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

Bin Ma†, Jinwen Chen‡, Xiaomin Yang§, Jingya Bai¶, Siwei Ouyang†, Xiaodan Mo†, Wangsheng Chen†, Chuan-Chao Wang§,¶,¶ and Xiangjun Hai†,¶

1Key Laboratory of Environmental Ecology and Population Health in Northwest Minority Areas, Northwest Minzu University, Lanzhou, China, 2State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen, China, 3Department 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, 4State 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, 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, and Salar, as well as the Hui group. Analysis based on allele-sharing and haplotype-sharing were used to elucidate the population history of Northeast 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 flourished 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
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; 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 have 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 Eastern 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; Tangkanchanapras 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 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; Maramovitch 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; Antwerpens 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 Chemurche 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 et al., 2021)....
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 included individuals were required to be indigenous self-declared, following the criteria 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.

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) software 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
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 covering 72,541 SNPs when merged our 140 samples with the previously published data from Human Origin Dataset and 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).
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 f3-statistics of the form f3 (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-f3 statistics in the form of f3 (X, Y; Target) for all pairs of reference populations to make an evaluation of the possible admixture signals for the target populations. We conducted the heatmap visualization of the outgroup-f3 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 f3-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 (Hellerthal et al., 2014) analysis for our studied groups to obtain haplotype-sharing
based evidence of admixture. Using these haplotypes from SHAPEIT2, the “chunk length” output was obtained by running ChromoPainterv2 across all chromosomes. 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.
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. Notably, 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 coancestry 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
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 $f_3$ 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.
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 they provided 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_Afanasiyevo, Mongolia_N_East, and Russia_MLBA_Sintashta.

### Table 1 | Admixture $f_3$ statistics of the form (Source1, Source2; Target) with the lowest $f_3$ 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       | Shinao_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 | 169864 |
| Russia_Alan.SG             | WLR_LN                 | Dongxiang    | -0.014067 | -0.01139 | -12.348 | 138809 |
| CHB.SG                    | Anatolia_N             | Dongxiang    | -0.013755 | -0.00316 | -43.553 | 172663 |
| CHB.SG                    | Russia_MLBA_Sintashta  | Dongxiang    | -0.013573 | 0.000324 | -41.928 | 171061 |
| Anatolia_N                | YR_LBIA                | Dongxiang    | -0.013525 | -0.00062 | -21.527 | 170054 |
| Kazakhstan_Kangju.SG       | Shinao_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       | Shinao_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  | 29813  |
| Russia_MLBA_Sintashta     | Miaozigou_MN           | Yugur        | -0.008676 | 0.001367 | -6.346  | 49569  |
| Kazakhstan_Kangju.SG       | Shinao_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       | Shinao_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             | Shinao_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.
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_Andro novo 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 f3 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 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, D1a1alala2a~−, O2a2b1a2a1a2, 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
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 matrilineral 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 population 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.
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 is one 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 prevailing 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...
The true admixture happened not immediately after arrival or multiple times 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.

REFERENCES

Abelson, A. (1978). Population Structure in the Western Pyrenees: Social Class, Migration and the Frequency of Consangunous Marriage, 1850 to 1910. Ann. Hum. Biol. 5, 165–178. doi:10.1080/03014467800002761
Alexander, D. H., November, 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/s41368-020-0796-9
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

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.
Genetic Continuity from the Late Neolithic and Additional Genetic Affinities with the Steppe Populations. *Am. J. Phys. Anthropol.* 172, 176–188. doi:10.1002/ampa.24097

Khaz, 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:10.2669/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., Teller, 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* 173, 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 Tuja Population via 19 X-STRA. *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

Malychuk, B. A., Derenko, M., Denisova, G., Wozniak, M., Rogalla, U., Dambueva, I., et al. (2016). Y Chromosome Haploype 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., Kossol, 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. Natl. Acad. Sci. USA* 92, 258–267.

Medugorac, L. Kustermann, W., Lazar, P., Russ, L., and Pincher, F. (1994). Marker-derived Phylogeny of European Cattle Supports Demic Expansion of Agriculture. *Anim. Genet.* 25 (Suppl. 1), 19–27. doi:10.1111/1365-2052.1995.tb0399x

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., Bernardoz, 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-
