JSE Journal of Systematics and Evolution

doi: 10.1111/jse.12957

Research Article

Genetic admixture of Chinese Tajik people inferred from genome-wide array genotyping and mitochondrial genome sequencing

Jing Zhao[1,2†], Qiao Wu[1†], Xinhong Bai[3], Edward Allen[4], Mengge Wang[2], Guanglin He[2], Jianxin Guo[2], Xiaomin Yang[2], Jianxue Xiong[4,5], Zixi Jiang[4,5], Xiaoyan Ji[4,5], Hui Wang[4,5], Jingze Tan[1*], Shaoqing Wen[1,4,5*], and Chuan-Chao Wang[2,6,7,8*]

[1]Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Fudan University, Shanghai 200433, China
[2]Department of Anthropology and Ethnology, Institute of Anthropology, School of Sociology and Anthropology, Xiamen University, Xiamen 361005, China
[3]The Shanghai Anthropological Association, Shanghai 200433, China
[4]Institute of Archaeological Science, Fudan University, Shanghai 200433, China
[5]Center for the Belt and Road Archaeology and Ancient Civilizations (BRAAC), Fudan University, Shanghai 200433, China
[6]State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen 361102, China
[7]State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen 361102, China
[8]Institute of Artificial Intelligence, Xiamen University, Xiamen, China
[†]These authors contributed equally to this work.
*Authors for correspondence. Jingze Tan. E-mail: jztan@fudan.edu.cn; Shaoqing Wen. E-mail: wenshaoqing@fudan.edu.cn; Chuan-Chao Wang. E-mail: wang@xmu.edu.cn
Received 26 August 2022; Accepted 21 April 2023; Article first published online xxxx

Abstract Chinese Tajiks are an Indo-Iranian-speaking population in Xinjiang, northwest China. Although the complex demographic history has been characterized, the ancestral sources and genetic admixture of Indo-Iranian-speaking groups in this region remain poorly understood. We here provide the genome-wide genotyping data for over 700 000 single-nucleotide polymorphisms (SNPs) and mtDNA multiplex sequencing data in 64 Chinese male Tajik individuals from two dialect groups, Wakhi and Selekur. We applied principal component analysis (PCA), ADMIXTURE, *f*-statistics, treemix, *qpWave*/*qpAdm*, Admixture-induced Linkage Disequilibrium for Evolutionary Relationships (ALDER), and *Fst* analyses to infer a fine-scale population genetic structure and admixture history. Our results reveal that Chinese Tajiks showed the closest affinity and similar genetic admixture pattern with ancient Xinjiang populations, especially Xinjiang samples in the historical era. Chinese Tajiks also have gene flow from European and Neolithic Iran farmers-related populations. We observed a genetic substructure in the two Tajik dialect groups. The Selekur-speaking group who lived in the county had more gene flow from East Asians than Wakhi-speaking people who inhabited the village. These results document the population movements contributed to the influx of diverse ancestries in the Xinjiang region.

Key words: Chinese Tajiks, East Asia, genetic structure, population admixture, population history.

1 Introduction

Known as a "crossroads of civilizations," the history of East–West communications in Central Asia stretches back for millennia (Barthold, 1962; Cilli et al., 2019). This vast area contains a high linguistic, genetic, and ethnic diversity (Comas et al., 1998; Chaix et al., 2007), indicating a complex network of settlement and occupation. Among all the Central Asian ethnic groups, the Tajiks are considered an ideal choice to study broader patterns in Central Asian history. With a global population of 15–20 million, the Tajiks are one of the ancient indigenous peoples of Central Asia. Tajik people live mainly in Tajikistan, Afghanistan, Uzbekistan, and China's Xinjiang Uygur Autonomous Region. The population is generally divided into highland Tajiks and lowland Tajiks (Li & Bi, 2015; Palstra et al., 2015). The highland Tajiks mostly occupy the Hindu Kush Mountains and Pamir Plateau regions. At the same time, lowland Tajiks reside in the Central Asia plains, including the urban centers of Herat, Balkh, Bukhara, and Samarkand. Highland Tajiks primarily engage in animal husbandry, while their advanced handicraft industry and commerce mark lowland Tajiks in agricultural production. Both the highland and lowland Tajiks languages belong to the Indo-Iranian group of the Indo-European language family.

Chinese Tajiks belong primarily to the Tajik highland subdivision and are among 56 officially recognized ethnic

groups in China. The most recent census data shows China's present-day Tajik population is around 51 000. The Chinese Tajik language has absorbed many Uyghur and Chinese vocabulary due to long-term interaction between Chinese Tajik and surrounding ethnic groups (Li & Bi, 2015). About 80% of the Chinese Tajiks have been residents in the Tashkurgan Tajik Autonomous County (average altitude >4000 m) for more than three generations. Tashkurgan sits on the highest elevation of the Pamir Plateau. The "father of ice peaks," Mustagh Peak, stands north of Tashkurgan, while K2, the second-highest mountain on earth, straddles the China–Pakistan border to the south. This unique natural setting has allowed the Tajiks of Tashkurgan to retain numerous indigenous linguistic and cultural attributes and largely prohibited frequent interethnic marriages (Gao, 1996; Zeng, 2005; Yan et al., 2006; Liu et al., 2010; Malyarchuk et al., 2013; Khitrinskaia et al., 2014; Li et al., 2014; Li & Bi, 2015; Palstra et al., 2015; Peng et al., 2018).

Many previous studies have suggested extensive genetic admixture between East and West Eurasians in northwest China (Ning et al., 2019; Wang et al., 2019, 2021; Adnan et al., 2020, 2021; Wen et al., 2020; Zhao et al., 2020; He et al., 2021; Ma et al., 2021; Yang et al., 2021; Yao et al., 2021; Zhang et al., 2021), but a few focus on Tajiks. A 2018 paper by Min-Sheng Peng et al. used mitochondrial genome analysis to understand the maternal origin of multiple ethnic groups on the Pamir Plateau. Their study pointed to substantial genetic differentiation among different highland Tajik populations and the complex history behind the peopling of the Pamirs (Peng et al., 2018). However, current studies on the Tajiks have mainly focused on the Central Asian Tajiks. The genetic structure and population admixture of Chinese Tajiks largely remain unknown. Therefore, we carried out this research on the Tashkurgan Tajik population in China to shed more light on the genetic diversity and admixture pattern of the Tajik community and Chinese Tajiks.

2 Material and Methods

2.1 Sample collection

We collected saliva samples from 64 male Tajiks from the Taxkorgan Tajik Autonomous County in Xinjiang Uygur Autonomous Region, northwest China. These samples included two dialect groups, 21 Wakhi-speaking (living exclusively in Darbudar village) and 43 Selekur-speaking individuals. Detailed sample information is listed in Table S1. The geographical distribution of sampling sites is marked by the black triangle in the red area in Fig. 1. We collected the samples randomly from unrelated participants whose parents and grandparents had married in a nonconsanguineous fashion within the same ethnic group for at least three generations. Our study and sample collection were reviewed and approved by the Medical Ethics Committee of Xiamen University (Approval Number: XDYX2019009) and were in accordance with the recommendations provided by the revised Helsinki Declaration of 2000. The participants provided their written informed consent before they were invited to have participated in this study.

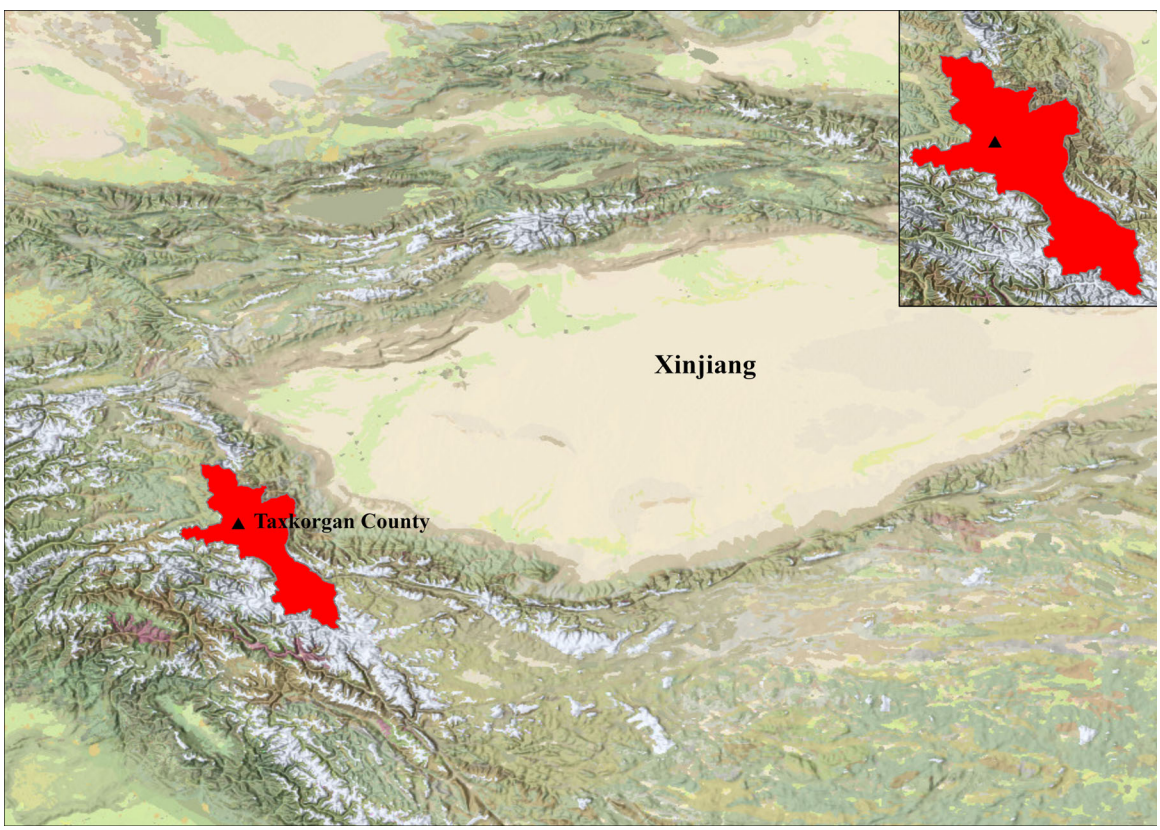

Fig. 1. Geographical distribution of sampling sites marked by the black triangle in the red area.

2.2 Data preparation

The genomic DNA of 64 samples was extracted using LD664 or LD607 Kits (Enlighten Biotech,, Shanghai, China). Genotyping was performed on the Affymetrix Inc. 23MF_v2 high-density SNP arrays at the 23 Mofang Laboratory, Chengdu. Filtering strategies were conducted with parameters: -hwe 1e-6, -geno 0.05, -mind 0.05 to obtain the quality-controlled raw data via the Plink 1.9 (Purcell et al., 2007). Finally, we had 705 767 SNPs for the subsequent analysis. We merged our newly generated data with publicly available data (Human Origin array and 1240K data set) curated by the David Reich lab (Patterson et al., 2012, Lazaridis et al., 2014; 1000 Genomes Project Consortium et al., 2015; Mallick et al., 2016) based on their overlapping SNPs to perform the corresponding population genomic analyses. The link can find the detailed information for the curated data set: https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data. The list of the modern and ancient populations co-analyzed with new data is in Table S10. The mtDNA libraries were prepared by using an MtDNA Library Preparation Kit 2.0 (Enlighten Biotech) and a WhoChrMT kit (Enlighten Biotech). The optimized multiplex polymerase chain reaction (PCR) amplification reaction conditions were according to the published reference (Wang et al., 2022).

2.3 Analysis of genetic relationship and population structure

We performed principal component analysis (PCA) at the individual level using smartpca, part of the EIGENSOFT package (Patterson et al., 2006). We employed the default parameters where two additional parameters were used: numoutlieriter: 0 and lsqproject: YES. Ancient reference populations were projected onto the two-dimensional plots based on their genetic variations to modern people. ADMIXTURE (Alexander et al., 2009) analysis was carried out after pruning SNPs exhibiting strong linkage disequilibrium using PLINK tools with the parameters "-indep-pairwise 200 25 0.4." There were 77 503 SNPs left after LD pruning. We then ran the unsupervised ADMIXTURE with a *K* value (number of assumed ancestral components) ranging from 2 to 10, using 100 bootstrap iterations with different random seeds. Qp3pop package in ADMIXTOOLS (Patterson et al., 2012) was used to carry out outgroup f_3-statistics (Source1, Source2; Mbuti) and admixture f_3-statistics (Source1, Source2; Target population). The outgroup f_3-statistics were used as an index for evaluating the shared genetic drift, and the admixture f_3-statistics were used for exploring admixture signatures with different ancestral sources. QpDstat package in ADMIXTOOLS (Patterson et al., 2012) with f_4 Mode (f_4: YES) was used to explore the derived-allele sharing relative to their reference pairs.

2.4 Streams of ancestry and inference of admixture proportions

We used *qpAdm* and *qpWave* (Haak et al., 2015) as implemented in *ADMIXTOOLS* to determine the number of ancestry sources and estimate the admixture proportions of Chinese Tajiks. Outgroups were listed in Table S5A and B, which had no recent genetic admixture and were differentially related to the ancestral sources of Chinese Tajik people.

2.5 Uniparental haplogroup assignment and mtDNA multiplex sequencing analysis

We assigned Y chromosomal haplogroups using the captured SNPs on the Y chromosome. We determined haplogroups by identifying the most derived allele upstream and the most ancestral allele downstream in the phylogenetic tree in the ISOGG version 11.89 (http://www.isogg.org/tree/). The mtDNA haplogroup assignment was performed using the online tools available with HaploGrep version 2.4.0 software (Kloss-Brandstätter et al., 2011).

MtDNA multiplex sequencing data were aligned and assembled in contigs, and consensus sequences were compared to the revised Cambridge Reference Sequence (Andrews et al., 1999). The quality of the sequences was examined manually, and two analysts independently annotated deviations from the reference sequence. For all analyses, cytosine (C) insertions/deletions at positions 16 193, 309, and 315, as well as adenine (A) to C transversions at positions 16 182 and 16 183, were ignored.

2.6 Genetic distance (*Fst*) analysis

Genetic distance was calculated according to Tajima–Nei (Tajima, 1989). The pairwise *Fst* was computed using Arlequin version 3.5.1.2 software (Excoffier et al., 2005) and is listed in Table S9. We used R Statistical Software v3.0.2 to plot the heatmaps based on the population pairwise *Fst*.

2.7 Treemix and Admixture-induced Linkage Disequilibrium for Evolutionary Relationships (ALDER)

We used TreeMix (Pickrell & Pritchard, 2012) to study population splits and gene flow events. ALDER (Loh et al., 2013) was used to estimate the time of potential ancestral source admixture events with 28 years per generation.

3 Results

3.1 Genetic structure of Chinese Tajik people

We merged our newly generated genome-wide data with the publicly available modern and ancient reference genomes. We obtained two different data sets, a low-density one with 94 755 overlapping SNPs, and a high-density one including 243 549 overlapping SNPs. We used these two data sets to reconstruct the Chinese Tajiks' population history.

We explored the patterns of genetic relationship among Chinese Tajiks and Central Asian, Western Eurasian, ancient Xinjiang and East Asian populations via PCA and unsupervised model-based ADMIXTURE clustering. Ancient individuals were projected onto the modern genetic landscape. In Fig. 2A, we observed three main genetic clines, which were in accordance with the geographical and linguistic divisions. These included the South China cline extending from Hmong Mien to Austronesian and ancient southern populations, the Central Asian/West Eurasian cline associated with not only Central Asian and Western Eurasian populations but also ancient Xinjiang groups, Sinitic/Tibetan Burman cline including Sinitic and Tibetan Burman populations as well as Yellow River farmers. We observed that Chinese Tajiks were located in the Central Asian/West Eurasian cline and separated from East Asian-related clusters. When focused on the Central Asian/West Eurasian cline (Fig. 2B), Chinese

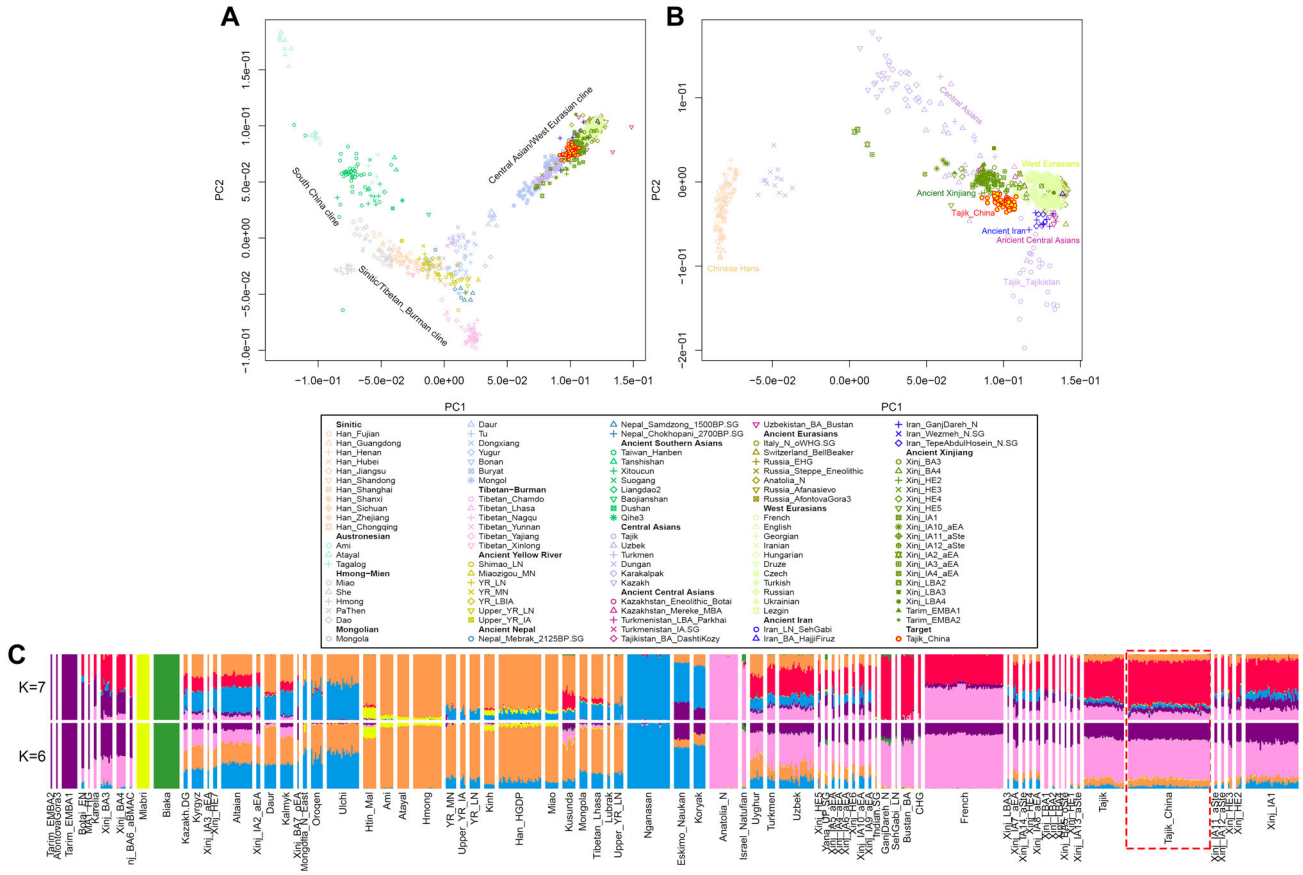

Fig. 2. Overview of genetic structure. **A**, Principal component analysis (PCA) of Chinese Tajiks with Central Asian, Western Eurasian, ancient Xinjiang and East Asian populations. Ancient individuals were projected onto the modern genetic landscape. **B**, PCA analysis focused on Central Asian and Western Eurasian populations. **C**, ADMIXTURE analysis of newly generated Chinese Tajik data (Tajik_China) together with worldwide representative modern and ancient non-Africans.

Tajiks clustered next to ancient Xinjiang people and were close to Western Eurasian and Central Asian populations. In addition, Chinese Tajiks did not cluster together with Tajiks in Tajikistan, which suggested there was a genetic substructure between Tajiks from different locations.

We ran the unsupervised model-based ADMIXTURE with 716 individuals from 78 modern and ancient non-African populations. We found Chinese Tajiks had more Western Eurasian-related ancestral components (blue and red bands) as shown in Fig. S1. Under the best fitting model at $K = 6$ with the smallest cross-validation errors, we observed Chinese Tajiks had most of the light purple ancestral components (Fig. 2C), which was maximized in the Western Eurasian group (Anatolia_N). Additionally, when $K = 7$, we noticed that Chinese Tajiks had a large proportion of red ancestral components, which were maximized in the Neolithic Iran farmers, Uzbekistan_BA_Bustan and Caucasus hunter-gatherers (CHG). Other ancestral components ($K = 6$) aligned with Tarim_EMBA1 (deep purple) and East Asian (orange) only accounted for a small fraction. In general, Chinese Tajiks had the most significant Western Eurasian-related ancestral components, followed by a small amount of East Asian-related ancestry, the patterns of which were similar to ancient Xinjiang populations and Tajiks in Tajikistan.

3.2 Population continuity and admixture in Chinese Tajik people

We used the ancient and modern reference populations from the Human Origin data set (Patterson et al., 2012; Lazaridis et al., 2014) as the plausible source proxies to perform the f_3- and f_4-statistics. There were a total of 94 755 SNPs left after data merging. We first conducted the outgroup f_3 (Mbuti; Y_facet, Tajik_China) to further confirm the genetic affinity between Chinese Tajiks and Eurasian populations. We showed populations with the top 40 f_3 values in Fig. 3A and listed the more comprehensive results in Table S2. We observed that Chinese Tajiks shared more genetic drift with Early-Middle Bronze Age Tarim people (Tarim_EMBA1), other ancient Xinjiang groups, Central Asians and Steppe populations.

We then explored the admixture signals (Z-scores < -3) in Chinese Tajiks using admixture-f_3 (Source1, Source2; Target). The negative f_3 values mean the allele frequencies are intermediate between two potential source populations. We listed all the potential related sources with the most significant negative f_3 values in Table S3. Most of the statistically significant source candidates were pairs from one of the Western populations, while the other was from East Asians. The maximum negative f_3-value was observed when French was the Western Eurasian ancestral source and

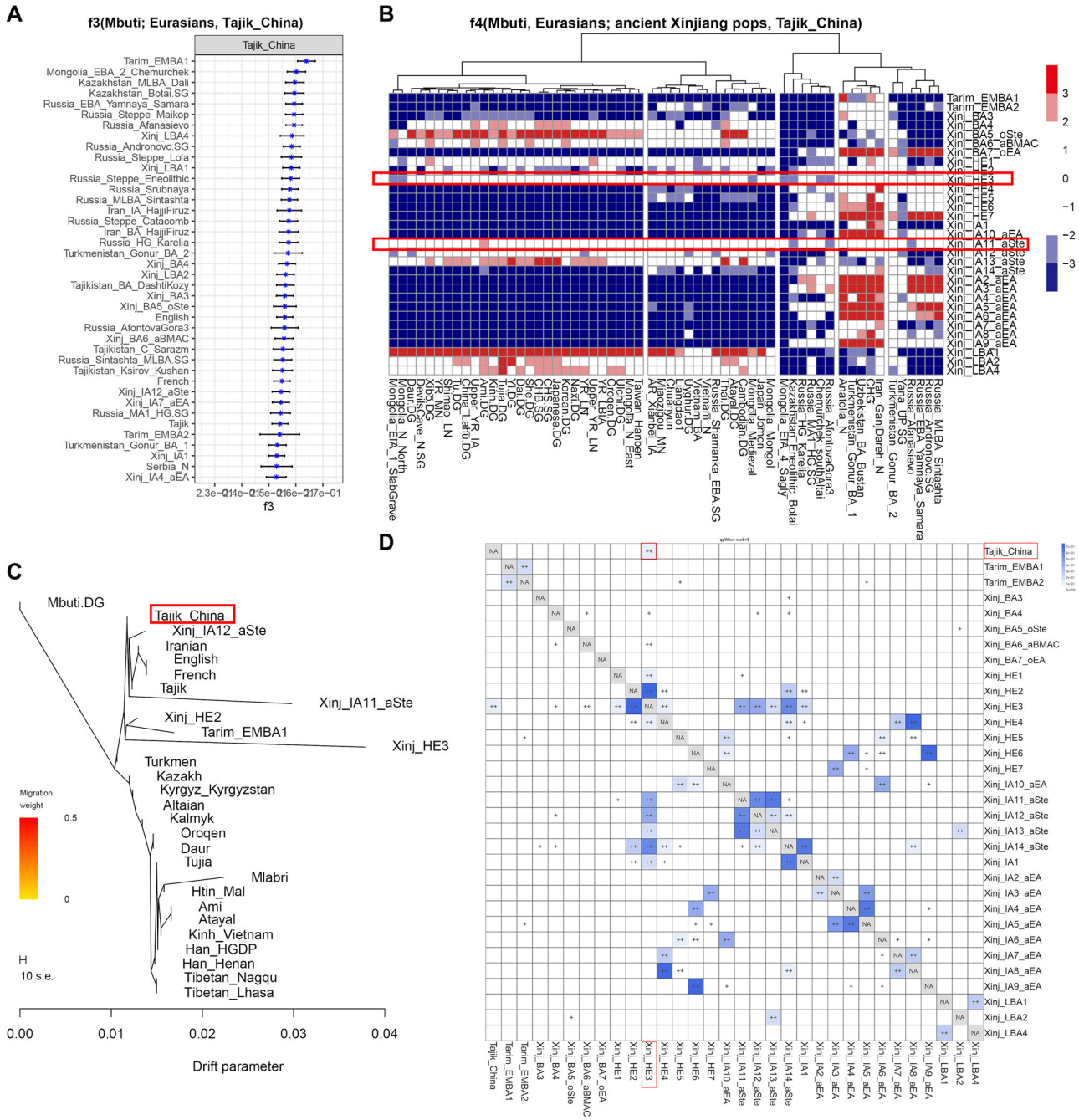

Fig. 3. Signals of admixed sources and the shared genetic drifts revealed from the outgroup-f_3 statistics, f_4-statistics, TreeMix and *qpWave* analysis. **A,** Top admixture signals for ancient and modern Eurasians as possible sources via outgroup-f_3 (Mbuti; Eurasians, Tajik_China). **B,** Investigation of the relatedness between Chinese Tajiks and ancient Xinjiang populations in f_4-statistics in the form of f_4 (Mbuti, reference populations; ancient Xinjiang populations, Tajik_China). **C,** TreeMix analysis to further explore the potential admixture. **D,** Test of the homogeneity between Chinese Tajiks and ancient Xinjiang populations by *qpWave* analysis.

Han_HGDP was the East Asian source, suggesting that Chinese Tajiks had the East–West admixture signals. Besides, the result from the negative f_3-values in f_3 (Xinj_LBA1, Han_NChina; Tajik_China) also suggested the impactful gene flow from the local Xinjiang and Chinese Han to the Chinese Tajiks.

In addition, motivated by the PCA, ADMIXTURE, and f_3-statistics results, Chinese Tajiks showed a close affinity with ancient Xinjiang populations, so we performed f_4-statistics in the form of f_4(Mbuti, reference populations; ancient Xinjiang populations, Tajik_China) to investigate the relatedness between them. In Fig. 3B, we observed that most ancient

Xinjiang populations presented significant negative values (Z-scores < −3) marked with blue color, which suggested that these ancient Xinjiang people had more gene flow from East Asians than Chinese Tajiks. The detailed results were listed in Table S4. Furthermore, there were no significant values (−3 < Z-scores < 3) between Chinese Tajiks and Xinj_HE3 and Xinj_IA11_aSte, suggesting that Chinese Tajiks had a close relationship with these two ancient Xinjiang groups.

Consistent with f_3 statistics, potential admixture was further demonstrated by the allele frequency-based TreeMix analysis. Among 28 populations in Fig. 3C, Chinese Tajiks clustered with ancient Xinjiang, Iranian, Tajikistan Tajiks, French, and English populations while separated from East Asian-related clades, suggesting that Chinese Tajiks showed a closer affinity with ancient Xinjiang and Western populations.

To further test the homogeneity between Chinese Tajiks and ancient Xinjiang populations, we performed a *qpWave* analysis using Mbuti, Ami, Anatolia_N, Iran_GanjDareh_N, Israel_Natufian, Italy_North_Villabruna_HG, Mixe, Onge, Russia_Kostenki14, Tianyuan, and Ust_Ishim as outgroups. The results (Fig. 3D) showed that a significant *P* value for rank = 0 ($P > 0.05$, indicated with double+) was presented between Chinese Tajiks and Xinj_HE3, suggesting those two groups were genetically homologous, which was also demonstrated in f_4-statistics above.

3.3 Genetic substructure between Tajik dialect groups

There are Wakhi and Selekur-speaking groups in Chinese Tajiks. We performed a series of analyses to explore the possible genetic substructure between these two groups. From the PCA and ADMIXTURE perspectives, we have not observed obvious genetic differences at the individual level (Figs. S2a, S2b). However, we found the significant negative values (Z-scores < −3) marked with blue color when we performed f_4-statistics in the form of f_4 (Mbuti, reference populations; Tajik_Selekur, Tajik_Wakhi), which suggested that Tajik_Selekur had more gene flow from East Asians than Tajik_Wakhi (Fig. S2c). To further confirm the genetic difference between Tajik_Selekur and Tajik_Wakhi, we conducted *qpWave* analysis together with relevant reference populations. There was no significant *P* value ($P > 0.05$, indicated with double+) between Tajik_Selekur and Tajik_Wakhi shown in Fig. S2d, which is consistent with the above f_4-statistics.

3.4 Genetic sources of Chinese Tajik people

To further investigate potential ancestral sources and genetic variations of Chinese Tajiks, we modeled the well-fitted two- and three-way admixture models in *qpAdm*. We used the published ancient Xinjiang populations (Kumar et al. 2022), millet farmers in the Yellow River basin, western and eastern Steppe populations and Neolithic Iran farmers as the potential sources to explore genetic admixture patterns of Chinese Tajiks (Figs. 4A–4J; Table S5). In the two-way admixture models (Figs. 4A–4C), Chinese Tajiks could be modeled as an admixture of Xinjiang_HE3 (80.4%–85.9%) and Turkmenistan_Gonur_BA (19.6%) or Iran_GanjDareh_N (16.8%) or CHG (14.1%). When we used other ancient Xinjiang groups (Xinj_IA7_aEA and Xinj_HE5) as sources, the admixture proportions were down to 56.1%~61.5% (Figs. 4D, 4E), indicating a closer affinity of Chinese Tajiks with Xinjiang_HE3 than with other ancient Xinjiang populations, which was also attested in f_4-statistics (Fig. 3B) and *qpWave* analysis (Fig. 3D). Moreover, in the three-way admixture models (Figs. 4F–4J), Chinese Tajiks could be modeled as ancient Xinjiang groups with additional ancestry from Turkmenistan_Gonur_BA, Russia_Andronovo, Russia_MLBA_Sintashta or Iran_GanjDareh_N. We also observed that the admixture proportions from Yellow River farmers (YR_MN) were less than 9% (Fig. 4H) in Chinese Tajiks.

When we focused on Wakhi and Selekur-speaking groups of Chinese Tajiks separately (Fig. S3a), Tajik_Selekur could be successfully modeled as the ancient Xinjiang population (Xinj_IA11_aSte, 93.6%–93.9%) and Yellow River farmers (6.4%) or Chinese Han (6.1%). However, Tajik_Wakhi failed in the two-source admixture models. In the three-source admixture models, there were slightly different between Tajik_Selekur and Tajik_Wakhi, indicating that Tajik_Selekur had a little more gene flow from Yellow River farmers than Tajik_Wakhi with a proportion of 5.3% and 3% respectively.

3.5 Admixture time estimation via the decay of linkage disequilibrium

We used different eastern and western ancestral sources to assess their possibilities of the admixture process in the formation of Chinese Tajik people by estimating ALDER-based admixture time. Considering the observed genetic structure, we used East Asians, Central Asians, Iranian and Western Eurasians as Eastern and Western ancestral proxies (Table S6). We obtained 22 pairs of best-fitted models after excluding failed models or models with lower Z-scores/*P*-values. Chinese Tajiks could be modeled as the admixture between ancient Xinjiang populations around 577.08 ± 187.32 to 1615.6 ± 511.56 years ago (ya). Chinese Tajiks also could be modeled as the admixture of ancient Xinjiang populations and Chinese Han around 695.8 ± 73.64 to 1141.28 ± 110.04 ya. In addition, we observed that Chinese Tajiks also could be modeled as a mixed group from ancient Xinjiang and European groups (Anatolia_N and CHG) from 1034.6 ± 194.04 to 2954.28 ± 509.32 @@ya, as well as Kyrgyz_Tajikistan (720.72 ± 196.28 ya) and Iranian (~551.04 ± 93.8 ya). When we used Anatolia_N and Tarim_EMBA1 as the two sources, the estimated admixture time became older (~4720.52 ± 595 ya).

3.6 Y chromosomal haplogroup analysis

Chinese Tajiks exhibited high diversity in their paternal gene pool (Fig. 4k; Table S7). A high percentage of Y-chromosome haplogroup R1a1-M17 was found among Chinese Tajiks, accounting for 53.13% of the total. In previous studies, the R1a1-M17 primarily appeared in high frequencies throughout western Xinjiang groups in Iron Age and Historical Era and Steppe_MLBA populations (Wells et al., 2001; Sharma et al., 2009; Kumar et al., 2022). In addition, the derived branch R1a1a1b2a2-F2935 is impressively enriched among the Bashkirs (Karmin et al., 2015), accounting for 14.06%, as shown in Table S7. The Bashkirs belong to one of the indigenous Russian groups speaking various Turkic languages. The other derived branch, R1a1a1b2a1-L657.1, accounting for 7.81%, has a large proportion among Uyghur and Kazakh populations (Bergström et al., 2020). Moreover, haplogroup R2a1-L263 is widely distributed among the

Fig. 4. Potential ancestral sources and paternal and maternal lineages inferred from the pairwise *qpAdm*, Y-chromosome and mitochondrial DNA haplogroups. **A**–**J**, Potential ancestral sources and genetic variations of Chinese Tajiks with well-fitted two- and three-way admixture models in *qpAdm*. **K**, Y-chromosome haplogroups distributed in the Chinese Tajik lineages. **L**, MtDNA haplogroups distributed in the Chinese Tajik gene pool.

Burusho living in the mountains of northern Pakistan. Ethnographic studies have previously found similarities between Burusho and Plateau Tajiks regarding appearance, dress, customs, and music (Bergström et al., 2020). Besides, the minor haplogroup J-M304, accounting for 12.5%, was suggested to being evolved in western Asia and at high frequencies in the Middle East, North Africa, the Horn of Africa, and Caucasus (Semino et al., 1996, 2000; Quintana-Murci et al., 2001). Haplogroup Q1b-M346 (6.25%) is mainly distributed among the Hazara of Afghanistan (Bergström et al., 2020), a group formed by mixing European and Mongolian ancestry. Haplogroup C2a-L1373, D-CTS3946, O2a-M324, and their subclades are common among East Asian populations (Bergström et al., 2020; Wang & Li, 2013; Yan et al., 2014) but at tiny proportions in Chinese Tajiks. In this study's Selekur- and Wakhi-speaking groups, we found a low frequency of the haplogroups associated with East Asian populations, such as C2a1a3a1-FGC16594, D-CTS3946 and O2a-M324.

From the perspective of Y-chromosome haplogroups above, we also observed the East–West admixture signals in Chinese Tajiks. The frequent haplogroup R1a1 was also predominant in western Xinjiang groups in Iron Age and Historical Era as well as Steppe_MLBA populations. However, the typical East Asian-related haplogroups were at low frequencies in Chinese Tajiks. Additionally, there was no significant difference in the paternal genetic profile between Selekur- and Wakhi-speaking groups of Chinese Tajiks.

3.7 MtDNA haplogroup analysis

There were 44 diverse maternal haplotypes determined within Chinese Tajik individuals using HaploGrep software (Kloss-Brandstätter et al., 2011). The results are shown in Fig. 4L and Table S8. We observed that the predominant maternal gene pool of Chinese Tajiks (U, H, and T) was also at high frequencies in ancient Xinjiang populations from Bronze Age to Historical Era (Kumar et al., 2022) and Western Eurasians (Richards et al., 1998, 2000; Bramanti et al., 2009; Bollongino et al., 2013; Brotherton et al., 2013; Fu et al., 2013; Davidovic et al., 2017). East Asian-related haplogroups (van Oven & Kayser, 2009) accounted for 9.4% of Chinese Tajik with haplogroups G (4.6%, 3/64), C (3.1%, 2/64), and A (1.6%, 1/64), which were only found in the Selekur-speaking population and consistent with the observation of f_4-statistics in the form of f_4 (Mbuti, reference populations; Tajik_Selekur, Tajik_Wakhi). No characteristic maternal lineages associated with East Asians appeared in the Wakhi-speaking population.

Motivated by investigating maternal lineages from the genome-wide array genotyping data above, we subsequently performed the mitochondrial multiplex PCR amplification. We used Arlequin software (Excoffier et al., 2005) to calculate the genetic distance (*Fst*) based on the mtDNA multiplex sequencing data. We compared the two dialect groups of Chinese Tajiks (Selekur and Wakhi dialects) with relevant populations from East Asia, Central Asia, and Eurasia. A heatmap based on the *Fst* values was shown in Fig. S3b, and detailed information was listed in Table S9. In Fig. S3b, the blue color indicated a close genetic distance, while the red showed a distant genetic distance. We found that both Selekur- and Wakhi-speaking groups showed a close genetic distance from each other and with published Tajik people in Taxkorgan and Pamiri, which was consistent with geographical locations. These two Chinese Tajik groups also had a close affinity with Israel and modern Iranian groups (marked by the blue wireframe).

4 Discussion

In this study, we analyzed genomic data of Chinese Tajiks with modern and ancient relevant Eurasian populations to trace the population genetic history. We found Chinese Tajiks contained two major ancestries, one was from ancient Xinjiang populations (Kumar et al., 2022), and the other was related to Western Eurasians probably from Europeans and Neolithic Iran farmers. The finding further supports the movement and admixture of Steppe, Central Asian, and East Asian people into the Xinjiang region increased in the Bronze Age and were still prevalent in both the historical period (HE) and present-day Xinjiang populations (Kumar et al., 2022). Specifically, Chinese Tajiks showed a large-scale genetic continuity with some ancient Xinjiang populations in historical periods (e.g., Xinj_HE3; Figs. 3B, 3D). The proportions of Xinj_HE3-related ancestry reached up to 80.4%–85.9% in Chinese Tajiks (Figs. 4A–4C), consistent with the admixture time evaluation that the admixture events between Chinese Tajiks and ancient Xinjiang populations occurred within the historical period (Table S6).

The genome-wide research on the Chinese Tajiks helps us understand the Chinese Tajik population's genetic diversity and admixture patterns. It provides clues for further exploring population dynamics in western China, which is in the exchange of material culture, agriculture, and technology between the West and East Eurasian populations. Generally, the diffusion of culture is not always accompanied by population movements (Posth et al., 2018). However, our findings give genomic evidence for introducing the East-Iranian language, one of the Indo-European languages, along with the population movements and genetic admixture. The Selekur- and Wakhi-speaking groups of Chinese Tajiks are major Indo-Iranian-speaking populations in Xinjiang. Although these two groups showed a great affinity with the local people in Xinjiang (Figs. 2–4), there are slight differences between them. The Selekur-speaking inhabitants in Taxkorgan County of Xinjiang were characterized by slightly more gene flow and admixture from East Asian populations than Wakhi-speaking inhabitants in Darbudar Village of Taxkorgan County (Fig. S2c). This inference was also supported by maternal haplogroup analysis that these East Asian-related haplogroups were only found in the Selekur-speaking group and did not appear in Wakhi-speaking people. This suggested that the closed-off rural lifestyle might reduce their contact with surrounding populations, while the broader demographic communications would aid in the population admixture.

In Xinjiang, there is a complex demographic history and the coexistence of populations with diverse cultural, linguistic, and genetic backgrounds. This study documented the dynamic interactions of Indo-Iranian languages in the Xinjiang region and uncovered similar genetic admixture patterns with surrounding populations. We observed admixed ancestries related to Xinjiang, European, Central and East Asian populations and Neolithic Iran farmers in present-day Chinese Tajik people, suggesting the formation of Tajik was accompanied by widespread population movements. These inferences have provided important information further to understand the Indo-Iranian languages in the Xinjiang region. However, the absence of female samples in our sample collection may result in potentially underestimating the X-chromosome genetic diversity. Further sampling of female populations will be necessary to characterize the formation of Chinese Tajiks' genetic makeup.

Acknowledgements

We thank all the volunteers and the local guide, Baohua Zhu, for the sample collection. S. Fang and Z. Xu from the Information and Network Center of Xiamen University are acknowledged for their help with high-performance computing. This work was funded 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, 32070576, 32270667, 31801040, and 32111530227), Major Research Program of National Natural Science Foundation of China (91731303), the Major Project of National Social Science Foundation of China granted to Chuan-Chao

Wang (21&ZD285), Xiaohua Deng (20&ZD248), and Shaoqing Wen (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), Nanqiang Outstanding Young Talents Program of Xiamen University (X2123302), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), the 111 Project (B13016), and European Research Council (ERC) grant (ERC-2019-ADG-883700-TRAM).

Author Contributions

Jing Zhao, Shaoqing Wen, and Chuan-Chao Wang conceived the idea for the study. Qiao Wu, Xinhong Bai, Jianxue Xiong, Zixi Jiang, Xiaoyan Ji, Hui Wang, Jingze Tan, and Shaoqing Wen performed or supervised the wet laboratory work. Jing Zhao, Mengge Wang, Guanglin He, Jianxin Guo, Xiaomin Yang, and Chuan-Chao Wang analyzed the data. Jing Zhao, Qiao Wu, Edward Alle, Shaoqing Wen, and Chuan-Chao Wang wrote and edited the manuscript. All authors contributed to the article and approved the submitted version.

Conflict of Interest

The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

Data Availability Statement

The variation data reported in this paper had been deposited in the Zenodo: https://doi.org/10.5281/zenodo.5872235.

Ethics Statement

Research involving human participants was reviewed and approved by the Medical Ethics Committee of Fudan University and Xiamen University (Approval Number: XDYX2019009) and followed the recommendations provided by the revised Helsinki Declaration of 2000. The participants provided their written informed consent to participate in this study.

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. *Nature* 526(7571): 68–74.

Adnan A, Anwar A, Simayijiang H, Farrukh N, Hadi S, Wang CC, Xuan JF. 2021. The heart of silk road "Xinjiang" its genetic portray and forensic parameters inferred from autosomal STRs. *Frontiers in Genetics* 12: 760760.

Adnan A, He G, Rakha A, Kasimu K, Guo J, Hassan SE, Hadi S, Wang CC, Xuan JF. 2020. Phylogenetic relationship and genetic history of Central Asian Kazakhs inferred from Y-chromosome and autosomal variations. *Molecular Genetics and Genomics* 295: 221–231.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19: 1655–1664.

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge reference. *Nature Genetics* 23: 147.

Barthold VV. 1962. *Four studies on the history of Central Asia*. Leiden: Brill.

Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J, Blanché H, Deleuze JF, Cann H, Mallick S, Reich D, Sandhu MS, Skoglund P, Scally A, Xue Y, Durbin R, Tyler-Smith C. 2020. Insights into human genetic variation and population history from 929 diverse genomes. *Science* 367: aay5012.

Bollongino R, Nehlich O, Richards MP, Orschiedt J, Thomas MG, Sell C, Fajkosová Z, Powell A, Burger J. 2013. 2000 years of parallel societies in Stone Age Central Europe. *Science* 342: 479–481.

Bramanti B, Thomas MG, Haak W, Unterlaender M, Jores P, Tambets K, Antanaitis-Jacobs I, Haidle MN, Jankauskas R, Kind CJ, Lueth F, Terberger T, Hiller J, Matsumura S, Forster P, Burger J. 2009. Genetic discontinuity between local hunter-gatherers and central Europe′s first farmers. *Science* 326: 137–140.

Brotherton P, Haak W, Templeton J, Brandt G, Soubrier J, Jane Adler C, Richards SM, Der Sarkissian C, Ganslmeier R, Friederich S, Dresely V, van Oven M, Kenyon R, Van der Hoek MB, Korlach J, Luong K, Ho SYW, Quintana-Murci L, Behar DM, Meller H, Alt KW, Cooper A Genographic Consortium. 2013. Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. *Nature Communications* 4: 1764.

Chaix R, Quintana-Murci L, Hegay T, Hammer MF, Mobasher Z, Austerlitz F, Heyer E. 2007. From social to genetic structures in central Asia. *Current Biology* 17: 43–48.

Cilli E, Sarno S, Gnecchi Ruscone GA, Serventi P, De Fanti S, Delaini P, Ognibene P, Basello GP, Ravegnini G, Angelini S, Ferri G, Gentilini D, Di Blasio AM, Pelotti S, Pettener D, Sazzini M, Panaino A, Luiselli D, Gruppioni G. 2019. The genetic legacy of the Yaghnobis: A witness of an ancient Eurasian ancestry in the historically reshuffled central Asian gene pool. *American Journal of Physical Anthropology* 168: 717–728.

Comas D, Calafell F, Mateu E, Pérez-Lezaun A, Bosch E, Martínez-Arias R, Clarimon J, Facchini F, Fiori G, Luiselli D, Pettener D, Bertranpetit J. 1998. Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. *American Journal of Human Genetics* 63: 1824–1838.

Davidovic S, Malyarchuk B, Aleksic J, Derenko M, Topalovic V, Litvinov A, Skonieczna K, Rogalla U, Grzybowski T, Stevanovic M, Kovacevic-Grujicic N. 2017. Mitochondrial super-haplogroup U diversity in Serbians. *Annals Human Biology* 44: 408–418.

Excoffier L, Laval G, Schneider S. 2005. Arlequin (version 3,0): An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* 1: 47–50.

Fu Q, Mittnik A, Johnson PLF, Bos K, Lari M, Bollongino R, Sun C, Giemsch L, Schmitz R, Burger J, Ronchitelli AM, Martini F, Cremonesi RG, Svoboda J, Bauer P, Caramelli D, Castellano S, Reich D, Pääbo S, Krause J. 2013. A revised timescale for human evolution based on ancient mitochondrial genomes. *Current Biology* 23: 553–559.

Gao EQ. 1996. *Tajik-Chinese dictionary*. Chengdu: Sichuan Nationalities Publishing House.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, Fu Q, Mittnik A, Bánffy E, Economou C, Francken M, Friederich S, Pena RG,

Hallgren F, Khartanovich V, Khokhlov A, Kunst M, Kuznetsov P, Meller H, Mochalov O, Moiseyev V, Nicklisch N, Pichler SL, Risch R, Rojo Guerra MA, Roth C, Szécsényi-Nagy A, Wahl J, Meyer M, Krause J, Brown D, Anthony D, Cooper A, Alt KW, Reich D. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522: 207–211.

He G, Wang M, Zou X, Chen P, Wang Z, Liu Y, Yao H, Wei LH, Tang R, Wang CC, Yeh HY. 2021. Peopling history of the tibetan plateau and multiple waves of admixture of Tibetans Inferred from both ancient and modern genome-wide data. *Frontiers in Genetics* 12: 725243.

Karmin M, Saag L, Vicente M, Wilson Sayres MA, Järve M, Talas UG, Rootsi S, Ilumäe AM, Mägi R, Mitt M, Pagani L, Puurand T, Faltyskova Z, Clemente F, Cardona A, Metspalu E, Sahakyan H, Yunusbayev B, Hudjashov G, DeGiorgio M, Loogväli EL, Eichstaedt C, Eelmets M, Chaubey G, Tambets K, Litvinov S, Mormina M, Xue Y, Ayub Q, Zoraqi G, Korneliussen TS, Akhatova F, Lachance J, Tishkoff S, Momynaliev K, Ricaut FX, Kusuma P, Razafindrazaka H, Pierron D, Cox MP, Sultana GN, Willerslev R, Muller C, Westaway M, Lambert D, Skaro V, Kovačevic L, Turdikulova S, Dalimova D, Khusainova R, Trofimova N, Akhmetova V, Khidiyatova I, Lichman DV, Isakova J, Pocheshkhova E, Sabitov Z, Barashkov NA, Nymadawa P, Mihailov E, Seng JW, Evseeva I, Migliano AB, Abdullah S, Andriadze G, Primorac D, Atramentova L, Utevska O, Yepiskoposyan L, Marjanovic D, Kushniarevich A, Behar DM, Gilissen C, Vissers L, Veltman JA, Balanovska E, Derenko M, Malyarchuk B, Metspalu A, Fedorova S, Eriksson A, Manica A, Mendez FL, Karafet TM, Veeramah KR, Bradman N, Hammer MF, Osipova LP, Balanovsky O, Khusnutdinova EK, Johnsen K, Remm M, Thomas MG, Tyler-Smith C, Underhill PA, Willerslev E, Nielsen R, Metspalu M, Villems R, Kivisild T. 2015. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Research* 25: 459–466.

Khitrinskaia IY, Kharkov VN, Voevoda MI, Stepanov VA. 2014. Genetic diversity and relationships of northern eurasia populations for polymorphic Alu-insertions. *Molecular Biology* 48(1): 69–80.

Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F. 2011. HaploGrep: A fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Human Mutation* 32: 25–32.

Kumar V, Wang W, Zhang J, Wang Y, Ruan Q, Yu J, Wu X, Hu X, Wu X, Guo W, Wang B, Niyazi A, Lv E, Tang Z, Cao P, Liu F, Dai Q, Yang R, Feng X, Ping W, Zhang L, Zhang M, Hou W, Liu Y, Bennett EA, Fu Q. 2022. Bronze and Iron Age population movements underlie Xinjiang population history. *Science* 376: 62–69.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, Berger B, Economou C, Bollongino R, Fu Q, Bos KI, Nordenfelt S, Li H, de Filippo C, Prüfer K, Sawyer S, Posth C, Haak W, Hallgren F, Fornander E, Rohland N, Delsate D, Francken M, Guinet JM, Wahl J, Ayodo G, Babiker HA, Bailliet G, Balanovska E, Balanovsky O, Barrantes R, Bedoya G, Ben-Ami H, Bene J, Berrada F, Bravi CM, Brisighelli F, Busby GB, Cali F, Churnosov M, Cole DE, Corach D, Damba L, van Driem G, Dryomov S, Dugoujon JM, Fedorova SA, Gallego Romero I, Gubina M, Hammer M, Henn BM, Hervig T, Hodoglugil U, Jha AR, Karachanak-Yankova S, Khusainova R, Khusnutdinova E, Kittles R, Kivisild T, Klitz W, Kučinskas V, Kushniarevich A, Laredj L, Litvinov S, Loukidis T, Mahley RW, Melegh B, Metspalu E, Molina J, Mountain J, Näkkäläjärvi K, Nesheva D, Nyambo T, Osipova L, Parik J, Platonov F, Posukh O, Romano V, Rothhammer F, Rudan I, Ruizbakiev R, Sahakyan H, Sajantila A, Salas A, Starikovskaya EB, Tarekegn A, Toncheva D, Turdikulova S, Uktveryte I, Utevska O, Vasquez R, Villena M, Voevoda M, Winkler CA, Yepiskoposyan L, Zalloua P, Zemunik T, Cooper A, Capelli C, Thomas MG, Ruiz-Linares A, Tishkoff SA, Singh L, Thangaraj K, Villems R, Comas D, Sukernik R, Metspalu M, Meyer M, Eichler EE, Burger J, Slatkin M, Pääbo S, Kelso J, Reich D, Krause J. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513: 409–413.

Li D, Bi X. 2015. *Encyclopedia of Chinese ethnicities 14 Kazakh Kirgiz Tajik Tatar volume M*. Xi'an: World Book Publishing Xi'an Co, Ltd. 387–389.

Li J, Lou H, Yang X, Lu D, Li S, Jin L, Pan X, Yang W, Song M, Mamatyusupu D, Xu S. 2014. Genetic architectures of ADME genes in five Eurasian admixed populations and implications for drug safety and efficacy. *Journal of Medical Genetics* 51: 614–622.

Liu XL, Zhang FL, Zhou ZY, Zhao HL, Shen GM, Baohan WY, Duan ZY, Li W, Zhang JW. 2010. Analysis of two sequence variants in peroxisome proliferator activated receptor gamma gene in Tajik population at high altitudes and Han population at low altitudes in China. *Molecular Biology Reports* 37: 179–184.

Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. 2013. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193: 1233–1254.

Ma B, Chen J, Yang X, Bai J, Ouyang S, Mo X, Chen W, Wang CC, Hai X. 2021. The genetic structure and east-west population admixture in northwest China inferred from genome-wide array genotyping. *Frontiers in Genetics* 12: 795570.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, Skoglund P, Lazaridis I, Sankararaman S, Fu Q, Rohland N, Renaud G, Erlich Y, Willems T, Gallo C, Spence JP, Song YS, Poletti G, Balloux F, van Driem G, de Knijff P, Romero IG, Jha AR, Behar DM, Bravi CM, Capelli C, Hervig T, Moreno-Estrada A, Posukh OL, Balanovska E, Balanovsky O, Karachanak-Yankova S, Sahakyan H, Toncheva D, Yepiskoposyan L, Tyler-Smith C, Xue Y, Abdullah MS, Ruiz-Linares A, Beall CM, Di Rienzo A, Jeong C, Starikovskaya EB, Metspalu E, Parik J, Villems R, Henn BM, Hodoglugil U, Mahley R, Sajantila A, Stamatoyannopoulos G, Wee JT, Khusainova R, Khusnutdinova E, Litvinov S, Ayodo G, Comas D, Hammer MF, Kivisild T, Klitz W, Winkler CA, Labuda D, Bamshad M, Jorde LB, Tishkoff SA, Watkins WS, Metspalu M, Dryomov S, Sukernik R, Singh L, Thangaraj K, Pääbo S, Kelso J, Patterson N, Reich D. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538: 201–206.

Malyarchuk B, Derenko M, Wozniak M, Grzybowski T. 2013. Y-chromosome variation in Tajiks and Iranians. *Annals of Human Biology* 40: 48–54.

Ning C, Wang CC, Gao S, Yang Y, Zhang X, Wu X, Zhang F, Nie Z, Tang Y, Robbeets M, Ma J, Krause J, Cui Y. 2019. Ancient genomes reveal Yamnaya-related ancestry and a potential source of Indo-European speakers in iron age Tianshan. *Current Biology* 29: 2526–2532.

Palstra F, Heyer E P, Austerlitz F. 2015. Statistical inference on genetic data reveals the complex demographic history of human populations in central Asia. *Molecular Biology and Evolution* 32: 1411–1424.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. *Genetics* 192: 1065–1093.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genetics* 2: e190.

Peng MS, Xu W, Song JJ, Chen X, Sulaiman X, Cai L, Liu HQ, Wu SF, Gao Y, Abdulloevich NT, Afanasevna ME, Ibrohimovich KB, Chen X, Yang WK, Wu M, Li GM, Yang XY, Rakha A, Yao YG, Upur H, Zhang YP. 2018. Mitochondrial genomes uncover the maternal

history of the Pamir populations. *European Journal of Human Genetics* 26: 124–136.

Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics* 8: e1002967.

Posth C, Nägele K, Colleran H, Valentin F, Bedford S, Kami KW, Shing R, Buckley H, Kinaston R, Walworth M, Clark GR, Reepmeyer C, Flexner J, Maric T, Moser J, Gresky J, Kiko L, Robson KJ, Auckland K, Oppenheimer SJ, Hill AVS, Mentzer AJ, Zech J, Petchey F, Roberts P, Jeong C, Gray RD, Krause J, Powell A. 2018. Language continuity despite population replacement in remote Oceania. *Nature Ecology & Evolution* 2: 731–740.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. 2007. PLINK: A tool set for whole-genome association and population based linkage analyses. *American Journal of Human Genetics* 81: 559–575.

Quintana-Murci L, Krausz C, Zerjal T, Sayar SH, Hammer MF, Mehdi SQ, Ayub Q, Qamar R, Mohyuddin A, Radhakrishna U, Jobling MA, Tyler-Smith C, McElreavey K. 2001. Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. *The American Journal of Human Genetics* 68: 537–542.

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Gölge M, Dimitrov D, Hill E, Bradley D, Romano V, Calì F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Nørby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozari R, Torroni A, Bandelt HJ. 2000. Tracing European founder lineages in the near eastern mtDNA pool. *The American Journal of Human Genetics* 67: 1251–1276.

Richards MB, Macaulay VA, Bandelt HJ, Sykes BC. 1998. Phylogeography of mitochondrial DNA in western Europe. *Annals of Human Genetics* 62: 241–260.

Semino O, Passarino G, Brega A, Fellous M, Santachiara-Benerecetti AS. 1996. A view of the neolithic demic diffusion in Europe through two Y chromosome-specific markers. *The American Journal of Human Genetics* 59: 964–968.

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Underhill PA. 2000. The genetic legacy of Paleolithic *Homo sapiens* in extant Europeans: A Y chromosome perspective. *Science* 290: 1155–1159.

Sharma S, Rai E, Sharma P, Jena M, Singh S, Darvishi K, Bhat AK, Bhanwer AJ, Tiwari PK, Bamezai RN. 2009. The Indian origin of paternal haplogroup R1a1* substantiates the autochthonous origin of Brahmins and the caste system. *Journal of Human Genetics* 54: 47–55.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.

van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation* 30: E386–394.

Wang CC, Li H. 2013. Inferring human history in East Asia from Y chromosomes. *Investigative Genetics* 4: 11.

Wang CC, Lu Y, Kang L, Ding H, Yan S, Guo J, Zhang Q, Wen SQ, Wang LX, Zhang M, Tong X, Huang X, Nie S, Deng Q, Zhu B, Jin L, Li H. 2019. The massive assimilation of indigenous East Asian populations in the origin of Muslim Hui people inferred from paternal Y chromosome. *American Journal of Physical Anthropology* 169: 341–347.

Wang CC, Yeh HY, Popov AN, Zhang HQ, Matsumura H, Sirak K, Cheronet O, Kovalev A, Rohland N, Kim AM, Mallick S, Bernardos R, Tumen D, Zhao J, Liu YC, Liu JY, Mah M, Wang K, Zhang Z, Adamski N, Broomandkhoshbacht N, Callan K, Candilio F, Carlson KSD, Culleton BJ, Eccles L, Freilich S, Keating D, Lawson AM, Mandl K, Michel M, Oppenheimer J, Özdoğan KT, Stewardson K, Wen S, Yan S, Zalzala F, Chuang R, Huang CJ, Looh H, Shiung CC, Nikitin YG, Tabarev AV, Tishkin AA, Lin S, Sun ZY, Wu XM, Yang TL, Hu X, Chen L, Du H, Bayarsaikhan J, Mijiddorj E, Erdenebaatar D, Iderkhangai TO, Myagmar E, Kanzawa-Kiriyama H, Nishino M, Shinoda KI, Shubina OA, Guo J, Cai W, Deng Q, Kang L, Li D, Li D, Lin R, Nini, Shrestha R, Wang LX, Wei L, Xie G, Yao H, Zhang M, He G, Yang X, Hu R, Robbeets M, Schiffels S, Kennett DJ, Jin L, Li H, Krause J, Pinhasi R, Reich D. 2021. Genomic insights into the formation of human populations in East Asia. *Nature* 591: 413–419.

Wang CZ, Yu XE, Shi MS, Li H, Ma SH. 2022. Whole mitochondrial genome analysis of the Daur ethnic minority from Hulunbuir in the Inner Mongolia Autonomous Region of China. *BMC Ecology and Evolution* 22: 66.

Wells RS, Yuldasheva N, Ruzibakiev R, Underhill PA, Evseeva I, Blue-Smith J, Jin L, Su B, Pitchappan R, Shanmugalakshmi S, Balakrishnan K, Read M, Pearson NM, Zerjal T, Webster MT, Zholoshvili I, Jamarjashvili E, Gambarov S, Nikbin B, Dostiev A, Aknazarov O, Zalloua P, Tsoy I, Kitaev M, Mirrakhimov M, Chariev A, Bodmer WF. 2001. The Eurasian Heartland: A continental perspective on Y-chromosome diversity. *Proceedings of the National Academy of Sciences of the United States of America* 98: 10244–10249.

Wen SQ, Sun C, Song DL, Huang YZ, Tong XZ, Meng HL, Yao HB, Du PX, Wei LH, Wang LX, Wang CC, Shi MS, Lan YM, Wang JC, Jin L, Zhabagin M, Xie XD, Li H. 2020. Y-chromosome evidence confirmed the Kerei-Abakh origin of Aksay Kazakhs. *Journal of Human Genetics* 65: 797–803.

Yan L, Zhu F, He J, Sandler SG. 2006. Human platelet alloantigen systems in three Chinese ethnic populations. *Immunohematology* 22: 6–10.

Yan S, Wang CC, Zheng HX, Wang W, Qin ZD, Wei LH, Wang Y, Pan XD, Fu WQ, He YG, Xiong LJ, Jin WF, Li SL, An Y, Li H, Jin L. 2014. Y chromosomes of 40% Chinese descend from three Neolithic super-grandfathers. *PLoS One* 9: e105691.

Yang X, Sarengaowa HeG, Guo J, Zhu K, Ma H, Zhao J, Yang M, Chen J, Zhang X, Tao L, Liu Y, Zhang XF, Wang CC. 2021. Genomic insights into the genetic structure and natural selection of Mongolians. *Frontiers in Genetics* 12: 735786.

Yao H, Wang M, Zou X, Li Y, Yang X, Li A, Yeh HY, Wang P, Wang Z, Bai J, Guo J, Chen J, Ding X, Zhang Y, Lin B, Wang CC, He G. 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. *Molecular Genetics and Genomics* 296: 631–651.

Zeng F. 2005. *900 sentences in Chinese English and Tajik*. Beijing: Nationalities Publishing House.

Zhang X, He G, Li W, Wang Y, Li X, Chen Y, Qu Q, Wang Y, Xi H, Wang CC, Wen Y. 2021. Genomic insight into the population admixture history of Tungusic-speaking Manchu people in northeast China. *Frontiers in Genetics* 12: 754492.

Zhao J, Wurigemule, Sun J, Xia Z, He G, Yang X, Guo J, Cheng HZ, Li Y, Lin S, Yang TL, Hu X, Du H, Cheng P, Hu R, Chen G, Yuan H, Zhang XF, Wei LH, Zhang HQ, Wang CC. 2020. Genetic substructure and admixture of Mongolians and Kazakhs inferred from genome-wide array genotyping. *Annals of Human Biology* 47: 620–628.

Supplementary Material

The following supplementary material is available online for this article at http://onlinelibrary.wiley.com/doi/10.1111/jse.12957/suppinfo:

Fig. S1. The complete ADMIXTURE analysis result assuming 2–10 ancestral populations.

Fig. S2. The genetic affinity between Wakhi- and Selekur-speaking populations. **a**, Principal component analysis (PCA) of Wakhi and Selekur-speaking groups of Chinese Tajiks with different shapes separately. **b**, Focus on the ADMIXTURE analysis only between Wakhi- and Selekur-speaking groups. **c**, The f_4-statistics in the form of f_4 (Mbuti, reference populations; Tajik_Selekur, Tajik_Wakhi). **d**, The exploration of homogeneity between Tajik_Selekur and Tajik_Wakhi together with relevant reference populations in *qpWave* analysis.

Fig. S3. a, The comparation of Wakhi and Selekur-speaking groups of Chinese Tajiks *in qpAdm* analysis. **b**, The assessment of genetic distance (*Fst*) with Arlequin software based on the mtDNA multiplex sequencing data.

Table S1. Summary of context for the 64 newly reported male samples of Tajiks in Xinjiang province, northwest China.

Table S2. Outgroup-f_3 results in the form of f_3 (Reference groups, Tajik_China; Mbuti) in the order of f_3 values from the largest to the smallest.

Table S3. Admixture-f_3 results in the form f_3 (X, Y; Tajik_China) with the significant negative f_3 values ($Z < -3$).

Table S4. The results of f_4-statistics in the form of f_4 (Mbuti, reference populations; Ancient Xinjiang, Tajik_China).

Table S5. Feasible distal *qpAdm* models with two and three sources for Chinese Tajiks.

Table S6. Admixture time estimated by Admixture-induced Linkage Disequilibrium for Evolutionary Relationships (ALDER).

Table S7. Y chromosomal DNA haplogroup assignment in Chinese Tajik people.

Table S8. A, MtDNA haplogroup assignment in Chinese Tajik people. **B**, The percentages of mtDNA haplogroups in Chinese Tajiks.

Table S9. Population pairwise *Fst* based on the method of Tajima and Nei.

Table S10. A, Previously reported ancient individuals co-analyzed with new data. **B**, Previously reported present-day individuals co-analyzed with new data.