007). Stable isotope and DNA evidence for ritual sequences in Inca child sacrifice. *Proc. Natl. Acad. Sci. U.S.A.* 104, 16456–16461. doi: 10.1073/pnas.0704276104

Yang, J. P. (1991). Several questions of the *tuntian* farming during the Western Han dynasty. *J. Chin. Soc. Econ. History* 4, 11–18.

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

Zhang, A., and Tian, H. (2015). The remains of the site the layout of the defense cities of the Han dynasty in the Western Regions. *J. Chin. Historic. Geogr.* 30, 47–55.

Zhang, F., Ning, C., Scott, A., Fu, Q., Bjørn, R., Li, W., et al. (2021). The genomic origins of the Bronze Age Tarim Basin mummies. *Nature* 599, 256-261. doi: 10.1038/s41586-021-04052-7

Zhu, S. H. (2012). Investigation of tuntian system during the Han dynasty. *J. Historic. Sci.* 10, 27–38.

frontiers | Frontiers in Genetics

TYPE Brief Research Report
PUBLISHED 22 August 2022
DOI 10.3389/fgene.2022.884612

OPEN ACCESS

EDITED BY
Maxat Zhabagin,
National Center for Biotechnology,
Kazakhstan

REVIEWED BY
Vagheesh Narasimhan,
University of Texas at Austin,
United States
John Hawks,
University of Wisconsin-Madison,
United States
Mário Vicente,
Stockholm University, Sweden

*CORRESPONDENCE
Perle Guarino-Vignon,
perle.gv@gmail.com
Céline Bon,
celine.bon@mnhn.fr

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

RECEIVED 26 February 2022
ACCEPTED 14 July 2022
PUBLISHED 22 August 2022

CITATION
Guarino-Vignon P, Marchi N, Chimènes A, Monnereau A, Kroll S, Mashkour M, Lhuillier J, Bendezu-Sarmiento J, Heyer E and Bon C (2022), Genetic analysis of a bronze age individual from Ulug-depe (Turkmenistan). *Front. Genet.* 13:884612. doi: 10.3389/fgene.2022.884612

Genetic analysis of a bronze age individual from Ulug-depe (Turkmenistan)

Perle Guarino-Vignon[1,2]*, Nina Marchi[3], Amélie Chimènes[1], Aurore Monnereau[1,4], Sonja Kroll[5], Marjan Mashkour[5], Johanna Lhuillier[6], Julio Bendezu-Sarmiento[1], Evelyne Heyer[1] and Céline Bon[1]*

[1]Eco-Anthropologie (EA), Muséum National D'Histoire Naturelle, CNRS, Université de Paris, Paris, France, [2]CAGT, UMR 5288, CNRS, Université Paul Sabatier Toulouse III, Toulouse, France, [3]CMPG, Institute of Ecology and Evolution, University of Berne, Berne, Switzerland, [4]BioArCh, Department of Archaeology, University of York, York, United Kingdom, [5]Archéozoologie, Archéobotanique Sociétés, Pratiques et Environnements (AASPE), Muséum National D'Histoire Naturelle, CNRS, Paris, France, [6]Archéorient, Environnements et Sociétés de L'Orient Ancien, CNRS/Université Lyon 2, Lyon, France

The Oxus Civilisation (or Bactrio-Margian Archaeological Complex, BMAC) was the main archaeological culture of the Bronze Age in southern Central Asia. Paleogenetic analyses were previously conducted mainly on samples from the eastern part of BMAC. The population associated with BMAC descends from local Chalcolithic populations, with some outliers of steppe or South-Asian descent. Here, we present new genome-wide data for one individual from Ulug-depe (Turkmenistan), one of the main BMAC sites, located at the southwestern edge of the BMAC. We demonstrate that this individual genetically belongs to the BMAC cluster. Using this genome, we confirm that modern Indo-Iranian–speaking populations from Central Asia derive their ancestry from BMAC populations, with additional gene flow from the western and the Altai steppes in higher proportions among the Tajiks than the Yagnobi ethnic group.

KEYWORDS

Paleogenetics, Central Asia, Bactrian-Margian Archaeological Complex, Bronze Age, Turkmenistan

Introduction

Central Asia was one of the first regions outside of Africa populated by *Homo sapiens*. It has played a key role in human history, having served for millennia as a connection between Europe and Asia on the one hand, and Siberia and southern Eurasia on the other. This area exhibits a high genetic, linguistic, and ethnological diversity. Today, several ethnic groups are present in Central Asia belonging to two linguistic families: Indo-Iranian—composed of the Tajik and Yagnobi populations, traditionally sedentary agriculturalists—and Turco Mongol that encompasses Uzbeks, Turkmens, Karakalpaks, Kazakhs, and Kyrgyz, traditionally nomadic herders and organized into patrilineal descent groups.

Genetic analyses led by our team on about 40 Central Asian populations brought to light a major impact of social organization and cultural practices on Central Asia's genetic diversity. The patrilocal rule of residence favors women's migrations while increasing the spatial, linguistic, and ethnic structuring of men (Heyer et al., 2009; Marchi et al., 2017) and could explain the larger genetic differentiation between populations for the paternally inherited Y chromosome than for their maternal equivalent, the mitochondrial DNA. The male effective population size, but not the female one, is found to be smaller for the patrilineal Turco-Mongols than for the cognatic Indo-Iranian populations (Ségurel et al., 2008), suggesting an effect of the filiation rule. Indeed, the patrilineal filiation rule drastically decreases their masculine effective population size (Chaix et al., 2007; Marchi et al., 2017; Ségurel et al., 2008) and accentuates the male transmission of reproductive success Heyer et al., (2015). Eventually, by contrasting estimated genetic and given historical ages for ethnic groups and tribes, we unveiled that ethnicity and tribalism in Central Asia are likely cultural constructions rather than biological entities (Quintana-Murci et al., 2004; Heyer et al., 2009; Marchi et al., 2017). These cultural behaviors are expected to affect the demographic history of Central Asian groups (Aimé et al., 2013; Aimé, Heyer, and Austerlitz 2015; Balaresque et al., 2015).

Genetic analyses have also helped to understand the origin and history of these populations. First, they evidenced that the two cultural groups are genetically distinct (Martínez-Cruz et al., 2011; Heyer et al., 2009; Marchi et al., 2018, 2017). Indo-Iranian speakers have the greatest proximity to modern Western Eurasian populations, while the Turco-Mongols are mostly related to Eastern Eurasians (Northern Asia). However, the Turkmens stand out from this general conclusion: despite speaking a Turco-Mongol language, they are more related to Indo-Iranian populations, suggesting a recent change, both in their language and way of living (Guarino-Vignon et al., 2022). Second, using approximate Bayesian computation, our group inferred that the Indo-Iranian group resulted from the first prehistoric admixture between Western and Eastern Eurasian groups (Palstra, Heyer, and Austerlitz 2015). Then, some 2.3 ky ago, the Turco-Mongol group emerged from a second admixture between these proto–Indo-Iranians and Eastern Eurasians. However, genetic analyses based only on modern data can be skewed by the recent and population-specific demographic history of the populations taken as references.

The development of paleogenomics over the last 15 years has allowed a better understanding of the demographic events that shaped the history of these populations. In this region with a complex migration history, ancient DNA is precious for disentangling the different waves of migration. Furthermore, relying on precise archaeological context made possible the joint study of cultural and demographic changes through time.

Notably, the Neolithic way of life developed in the steppe territory in the North (corresponding to present-day Kazakhstan, Kyrgyzstan, and Northern Uzbekistan) around 4000 BC with the Botaï culture (de Barros Damgaard et al., 2018). Later, during the Middle Bronze Age (around 2000 BC), the Botaï culture was replaced by people associated with the Sintashta culture, responsible for the introduction of wheeled chariots and horse breeding in the steppes. Sintashta being related both from a cultural and genetic point of view to Western steppe groups, their presence in Northern Central Asia during the Bronze and Iron Age (third-first mill. BC) suggests some West-to-East migrations (Allentoft et al., 2015; Jeong et al., 2019), as well as great mobility linked to pastoralism (Bendezu Sarmiento, 2007; Bendezu Sarmiento, 2021a). During the Middle-Late Bronze Age, there seems to be a continuity with the Andronovo complex that was derived from the Sintashta horizon. At the end of the Bronze Age—the beginning of the Iron Age, some East Asian populations expanded in this area, reflecting the likely onset of Turco-Mongol expansion westwards (Hollard et al., 2014; Unterländer et al., 2017).

In the South (present-day southern Uzbekistan, Tajikistan, Turkmenistan, southern Kyrgyzstan, northern Afghanistan, and Northeastern Iran), agropastoral communities are present since 6000 BC (D. R. Harris, Gosden, and Charles 1996) and as early as the eighth millennium BC in Northeastern Iran (Harris 2011; Roustaei, Mashkour, and Tengberg 2015; David R). These groups are genetically similar to Neolithic Iranian communities (Broushaki et al., 2016; Lazaridis et al., 2016), which may suggest that the agricultural way of life was acquired through the expansion of Southwestern Eurasian farmer populations in the south of Central Asia or that the local hunter-gatherer ancestry was related to a vast population found in Iran and Caucasus (Shinde et al., 2019). Through the Chalcolithic and the Bronze Age, the development of agriculture is associated with increasing size of the villages and the beginning of irrigation, which culminated with the blossoming of the Bactrian-Margian Archaeological Complex (BMAC) also called the Oxus Civilization (Lyonnet and Dubova, 2020). Genetic data obtained from several archaeological sites show a strong genetic continuity between the Neolithic and the beginning of BMAC, with only a limited genetic contribution of other groups. BMAC displays the first structured proto-urban cities of the area gathering thousands of individuals, and a deep social structuring (Muradov, 2021). BMAC was part of the "Middle Asian Interaction Sphere," a dynamic network of cultural interaction and interregional exchanges with the Indus Civilisation (northern India and Pakistan), the Syro-Anatolian area, Mesopotamia, and the Iranian Plateau (Possehl, 2002; Mutin et al., 2017). During the BMAC period, and more frequently after the Middle Bronze Age, some outliers, coming from Southern Asian or steppe populations are evidenced, suggesting that the long-distance relationships seen in the material assemblage reflect on the genetic level. For still unknown reasons, the Late Bronze Age (ca. 1800–1500 BC) corresponds to a major cultural, economic, and ideological shift in southern Central Asia, leading to the disappearance of the Oxus Civilisation and is characterized by deterioration in the quality of the craft industry and the disappearance of long-distance exchanges within the Middle Asia Interaction Sphere. However, contacts with neighboring steppe Andronovo cultural community increased during the Late and Final Bronze Age period (Rouse and Cerasetti, 2014). The following Iron Age is characterized by a mosaic of cultures,

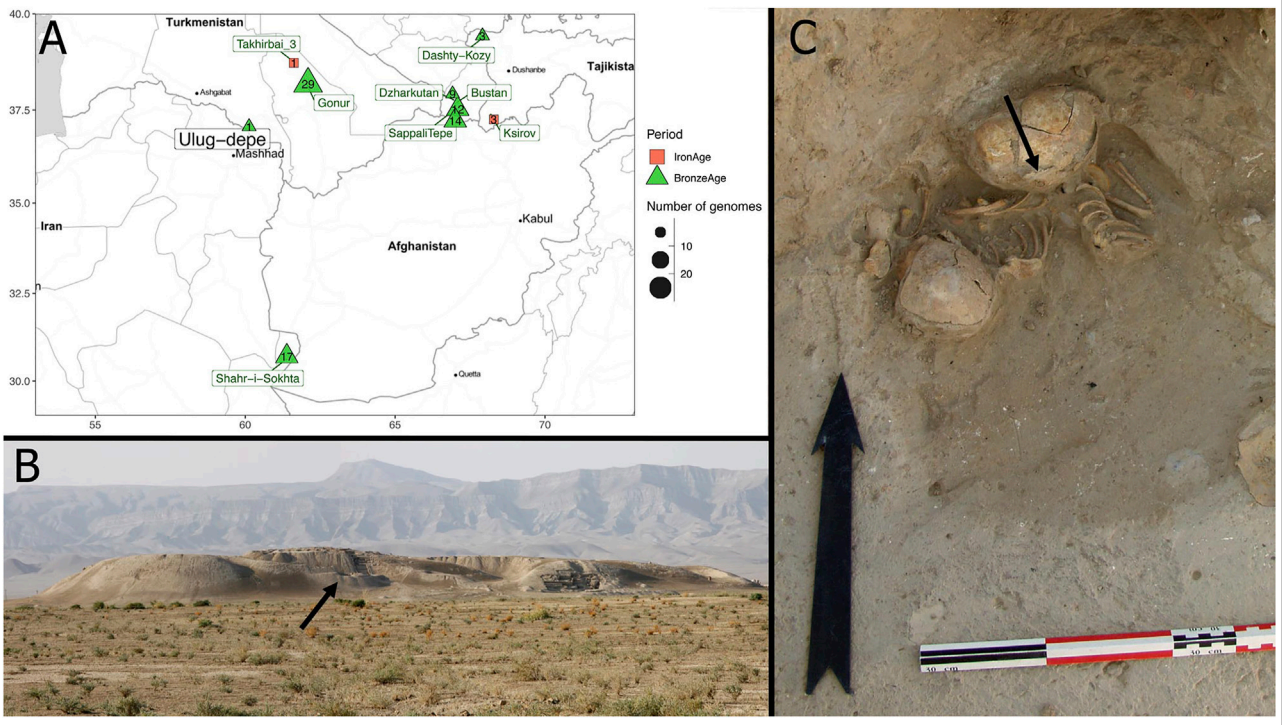


FIGURE 1
The archaeological context of *ULG75*. **(A)** Maps showing the different Turan Bronze Age (green) and Iron Age (red) sites from which genome-wide data were obtained. **(B)** View from Ulug-depe site from North-East. The place where *ULG75* was found is indicated by an arrow. **(C)** Photograph of the *ULG75* burial (grave 75). An arrow indicates the sample from which DNA was extracted.

characterized by specific handmade pottery, with red geometric designs, also known as "Yaz I cultures," which spread over a territory larger than the BMAC territory with a radical transformation of the settlement pattern, small sedentary villages replacing large proto-urban sites and spreading to new areas (Lhuillier, 2013; Lhuillier 2019). Because funerary practices change at this time, and inhumation is replaced by the exposition of corpses and defleshing by scavengers (Bendezu Sarmiento and Lhuillier 2013), the number of human remains for this period is low. For instance, genome-wide data have been published only for one individual from Turkmenistan Damgaard et al., (2018), whose ancestry is the result of an admixture between BMAC population and a steppe population related to Andronovo culture. Comparison of this individual with modern Indo-Iranians suggests genetic continuity since the Iron Age, as 90% of Yagnobis ancestry is inherited from BMAC, with only a limited pulse from an East Asian population and, for the Tajik group, from South Asia (Guarino-Vignon et al., 2022).

Despite the publication of genome-wide data from several BMAC archaeological sites, none has been published up to now from Ulug-depe (Turkmenistan). Ulug-depe is one of the biggest proto-urban sites of the BMAC (13 ha), and is located in the formative area of the BMAC, halfway between Namazga-depe and Altyn-depe (Figures 1A,B), at 175 km east of Ashgabat in Turkmenistan. The site was first studied by V.I. Sarianidi in the 1960s and by the MAFTur in the early 2000s. Ulug-depe displays the longest stratigraphical sequence of southern Central Asia, from Early Chalcolithic to the Middle Iron Age, making it a key site for the understanding of the genesis and the evolution of the BMAC.

No necropolis has been (yet) found in Ulug-depe, but up to 100 burials from the Bronze Age have been discovered inside the houses (Bendezu Sarmiento, 2021b). Inside a house in trench 1 Est (Figure 1C), a grave was installed in the corner of a room that was likely still occupied. The pit grave contained the remains of three perinatal juveniles (Julio, 2013; Bendezu Sarmiento, 2021b).

To better understand the genetic diversity of Bronze Age Ulug-depe, we performed a paleogenetic analysis of one of these individuals.

Materials and methods

Material

Thirteen samples from the BMAC period and three from Iron Age have been selected, based on the overall preservation (Supplementary Table S8). They span several occupancies of the

site from the Early Bronze Age (Namazga IV) to the Iron Age. The selected samples came from different areas of the skeleton: teeth, phalanx, and coastal fragments, as well as a petrous bone for one individual (grave 75). It belongs to an infant, around 10 months old, from the Middle Bronze Age (Namazga V period). The child was lying on the left side, in a crouched position, the head facing west. The bone deposit suggests that some disturbances have taken place, maybe when the last infant was buried.

Ancient DNA extraction

All the preamplification steps were carried out in the clean room, dedicated to ancient DNA, of the Paleogenomic and Molecular genetics platform, set in the Musée de l'Homme (Paris, France). Ancient DNA extraction was performed for the 15 samples using a protocol adapted from Dabney et al. (2013). Briefly, 50–100 mg of the petrous bone were powdered by drilling and incubated in 1 ml of lysis buffer (0.45 M EDTA, 10 mM Tris-HCl (pH 8.0), 0.1% SDS, 65 mM DTT, and 0.5 mg/ml proteinase K) at 37°C for 14 h. After centrifugation, 1 ml of supernatant was recovered and purified with 13 ml of binding buffer (5M GuHCl, 40% 2-propanol 0.05% Tween 20, 90 mM sodium acetate 2M, 1× phenol red). The mixture was then transferred on a High Pure Extender Assembly column (Roche High Pure Viral Nucleic Acid Large Volume Kit) and centrifuged. Then, the column was washed following manufacturer recommendations (briefly, 500 µl of inhibitor removal buffer, centrifugation, and 450 µl of wash buffer twice). DNA was eluted in 100 µl of elution buffer. The amount of human DNA in the samples was evaluated through PCR amplification. PCR was performed in a 12-µl reaction volume containing mock or ancient DNA extracts, 300 pM sense (TGGGGAAGCAGATTTGGGT) and antisense (TGG CTGGCAGTAATGTACG) primers targeting the mitochondrial DNA, 200 µM dNTP, 2.5 mM $MgCl_2$, 1 × PCR buffer II, and 1 U of AmpliTaq Gold DNA polymerase (Applied Biosystems). 1 µl of PCR product was loaded on a 2% agarose gel to check for positive results.

Library preparation

The ancient DNA extract was converted into a TruSeq Nano Illumina library following the manufacturer's protocol with slight modifications that account for the ancient DNA damage. First, DNA was not fragmented; only 25 µl of ancient DNA extract was used; after end-repair, the libraries were purified using a MinElute column (Qiagen ©); libraries were amplified using 10–12 PCR cycles and purified on a MinElute column (Qiagen ©). Analyses on a LabChip ® GX provided an estimated size distribution of fragments with a peak length of 150–250 bp.

Genome-wide data was successfully produced from the only sample (*ULG75* from grave 75) with more than 1% of endogenous DNA content. For this sample, a genomic capture was performed using the myBaits Expert Whole Genome Enrichment (WGE) kit (Arbor Biosciences) and the manufacturer's instructions were followed. Baits were formed from the genomic DNA of three individuals of different (African, European, and Asian) ancestries. After enrichment, libraries were re-amplified using 12 PCR cycles.

Sequencing

Captured libraries were sequenced on a NextSeq 500 (2 × 75 bp) instrument on the IGenSeq platform.

Data processing

We similarly processed the new sample as follows. We mapped sequencing reads to the human reference genome (GRCh37) using BWA-0.1.17 (Li and Durbin, 2009) with the *aln* command and the parameter "–l 1,024." Mapped reads were filtered out for mapping quality <25 with SAMtools 1.940. Duplicates were removed using Picard MarkDuplicates (http://picard. sourceforge.net), and the reads were realigned using GATK 3.5 IndelRealigner (Auwera et al., 2013). Finally, MapDamage 2 (Jónsson et al., 2013) was used to rescale base quality for all samples and take into account ancient DNA–specific damages. We restricted our analysis to known present-day DNA variants to minimize false positives. We used the *mpileup* command of SAMtools 1.9 (Li et al., 2009) to extract reads overlapping known variants from the v42.4 available at https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data covering 1,233,013 positions SNPs (1240k dataset). For positions with more than one base call, one allele was randomly chosen with a probability equal to the frequency of the base at that position.

We calculated the contamination rate on the genome using AuthentiCT (Peyrégne and Peter 2020).

Biological sex was determined using Rx (Mittnik et al., 2016) and Ry (Skoglund et al., 2013) statistics. Mitochondrial haplogroup was found using Haplogrep v2.4.0 (Kloss-Brandstätter et al., 2011) and compared to the AmtDB database (Ehler et al., 2019).

Merging genomic data

Among ancient human genomes from Eurasia from Paleolithic to Middle Age (Supplementary Table S1), DNA sequences were generated with whole genome shotgun or hybridization capture technics, from the 1240k dataset, and from Skourtanioti et al. (2020) and Jeong et al. (2020), we

retained 1,587 ancient unrelated Eurasian individuals with more than 10,000 SNP hits on the 1240 k panel from the Human Origin dataset. We merged our individual with these 1,587 published individuals using *mergeit* from eigensoft. This first merge is called the "1240 k-dataset" and includes 1,233,013 SNPs. We provide all metadata about the ancient dataset in Supplementary Table S1.

For some analyses, it was then merged with 22 modern Indo-Iranian genomes [as described in Guarino-Vignon et al. (2022)] and included 716,743 SNPs.

For analysis requiring more modern diversity, we selected 3,109 published modern genomes from Eurasia, Mbuti population (the latter to serve as outgroup), from the Human Origin dataset and we haploidized them by randomly selecting one allele per position. We merged the modern dataset with the 1240 k-dataset as before, in order to provide a frame of comparison regarding the modern Eurasian diversity. The final merge was called "HO-dataset" and included 597,573 SNPs for 4,696 individuals.

Descriptive analysis

We ran principal component analyses (PCA) with *smartpca* (Patterson, Price, and Reich 2006) on the HO-dataset for 1,309 European and Middle-Eastern modern individuals, and we projected all the ancient samples. We used the default parameters with *lsqproject: YES*, and *numoutlieriter: 0* settings.

We computed ADMIXTURE analysis (Alexander, Novembre, and Lange 2009) on the HO dataset, downsampling all populations to a maximum of 20 individuals, that is, for 1,266 Eurasian modern individuals, including East Asians, and 2,526 ancient samples. A subset of 365,075 SNPs were retained for the analysis after pruning for linkage disequilibrium done by using *plink 1.9* --indep-pairwise 200 25 0.4 function (Chang et al., 2015). We ran 10 replicated ADMIXTURE analyses for K between 2 and 15, from which we kept the most likely.

D-statistics and f3 statistics

We performed D-statistics on the "1,240 k dataset" using the *qpDstat* program of the ADMIXTOOLS package (Patterson et al., 2012). We computed D-statistics of the form D(Mbuti, Y; BMAC populations, *ULG75*), with *Y* being ancient Paleolithic, Mesolithic, or Neolithic populations from Western Eurasia and most particularly southern Central Asia to test the proximity of the Ulug individual *ULG75* with other BMAC and post-BMAC groups. We also computed all D (Mbuti, Y; Indo-Iranian, *ULG75*), with Y being protohistoric Eurasian populations, and Indo-Iranians being Tajiks and Yagnobis. We corrected the Z-score of every D-statistic accounting for the repetitive testing using Benjamini and Hochberg's (1995) method. We used *qp3pop* of the same package, using *inbreed: YES* parameter to compute f3-outgroup statistics f3 (Mbuti, BMAC population, ancient Eurasian population) to identify the Eurasian populations that share the highest genetic drift with Ulug-depe and four other BMAC groups. We only retained f3-statistics calculated on more than 50,000 SNPs.

qpAdm analysis

We performed rotating *qpAdm* analysis from ADMIXTOOLS package, using the "1,240 k dataset," to model the ancestry of *ULG75*, with the same set of populations as in Narasimhan et al. (2019).

We also performed the cladality test of *ULG75* with other BMAC populations compared to outlier populations from the region (Indus periphery pool, Shahr I Sokhta, and Seh Gabi) using qpAdm, with only BMAC as a source and the outliers added to the outgroups.

Uniparental markers

Y chromosome and mitochondrial haplogroups were determined by comparison of *vcf* files to databases: respectively, Phylotree 17 (Oven and Kayser, 2009) using the software Haplogrep v2.241 (Kloss-Brandstätter et al., 2011); and ISOGG version 11 July 2020.

Results

For 15 samples, no or weak PCR signal was observed, indicating that ancient DNA was not well preserved in these samples. PCR amplifications were robustly possible only for *ULG75*. We generated genome-wide data for this individual dated to the Bronze Age from the site of Ulug-depe in Turkmenistan, associated with BMAC. We obtained damage profile compatibles with ancient DNA (Supplementary Figure S1), and the estimation of contamination was low (0.047). The estimation of Rx (1.09+/−0.06) and Ry (0.0159+/−2 E-7) statistics shows that *ULG75* is a female. We obtained 133,761 SNPs, which is sufficient for genome-wide analyses.

Mitochondrial DNA analyses

Mitochondrial haplogroup of *ULG75* is HV + 3,197 + 12,358 +16,311. The HV haplogroup is widely present in western Eurasia since the Neolithic, but this haplotype is rare. It has currently not been evidenced in any ancient DNA database, but the HV + 16,311 haplotype is present in two Central European Bell Beaker individuals (Ehler et al., 2019).

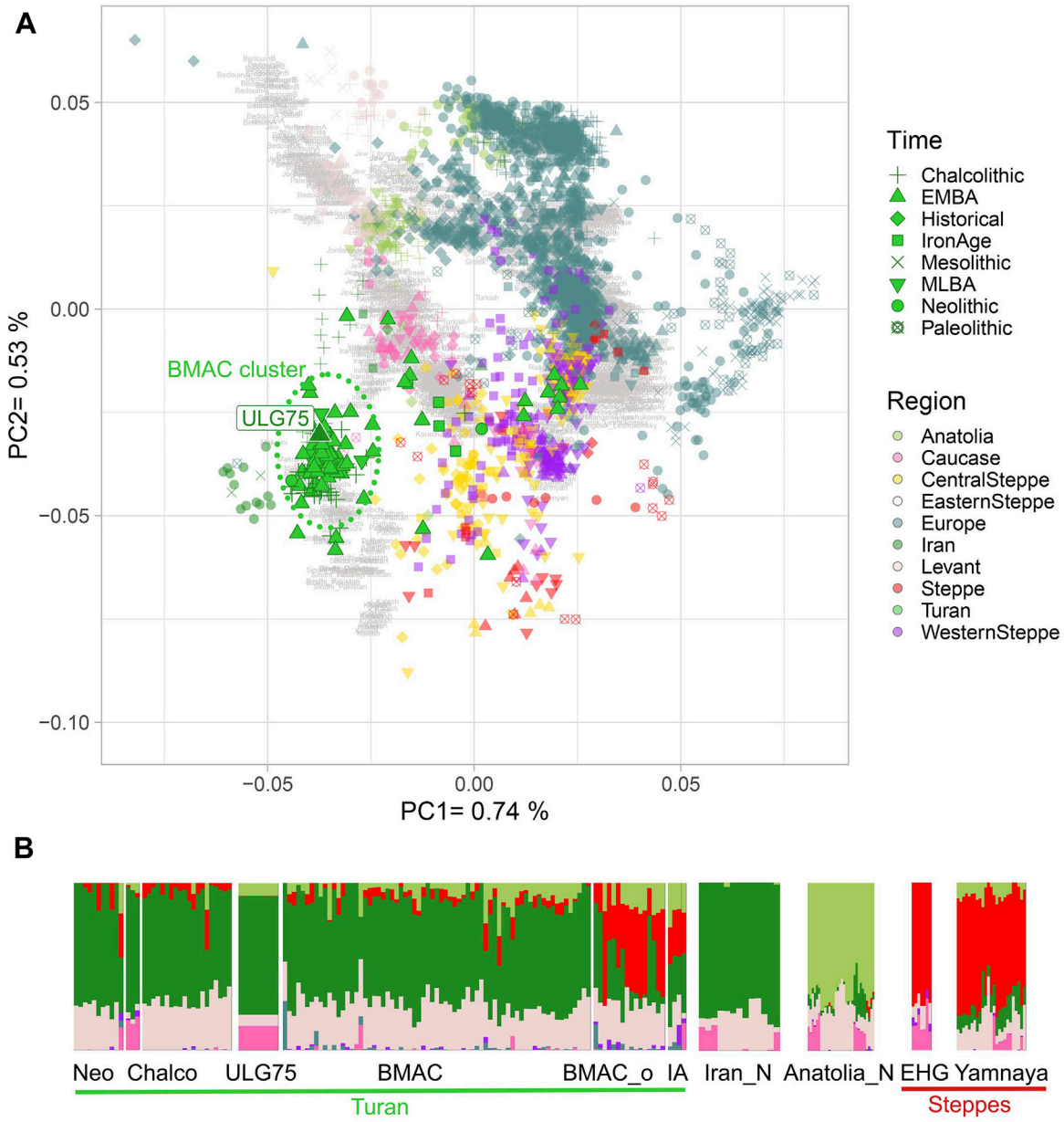


FIGURE 2
Genetic affinity of *ULG75* with other Turan individuals. **(A)** PCA is calculated on present-day individuals with ancient individuals projected onto it. Ancient Turan individuals are shown in non-transparent green with outlines. *ULG75* individual is labeled. **(B)** Unsupervised ADMIXTURE analysis was performed on the HO-dataset. Only *ULG75*, ancient Turan individuals, Neolithic Iranian Anatolian farmers, and Paleolithic and Eneolithic steppe populations are shown. Full ADMIXTURE is available in the supplementary data. Neo, Neolithic; Chalco, Chalcolithic; BMAC_o, outliers from BMAC sites; IA, Iron Age; Iran_N, Neolithic Iranian farmers; Anatolia_N, Neolithic Anatolian farmers; EHG, Eastern hunter-gatherers.

Genetic affinity of the Ulug individual with other BMAC individuals

To decipher the genetic relations between the new Ulug-depe individual and other ancient populations belonging to BMAC, we calculated a PCA on the modern individuals of the "HO-dataset" and projected ancient individuals (Supplementary Table S1): *ULG75* falls within all other individuals of the Oxus civilization (Figure 2), including those from Dzharkutan or Gonur-Depe. To further explore the genetic similarity of our Ulug sample with the other individuals from BMAC, we performed an ADMIXTURE analysis (Alexander, Novembre, and Lange 2009). Previous analyses of BMAC individuals have shown that these sites receive individuals from other origins

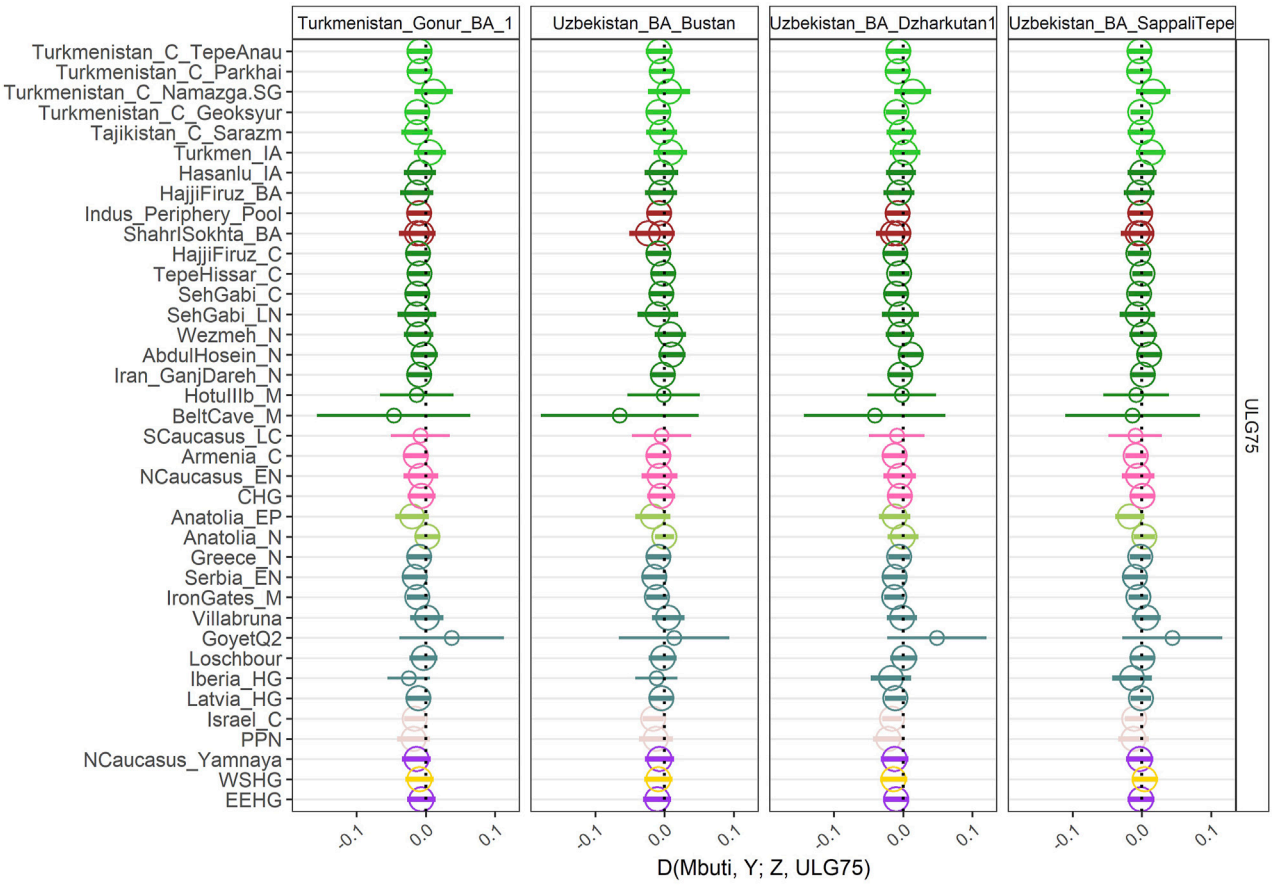


FIGURE 3
Genetic homogeneity of *ULG75* with other BMAC populations. D-statistics of the form D (Mbuti, Ancient Pop; BMAC, *ULG75*). The hollow point indicates a null D-statistic (|Z|<3). A small point indicates that the D-statistic was calculated with less than 50,000 SNPs.

(Steppe populations and South-Asian populations), later identified as outliers. *ULG75* presents a profile close to the other BMAC individuals and to ancient Neolithic and Chalcolithic Turan groups, without the Steppe component (red) that is found in Iron Age individuals from Turan and outliers from BMAC (Figure 2; Supplementary Figure S3). We also estimated with f3-outgroup statistics of the form f3 (Mbuti; *ULG75*, PopX) that *ULG75* shares the most genetic drift with Turan groups from Chalcolithic to Late Bronze Age (Supplementary Figure S3). By comparing these results to those of the f3 (Mbuti; BMAC populations, PopX), we find that values observed for Ulug-depe are highly correlated with those obtained for other BMAC individuals (r2 = 0.93, p = 0) (Supplementary Figure S4).

To confirm that our individual forms a homogeneous group with the already published BMAC individuals, we also calculated D-statistics D (Mbuti, Ancient population; *ULG75*, BMAC). For all the D-statistics, we only obtained null D-stats (Figure 3) further reinforcing that *ULG75* belongs to the BMAC cluster. To fully test the cladality of *ULG75* with BMAC regarding eastern outlier from the Bronze Age, we performed qpAdm analysis with only previously published BMAC individuals as a unique source for *ULG75* and outliers populations, such as Shahr I Sokhta, the Indus Periphery Pool cluster, and Chalcolithic individuals from Seh Gabi. We did obtain non-significant results for all the tests, meaning the clade formed by *ULG75* with BMAC could not be broken.

Eventually, we tried to model *ULG75* as the BMAC groups (formed by several individuals from different sites) were modeled in Narasimhan et al. (2019). All the models worked, but with notably bigger standard error for the admixture percentage estimation, probably because we only considered one individual of limited coverage. We were able to confidently model *ULG75* (p-values = 0.46 and 0.63) as the product of an admixture of 54–63% (+/−10%) of the Chalcolithic population from Geoksyur, 12% (+/−5%) of Chalcolithic population from Hajji Firuz or 19% (+/−9%) from Seh Gabi, both in Iran, and 24% (+/−7%) of populations from the Indus Periphery. For the rest of

the models, we obtained *p*-values above 0.05, but contributions from the admixed populations differed from what Narasimhan et al. (2019) obtained. Mostly, *ULG75* seems to be better modeled with only one contribution from the Bronze Age population, genetically part of the BMAC cluster, of Shahr-i-Shokhta (Iran), as nested models—inferring *ULG75* as 100% as Shahr-i-Shokhta—in our 3-way or 2-way admixtures could not be rejected (*p*-value > 0.05).

Genetic continuity between BMAC group and modern Indo-Iranian populations

By computing D-statistics D (Mbuti, A selection of Ancient Populations (Y); Indo-Iranians, *ULG75*), we observed genetic continuity between *ULG75* and modern Indo-Iranian (Supplementary Figure S5) with clear gene flows from different steppe populations in Tajiks. We first obtained negative D-statistics (Z < −3) for Tajiks and Alakul, Saka, and Eastern European hunter-gatherers (from the Pontic Steppe), which points to a gene flow originating from the western or central steppes. We also have negative D-statistics of the form D (Mbuti, XiongNu/Shamanka, Tajiks, *ULG75*), which indicates a gene flow from the Eastern steppe with a strong Baikal ancestry.

We do not obtain significantly negative D-stats for Yagnobis, but the closer genetic proximity between *ULG75* and Iranian farmers from Ganj Dareh and Chalcolithic populations from Turan than Yagnobis and Tajiks [positive D (Mbuti, Iran_N/Turan_C; *ULG75*, Tajiks/Yagnobis)] indicates that the latter lost Iranian Neolithic ancestry since Bronze Age, probably linked to a gene flow from a population with low Iranian Neolithic ancestry, like most of the Steppe populations. Nevertheless, the low coverage of *ULG75* may not give us enough sensitivity to identify all the gene flows.

Discussion and conclusion

Genetic homogeneity in bronze age BMAC

Ancient DNA analyses of a Middle Bronze Age individual from Ulug-depe, *ULG75*, show that it belongs to the genetic BMAC cluster, represented by individuals from Gonur-depe, Dzharkutan, Bustan, and Sappali-tepe. This reinforces the integration of Ulug-depe into the Oxus Civilization, at a population level. Strontium and oxygen isotope analyses of Early and Middle Bronze Age individuals from Ulug-depe (Kroll et al., n.d.) have shown remarkable mobility in earlier periods, which clearly decreased with the rise of the BMAC. This inter-site mobility, found by the isotopic results, during the beginning of the Oxus Civilization has certainly contributed to the observed genetic homogeneity between the BMAC settlements.

On the other hand, previous genetic analyses on Dzharkutan, Gonur-Depe, Bustan, and Sappali-tepe Bronze Age populations have shown that in addition to the BMAC genetic cluster, first- or second-generation migrants from various parts of Eurasia were present within the Oxus. Interestingly, these outliers cluster with two different populations: one is related to steppe populations, from Northern Central Asia, while the second to the population from the Indus periphery. Due to the poor preservation of DNA in Ulug-depe, a comparison of the genetic diversity inside the site was not feasible. However, the absence of genetic ancestry coming from either of these populations in *ULG75* suggests if these migrants were present in Ulug-depe, they were not part of the ancestors of this individual. From an isotope point of view, the individuals contemporaneous of *ULG75* revealed little mobility and extensive use of the direct surrounding; no foreigners could have been identified (Kroll et al., n.d.).

Origin and legacy of bronze age BMAC

As in Narasimhan et al. (2019), we observed that the BMAC individuals are largely derived from local, Chalcolithic populations similar to those found in Geoksyur. This genetic continuity matches with the long-term occupancy of several BMAC settlements and particularly Ulug-depe.

From a genetic point of view, the non-outlier BMAC individuals are strongly related to southerner populations such as the Iranian Neolithic, an observation that is also congruent with archaeological observations. The Turan archaeological sequence appears to mirror those found in South-western Asia. As early as the late seventh millennium BC, the Neolithic communities of South Central Asia have the same pattern of subsistence as those in the Near East and Iran; by the end of the fourth millennium, they also share the technological base for the production of pottery, metal, irrigation, etc., with numerous shreds of evidence showing material contact with the Iranian Plateau (Dani and Masson, 1992). These contacts may have implied gene flow as well and participate in the genetic homogeneity in this vast area.

This genetic continuity still goes on. A comparison of modern Indo-Iranian populations with the ancient genome (Guarino-Vignon et al., 2022) has shown genetic continuity with the BMAC cluster, with a high admixture with steppe populations, that occurred at the end of the BMAC. The first genetic outliers observed in several BMAC individuals may represent the very beginning of this demographic event. We also find evidence of gene flow from the East after the Iron Age, best modeled by XiongNu populations. This result is consistent with what has been found previously in Guarino-Vignon et al. (2022), reinforcing the XiongNu as a good proxy for modeling the gene flow that formed the Turko-Mongol populations and admixed with the local Indo-Iranian populations.

A more limited gene flow with the South-Asian populations has also been evidenced in Tajiks but not in Yagnobis, suggesting that it occurred mainly after the split between these two Indo-Iranian–speaking groups. On the other hand, outliers with a high

amount of the Indus Periphery group are found in several BMAC groups, such as Bustan, and the *ULG75* genome can be modeled with as much as 24% of its ancestry derived from Indus Periphery. Thus, the small amount of South-Asian ancestry found in the modern Turan population may be explained by an ancient, low, continuous, gene-flow.

This analysis shows the strength of adding ancient DNA data to better understand the evolution of genetic diversity in Central Asia. Several questions remain about the timing of the different demographic events and call for more paleogenetic data, despite the difficulties due to the low preservation of ancient DNA in this warm, arid region.

Data availability statement

The datasets presented in this study can be found in online repositories. Study accession number: ERP135732 or PRJEB51133.

Author contributions

PG-V, EH, and CB conceived the study. JL and JB-S collected the sample. CB, MM, and SK participated in the sample molecular characterization. AC and AM carried out molecular genetic studies. PG-V, NM, and CB performed the bioinformatic analyses. All the authors participated in the writing of the article.

Funding

This project has been funded by Sorbonne Université (PaleOxus project SU-16-R-EMR-06). Archaeological excavations were permitted thanks to the work of the French Archaeological Mission in Turkmenistan (MAFTUR) under the direction until 2013 of Olivier Lecomte (†) and the funding of the Ministry of Europe and Foreign Affairs (MEAE) and the Leon-Levy Foundation. PG-V is supported by a PhD grant provided by Ecole Normale Supérieure de Lyon.

Acknowledgments

We thank the French Archaeological Mission in Turkmenistan; Mission archéologique française au Turkménistan (MAFTUR) for the archaeological excavations, and all the people working on site. We thank the Paleogenomic and molecular genetics platform (P2GM) of the MNHN (Paris) as well as the IGenSeq platform of the Institut du Cerveau et de la Moelle (Paris).

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.884612/full#supplementary-material

References

Aimé, Carla, Heyer, Evelyne, and Austerlitz, Frédéric (2015). Inference of sex-specific expansion patterns in human populations from Y-chromosome polymorphism. *Am. J. Phys. Anthropol.* 157 (2), 217–225. doi:10.1002/ajpa.22707

Aimé, Carla, Laval, Guillaume, Patin, Etienne, Paul, Verdu, Ségurel, Laure, Chaix, Raphaëlle, et al. (2013). Human genetic data reveal contrasting demographic patterns between sedentary and nomadic populations that predate the emergence of farming. *Mol. Biol. Evol.* 30 (12), 2629–2644. doi:10.1093/molbev/mst156

Alexander, David H., Novembre, John, and Lange., Kenneth (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19 (9), 1655–1664. doi:10.1101/gr.094052.109

Allentoft, Morten E., Martin, Sikora, Sjögren, Karl Göran, Rasmussen, Simon, Rasmussen, Morten, Jesper, Stenderup, et al. (2015). Population genomics of Bronze age Eurasia. *Nature* 522 (7555), 167–172. doi:10.1038/nature14507

Auwera, Geraldine A., Carneiro, Mauricio O., Hartl, Christopher, Ryan, Poplin, Guillermo del AngelLevy-Moonshine, Ami, et al. (2013). From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* 43 (1), 11. doi:10.1002/0471250953.bi1110s43

Balaresque, Patricia, Poulet, Nicolas, Cussat-Blanc, Sylvain, Gerard, Patrice, Quintana-Murci, Lluis, Heyer, Evelyne, et al. (2015). Y-chromosome descent clusters and male differential reproductive success: Young lineage expansions dominate asian pastoral nomadic populations. *Eur. J. Hum. Genet.* 23 (10), 1413–1422. doi:10.1038/ejhg.2014.285

Bendezu Sarmiento, Julio (2013). Archéologie Funéraire et Bio-Anthropologie à Ulug Dépé et Dzharkutan. Âge Du Bronze Au Turkménistan et En Ouzbékistan. *L'archéologie Française En. Asie Centrale, Cahiers d'Asie centrale* 21/22, 501–532.

Bendezu Sarmiento, J., and Lhuillier, J. (2014). Sine sepulchro cultural complex of transoxiana (between 1500 and the Middle of the 1st millennium BC). Funerary

practices of the Iron age in southern central Asia: Recent work, old data, and new hypotheses. *Archäologische Mittl. Aus Iran. Und Turan (Amit)* 45, 44.

Bendezu Sarmiento, J. (2021a). "The first nomads in central Asia's steppes (Kazakhstan). An overview of major socio-economic changes, derived from funerary practices of the Andronovo and Saka populations of the Bronze and Iron ages (2nd-1st millennium BCE)," in *Nomad lives, natures en sociétés*. Editors A. Averbouh, N. Goutas, S. Méry, and MNHN (Paris: MNHN), 478–503.

Bendezu Sarmiento, J. (2021b). "Funerary rituals and archaeothanatological data from BMAC graves at Ulug depe (Turkmenistan) and dzharkutan (Uzbekistan)," in *The world of the Oxus civilization* (Oxfordshire, England, UK: Routledge), 405–424. Bertille Lyonnet and Nadezhda A. Dubova.

Bendezu Sarmiento, J (2007). Du Bronze À L'âge du fer Au Kazakhstan gestes funéraires et paramètres biologiques identités culturelles des populations Andronovo et Saka. *Memoires De. La Mission. Archeol. Francaise En. Asie Cent.* 1, 219–263.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. Available at: http://www.jstor.org/stable/2346101. doi:10.1111/j.2517-6161.1995.tb02031.x

Broushaki, Farnaz, Thomas, Mark G., Link, Vivian, López, Saioa, Lucy van DorpKirsanow, Karola, et al. (2016). Early neolithic genomes from the eastern fertile crescent. *Science* 353 (6298), 499–503. doi:10.1126/science.aaf7943

Chaix, Raphaëlle, Quintana-Murci, Lluís, Hegay, Tatyana, Hammer, Michael F., Mobasher, Zahra, Austerlitz, Frédéric, et al. (2007). From social to genetic structures in central Asia. *Curr. Biol.* 17 (1), 43–48. doi:10.1016/j.cub.2006.10.058

Chang, Christopher C., Chow, Carson C. Laurent C. A. M. Tellier, Vattikuti, Shashaank, Purcell, Shaun M., and Lee, James J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* 4 (1), 7. doi:10.1186/s13742-015-0047-8

Dabney, Jesse, Knapp, Michael, Glocke, Isabelle, Gansauge, Marie-Theres, Weihmann, Antje, Nickel, Birgit, et al. (2013). Complete mitochondrial genome sequence of a Middle pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U. S. A.* 110 (39), 15758–15763. doi:10.1073/pnas.1314445110

Damgaard, Barros, Peter, de, Martiniano, Rui, Kamm, Jack, Víctor Moreno-Mayar, J., Kroonen, Guus, 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

Dani, Ahmad Hasan, and Masson, V. M. (1992). *The history of civilzations of central Asia: The dawn of civilization: Earliest times to 700 BC*. Paris: Unesco.

Ehler, Edvard, Novotńy, Jiři, Anna, Juras, Chylénski, Maciej, Moravčik, Ondřej, Jan, Pačes, et al. (2019). AmtDB: A database of ancient human mitochondrial genomes. *Nucleic Acids Res.* 47 (D1), D29–D32. doi:10.1093/nar/gky843

Guarino-Vignon, Perle, Marchi, Nina, Bendezu-Sarmiento, Julio, Heyer, Evelyne, and Bon, Céline (2022). Genetic continuity of Indo-Iranian speakers since the Iron age in southern central Asia. *Sci. Rep.* 12 (1), 733. doi:10.1038/s41598-021-04144-4

Harris, David R. (2011). *Origins of agriculture in western central Asia an environmental-archaeological study*. Philadelphia: Incorporated University of Pennsylvania Press.

Harris, D. R., Gosden, Chris, and Charles, M. P. (1996). Jeitun: Recent excavations at an early neolithic site in southern Turkmenistan. *Proc. Prehist. Soc.* 62, 423–442. doi:10.1017/s0079497x00002863

Heyer, Evelyne, Balaresque, Patricia, Jobling, Mark a., Quintana-Murci, Lluis, Chaix, Raphaelle, Segurel, Laure, et al. (2009). Genetic diversity and the emergence of ethnic groups in central Asia. *BMC Genet.* 10, 49. doi:10.1186/1471-2156-10-49

Heyer, Evelyne, Tristan Brandenburg, Jean, Leonardi, Michela, Bruno, Toupance, Balaresque, Patricia, Hegay, Tanya, et al. (2015). Patrilineal populations show more male transmission of reproductive success than cognatic populations in central Asia, which reduces their genetic diversity. *Am. J. Phys. Anthropol.* 157 (4), 537–543. doi:10.1002/ajpa.22739

Hollard, Clémence, Keyser, Christine, Giscard, Pierre Henri, Tsagaan, Turbat, Bayarkhuu, Noost, Jan, Bemmann, et al. (2014). Strong genetic admixture in the Altai at the Middle Bronze age revealed by uniparental and ancestry informative markers. *Forensic Sci. Int. Genet.* 12, 199–207. doi:10.1016/j.fsigen.2014.05.012

Jeong, Choongwon, Balanovsky, Oleg, Lukianova, Elena, Kahbatkyzy, Nurzhibek, Flegontov, Pavel, Zaporozhchenko, Valery, 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, Choongwon, Wang, Ke, Shevan WilkinTaylor, William, Timothy, Treal, Miller, Bryan K., Bemmann, Jan 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

Jónsson, Hákon, Ginolhac, Aurélien, Schubert, Mikkel, Philip, L., Johnson, F., and Orlando, Ludovic (2013). MapDamage2.0: Fast approximate bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29 (13), 1682–1684. doi:10.1093/bioinformatics/btt193

Kloss-Brandstätter, Anita, Pacher, Dominic, Schönherr, Sebastian, Weissensteiner, Hansi, Binna, Robert, Specht, Günther, et al. (2011). HaploGrep: A fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum. Mutat.* 32 (1), 25–32. doi:10.1002/humu.21382

Kroll, S. K., Bendezu-Sarmiento, Julio, Lhuillier, Johanna, Luneau, Élise, Kaniuth, K., Teufer, M., et al. (n.d). Mobility and land use in the great khorasan civilization: Isotopic approaches (87Sr/86Sr, Δ18O) on human populations from southern central Asia. *Submitt. JAS*.

Lazaridis, Iosif, Nadel, Dani, Gary, Rollefson, Merrett, Deborah C., Rohland, Nadin, Mallick, Swapan, et al. (2016). Genomic insights into the origin of farming in the ancient Near East. *Nature* 536 (7617), 419–424. doi:10.1038/nature19310

Lhuillier, Johanna. (2013). Les cultures à céramique modelée peinte en Asie centrale méridionale. Dynamiques socio-culturelles à l'âge du Fer ancien (1500-1000 av. n.è.). De Boccard.

Lhuillier, J. (2019). "The settlement pattern in central Asia during the early Iron age," in Urban Cultures of Central Asia from the Bronze Age to the Karakhanids. Learnings and Conclusions from New Archaeological Investigations and Discoveries, Proceedings of the First International Congress on Central Asian Archaeology Held at the University Of Bern, February 4–6, 2016 edited by Christoph Baumer and Mirco Novak, Schriften (Harrassowitz Verlag) , 115–128.

Li, Heng, and Durbin, Richard (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25 (14), 1754–1760. doi:10.1093/bioinformatics/btp324

Li, Heng, Handsaker, Bob, Wysoker, Alec, Tim, Fennell, Ruan, Jue, Homer, Nils, et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25 (16), 2078–2079. doi:10.1093/bioinformatics/btp352

Lyonnet, Bertille, and Dubova, Nadezhda (2020). "The world of the Oxus civilization," in *Bertille lyonnet and nadezhda A* (Dubova: Routledge).

Marchi, Nina, Hegay, Tatyana, Mennecier, Philippe, Georges, Myriam, Laurent, Romain, Whitten, Mark, et al. (2017). Sex-specific genetic diversity is shaped by cultural factors in inner asian human populations. *Am. J. Phys. Anthropol.* 162 (4), 627–640. doi:10.1002/ajpa.23151

Marchi, Nina, Mennecier, Philippe, Georges, Myriam, Lafosse, Sophie, Hegay, Tatyana, Dorzhu, Choduraa, et al. (2018). Close inbreeding and low genetic diversity in inner asian human populations despite geographical exogamy. *Sci. Rep.* 8 (1), 9397. doi:10.1038/s41598-018-27047-3

Martínez-Cruz, Begoña, Vitalis, Renaud, Ségurel, Laure, Austerlitz, Frédéric, Georges, Myriam, Théry, Sylvain, et al. (2011). In the heartland of Eurasia: The multilocus genetic landscape of central asian populations. *Eur. J. Hum. Genet.* 19 (2), 216–223. doi:10.1038/ejhg.2010.153

Mittnik, Alissa, Wang, Chuan Chao, Svoboda, Jiří, and Krause, Johannes (2016). A molecular approach to the sexing of the triple burial at the upper paleolithic site of dolní vêstonice. *PLoS ONE* 11 (10), e0163019. doi:10.1371/journal.pone.0163019

Muradov, R. (2021). "The architecture of bactria-margiana archaeological culture," in *The word of the Oxus civilization* (London: Routledge), 145–177. Bertille Lyonnet and Nadezhda Dubova.

Mutin, Benjamin, Minc, Leah D., CarlKarlovsky, C. Lamberg, and Tosi, Maurizio (2017). Regional and long-distance exchange of an emblematic 'prestige' ceramic in the Indo-Iranian borderlands. Results of neutron activation analysis. *paleo.* 43 (1), 141–162. doi:10.3406/paleo.2017.5755

Narasimhan, Vagheesh M., Patterson, Nick, Priya, Moorjani, Rohland, Nadin, Bernardos, Rebecca, Mallick, Swapan, et al. (2019). The formation of human populations in South and central Asia. *Science* 365 (6457), eaat7487. doi:10.1126/science.aat7487

Oven, Mannis van, and Kayser, Manfred (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* 30 (2), 386–394. doi:10.1002/humu.20921

Palstra, Friso P., Heyer, Evelyne, and Austerlitz., Frédéric (2015). Statistical inference on genetic data reveals the complex demographic history of human populations in central Asia. *Mol. Biol. Evol.* 32 (6), 1411–1424. doi:10.1093/molbev/msv030

Patterson, Nick, Price, Alkes L., and Reich, David (2006). Population structure and eigenanalysis. *PLoS Genet.* 2 (12), e190. doi:10.1371/journal.pgen.0020190

Patterson, Nick, Priya, Moorjani, Luo, Yontao, Mallick, Swapan, Rohland, Nadin, Zhan, Yiping, et al. (2012). Ancient admixture in human history. *Genetics* 192 (3), 1065–1093. doi:10.1534/genetics.112.145037

Peyrégne, Stéphane, and Peter, Benjamin M. (2020). AuthentiCT: A model of ancient DNA damage to estimate the proportion of present-day DNA contamination. *Genome Biol.* 21 (1), 246. doi:10.1186/s13059-020-02123-y

Possehl, G. L. (2002). *The Indus civilization : A contemporary perspective*. Walnut Creek, CA: AltaMira P.

Quintana-Murci, Lluís, Chaix, Raphaëlle, Spencer Wells, R., Behar, Doron M., Sayar, Hamid, Scozzari, Rosaria, et al. (2004). Where west meets east: The complex MtDNA landscape of the southwest and central asian corridor. *Am. J. Hum. Genet.* 74 (5), 827–845. doi:10.1086/383236

Rouse, Lynne M., and Cerasetti, Barbara (2014). Ojakly: A late Bronze age mobile pastoralist site in the murghab region, Turkmenistan. *J. Field Archaeol.* 39 (1), 32–50. doi:10.1179/0093469013z.00000000073

Roustaei, Kourosh, Mashkour, Marjan, and Tengberg, Margareta (2015). Tappeh sang-e chakhmaq and the beginning of the neolithic in North-east Iran. *Antiquity* 89 (345), 573–595. doi:10.15184/aqy.2015.26

Ségurel, Laure, Martínez-Cruz, Begoña, Quintana-Murci, Lluis, Balaresque, Patricia, Georges, Myriam, Hegay, Tatiana, et al. (2008). Sex-specific genetic structure and social organization in central Asia: Insights from a multi-locus study. *PLoS Genet.* 4 (9), e1000200. doi:10.1371/journal.pgen.1000200

Shinde, V., Narasimhan, V. M., Rohland, N., Mallick, S., Mah, M., Lipson, M., et al. (2019). An ancient harappan genome lacks ancestry from steppe pastoralists or Iranian farmers. *Cell* 179 (3), 729–735. e10Epub 2019 Sep 5. PMID: 31495572; PMCID: PMC6800651. doi:10.1016/j.cell.2019.08.048

Skoglund, Pontus, Jan, Storå, Götherström, Anders, and Jakobsson, Mattias (2013). Accurate sex identification of ancient human remains using DNA shotgun sequencing. *J. Archaeol. Sci.* 40 (12), 4477–4482. doi:10.1016/j.jas.2013.07.004

Skourtanioti, Eirini, Erdal, Yilmaz S., Frangipane, Marcella, Balossi Restelli, Francesca, Yener, K. Aslıhan, Pinnock, Frances, et al. (2020). Genomic history of neolithic to Bronze age Anatolia, northern levant, and southern Caucasus. *Cell* 181 (5), 1158–1175. e28. doi:10.1016/j.cell.2020.04.044

Unterländer, Martina, Palstra, Friso, Lazaridis, Iosif, Pilipenko, Aleksandr, Hofmanová, Zuzana, Groß, Melanie, et al. (2017). Ancestry and demography and descendants of Iron age nomads of the eurasian steppe. *Nat. Commun.* 8, 14615. doi:10.1038/ncomms14615

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