



The Genetic Echo of the Tarim Mummies in Modern Central Asians

Shan-Shan Dai,^{†,1,15} Xierzhatijiang Sulaiman,^{†,1,2} Jainagul Isakova,^{†,3} Wei-Fang Xu,^{†,4} Najmudinov Tojiddin Abdulloevich,⁵ Manilova Elena Afanasevna,⁵ Khudoidodov Behruz Ibrohimovich,⁵ Xi Chen,^{6,7} Wei-Kang Yang,⁷ Ming-Shan Wang,⁸ Quan-Kuan Shen,^{1,15} Xing-Yan Yang,^{9,10} Yong-Gang Yao ,^{11,12,15} Almaz A. Aldashev,³ Abdusattor Saidov,⁵ Wei Chen,^{13,14} Lu-Feng Cheng,^{*,2} Min-Sheng Peng^{*,1,12,15} and Ya-Ping Zhang ^{*,1,12,15,16}

¹State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

²Department of Pharmacology, School of Pharmacy, Xinjiang Medical University, Urumqi, China

³Institute of Molecular Biology and Medicine, Bishkek, Kyrgyzstan

⁴Shenzhen Hospital of Guangzhou University of Chinese Medicine, Shenzhen, China

⁵E.N. Pavlovsky Institute of Zoology and Parasitology, Academy of Sciences of Republic of Tajikistan, Dushanbe, Tajikistan

⁶Research Center for Ecology and Environment of Central Asia, Chinese Academy of Sciences, Urumqi, China

⁷State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi, China

⁸Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA

⁹Key Laboratory of Chemistry in Ethnic Medicinal Resource, Yunnan Minzu University, Kunming, China

¹⁰School of Chemistry and Environment, Yunnan Minzu University, Kunming, China

¹¹Key Laboratory of Animal Models and Human Disease Mechanisms of the Chinese Academy of Sciences & Yunnan Province, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

¹²KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research in Common Diseases, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

¹³College of Agronomy and Biotechnology, Yunnan Agricultural University, Kunming, China

¹⁴State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan, Yunnan Agricultural University, Kunming, China

¹⁵Kunming College of Life Science, University of Chinese Academy of Sciences, Kunming, China

¹⁶State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan, School of Life Sciences, Yunnan University, Kunming, China

[†]These authors contributed equally to this work.

***Corresponding authors:** E-mails: lfcheng@xjmu.edu.cn; pengminsheng@mail.kiz.ac.cn; zhangyp@mail.kiz.ac.cn.

Associate editor: Connie Mulligan

Abstract

The diversity of Central Asians has been shaped by multiple migrations and cultural diffusion. Although ancient DNA studies have revealed the demographic changes of the Central Asian since the Bronze Age, the contribution of the ancient populations to the modern Central Asian remains opaque. Herein, we performed high-coverage sequencing of 131 whole genomes of Indo-European-speaking Tajik and Turkic-speaking Kyrgyz populations to explore their genomic diversity and admixture history. By integrating the ancient DNA data, we revealed more details of the origins and admixture history of Central Asians. We found that the major ancestry of present-day Tajik populations can be traced back to the admixture of the Bronze Age Bactria–Margiana Archaeological Complex and Andronovo-related populations. Highland Tajik populations further received additional gene flow from the Tarim mummies, an isolated ancient North Eurasian–related population. The West Eurasian ancestry of Kyrgyz is mainly derived from Historical Era populations in Xinjiang of China. Furthermore, the recent admixture signals detected in both Tajik and Kyrgyz are ascribed to the expansions of Eastern Steppe nomadic pastoralists during the Historical Era.

Key words: Central Asia, Kyrgyz, Tajik, admixture, genome, Steppe, Tarim.

© The Author(s) 2022. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

Introduction

Central Asia, located at the crossroads of Eurasia, is a key area for studying human evolution (Nei and Roychoudhury 1993). According to the historical and archaeological evidence, various migrations intertwining with language and culture shifts have intersected in the region (Harmatta et al. 1994; Bregel 2003; Findley 2004) and shaped human genetic diversity. Early studies based on microsatellite, Y-chromosomal, and mitochondrial DNA (mtDNA) markers showed that the genetic diversity of Central Asians is among the highest in Eurasia (Comas et al. 1998; Pérez-Lezaun et al. 1999; Hammer et al. 2001; Wells et al. 2001; Zerjal et al. 2002; Martínez-Cruz et al. 2011). Two hypotheses were proposed to explain this pattern. The Central Asian heartland hypothesis suggested Central Asia as a source for the genetic diversity of Eurasians (Wells et al. 2001), and the genetic admixture hypothesis proposed the Central Asians as the admixture of eastern and western Eurasians (Comas et al. 1998; Quintana-Murci et al. 2004; Yao et al. 2004; Palstra et al. 2015; Peng et al. 2018). The subsequent analyses for genome-wide single nucleotide polymorphisms (SNPs) genotyped with DNA chips further proposed a very complex scenario of genetic admixture in Central Asians, in which a hypothesis for multiple wave of gene flows from European, West Asian, and South Asian to the Central Asian gene pool was favored (Li et al. 2009; Feng et al. 2017) and some recent migration/admixture events were revealed (Hellenthal et al. 2014; Yunusbayev et al. 2015; Feng et al. 2017; Jeong et al. 2019).

In recent years, the ancient DNA investigation based on genome-wide level updates the genetic view of the evolutionary history of Central Asian populations (Allentoft et al. 2015; Unterländer et al. 2017; de Barros Damgaard et al. 2018; Järve et al. 2019; Narasimhan et al. 2019; Ning et al. 2019; Gneccchi-Ruscone et al. 2021; Kumar et al. 2021, 2022). In brief, the migration and admixture involving different genetic ancestries since the Bronze Age have been indicated. The Bactria–Margiana Archaeological Complex (BMAC) is characterized by genetic ancestries from Iranian early farmer-related ancestry (~60–65%) and smaller proportions of Anatolian

farmer–related ancestry (~20–25%) and West Siberian hunter-gatherer-related ancestry (~10%; Narasimhan et al. 2019), flourished in southern Central Asia (Dani and Masson 1992). Around 4,100 BP, Steppe-related ancestry appeared in Central Asia (Narasimhan et al. 2019). In the Iron Age, the genetic component of eastern nomads was found in the Central Steppe Scythians and Xinjiang populations (Unterländer et al. 2017; Wang Ding, et al. 2021; Guarino-Vignon et al. 2022; Kumar et al. 2022). Most recently, a novel genetic ancestry represented by the mummies of the Early and Middle Bronze Age from Tarim Basin (i.e., Tarim_EMBA1) located in Xinjiang of northwestern China was identified. Especially, the Tarim_EMBA1 was proposed to be mainly derived from Ancient Northern Eurasian (ANE) populations and isolated since the early Holocene (Zhang et al. 2021). It is still unclear about the history of the unique Xinjiang Bronze Age component as Tarim_EMBA1 in Central Asia. Although the complex demographic history of the Xinjiang populations during the past 5,000 years has been revealed (Kumar et al. 2022), how the admixture of the above genetic ancestries contributed to the modern Central Asian populations remains opaque.

The high-resolution of whole-genome sequencing (WGS) data facilitates the integration with ancient DNA data that are heterogenous in sequencing depth and quality to illustrate complex dynamics (Ioannidis et al. 2020; Almarri et al. 2021; Kivisild et al. 2021). In this study, we conducted high-depth WGS for a total of 131 individuals (table 1) from two representative ethnic groups in Central Asia: Kyrgyz (Turkic language speakers) and Tajik (Indo-European language speakers; fig. 1), which are underrepresented in the global WGS panels (1000 Genomes Project Consortium 2015; Bergström et al. 2020; Mallick et al. 2016; Pagani et al. 2016). By leveraging various population genomic approaches, we integrated massive ancient DNA and WGS data of modern populations to explore the admixture history of the Kyrgyz and Tajik populations. Our results refine the understanding of the origins and admixture dynamics of Central Asians.

Table 1. A List of the 252 Modern Samples Used for Joint SNP Genotyping.

Population	Size	Language	Location	Reference
Kyrgyz of Kyrgyzstan	42	Turkic	Kyrgyzstan	This study
Kyrgyz of China	30	Turkic	Xinjiang, China	This study
Sarikoli Tajik	19	Indo-European	Xinjiang, China	This study
Wakhi Tajik	20	Indo-European	Xinjiang, China	This study
Dushanbe Tajik	20	Indo-European	Dushanbe, Tajikistan	This study
Kashmiri	20	Indo-European	Azad Kashmir, Pakistan	Yang et al. (2021)
Balti	18	Sino-Tibetan	Gilgit-Baltistan, Pakistan	Yang et al. (2021)
Punjabi	2	Indo-European	Punjab, Pakistan	Yang et al. (2021)
Pamiri Tajik	20	Indo-European	Gorno-Badakhshan, Tajikistan	Yang et al. (2021)
Tibetan	8	Sino-Tibetan	Tibet, China	Yang et al. (2018)
Tibetan	33	Sino-Tibetan	Tibet, China	Lu et al. (2016)
Persian	20	Indo-European	Kerman, Iran	Charati et al. (2019)

Results

Population Genomic Variation

We performed WGS for 131 individuals from two representative Central Asian populations (fig. 1): Kyrgyz ($n = 72$, from two groups: Kyrgyz of China and Kyrgyz of Kyrgyzstan) and Tajik ($n = 59$, from three groups: Sarikoli Tajik, Wakhi Tajik, and Dushanbe Tajik). The average genomic sequencing depth for each sample is over 30X. We conducted the joint SNP genotyping with the published high-depth WGS data for 121 individuals from the surrounding regions of Central Asia (Lu et al. 2016; Yang et al. 2018; Charati et al. 2019; Yang et al. 2021) to output all sites of hg19 (Mallick et al. 2016). After quality control, we got a total of 11,000,006 SNPs (supplementary fig. S1, Supplementary Material online) in the newly sequenced 131 genomes. A total of 2,888,675 SNPs are not recorded in the dbSNP (version 138) database (Sherry et al. 2001). And 2,737,785 SNPs are not reported in the 1000 Genomes Project (1000 Genomes Project Consortium 2015; supplementary fig. S1, Supplementary Material online). This illustrates the importance of sequencing genetically underrepresented Central Asian populations.

To understand the genetic impact of ancient populations on Central Asians, we merged the data set of 252 genomes with previously published modern Eurasian, African, American, and ancient populations data (supplementary table S1, Supplementary Material online) to obtain a total of 659,765 SNPs for 2,456 individuals. The details for sampling, sequencing, SNP calling, data merging, and filtering were described in Materials and Methods.

Population Structure Analyses

We conducted principal component analysis (PCA; Patterson et al. 2006) to assess the genetic affinity between the ancient individuals and modern Central Asians by projecting ancient samples onto the context of genetic variation in present-day Eurasians (fig. 2 and supplementary fig. S2, Supplementary Material online). The PC1 distinguishes West Eurasians (i.e., Europeans and West Asians) and South Asians from East Eurasians (Eastern Asians and Siberians), and PC2 further splits South Asians from West Asians and Europeans. The Kyrgyz individuals from China/Kyrgyzstan distribute in the center of PCA and cluster largely according to their geographic locations. The Kyrgyz of China cluster closer with Europeans and South Asians indicating that they have a higher proportion of west Eurasian component than the Kyrgyz of Kyrgyzstan. The Tajik populations spread between South and West Asians, and cluster in line with their geographic locations: the Dushanbe Tajik west of the Pamirs separates from the Sarikoli Tajik, Pamiri Tajik, and Wakhi Tajik living in the Pamirs (supplementary fig. S2, Supplementary Material online). In the context of ancient DNA data, since the Paleolithic Age, the Tajik populations overlap with Russia_AfontnovaGora2 presenting a high proportion of ANE ancestry (Raghavan et al. 2014). The Kyrgyz populations are close to Mongolia_Salkhit_UP (supplementary

fig. S3, Supplementary Material online), which mainly harbors East Eurasian-related components (Massilani et al. 2020). The Dushanbe Tajik stretches toward Anatolia_N when compared with the other three Tajik populations (fig. 2A), likely suggesting a higher proportion of Anatolian farmer-related ancestry in the Dushanbe Tajik. Several Bronze Age Xinjiang individuals (e.g., Xinj_BA3, Xinj_BA4, Dzungaria_EBA1, and Dzungaria_EBA2) cluster closely with the four Tajik populations (fig. 2B), which may reflect a high genetic similarity between Tajik and Bronze Age Xinjiang populations. The Iron Age and Historical Era Central Asian and Steppe individuals are clearly separated from the four Tajik populations, except for the sporadic individuals from Xinjiang and South Asia (e.g., JEZK_IA3_oBMAC, LSH_IA2_oSte, Xinj_HE1, and Pakistan_RajaGira; fig. 2C and D). In contrast, a lot of Iron Age and Historical Era Xinjiang, Central Asian, and Mongolian individuals cluster with Kyrgyz. The patterns reflect different admixture histories involved in the formation of Kyrgyz and Tajik populations.

We then performed the model-based ADMIXTURE clustering analysis (Alexander et al. 2009) to get a profile of the ancestry admixture. We presented the result from the model of $K = 8$ with the lowest cross-validation error (supplementary fig. S4, Supplementary Material online). Under this model, the Kyrgyz and Tajik show East and West Eurasian-admixed profiles (fig. 3 and supplementary fig. S5, Supplementary Material online) with six distinctive components: the Anatolian Neolithic farmer-related ancestry (Anatolian_N; purple), the Iran Neolithic farmer-related ancestry (Iran_GanjDareh_N; pink), the ANE-related ancestry (dark blue) dominated in Tarim_EMBA1, the West European hunter-gatherer-related ancestry (WEHG; sky blue), the ancient East Asian-related ancestry (green), and the Baikal hunter-gatherer-related ancestry (yellow). In general, the six ancestry components are widely present in ancient DNA samples across Central Asia (supplementary figs S6 and S7, Supplementary Material online). The Tajiks have a higher proportion of ANE-related ancestry, which is dominant in Tarim_EMBA1 (dark blue), than any other modern Central Asians (fig. 3).

Genetic Contribution from Ancient Eurasians

To further depict the genetic affinities between ancient Eurasian populations and modern Central Asians (Kyrgyz and Tajik), we first calculated the f_4 -statistics (Reich et al. 2009; Patterson et al. 2012) in the form of $f_4(\text{Population1}, \text{Population2}; \text{Tajik/Kyrgyz}, \text{Mbuti})$, which shows that the Tajik has a greater affinity with Bronze Age populations from Central Asia and Steppe (e.g., Tarim_EMBA1, Xinj_BA3, Central_Steppe_EMBA, and Russia_Samara_EBA_Yamnaya) than with the Iron Age populations from the same regions (e.g., Tajikistan_Ksirov_Kushan, Sarmatians_450BCE, JEZK_IA2, JEZK_IA1_aSte, JEZK_IA3_oBMAC, and Turkmenistan_IA). The Kyrgyz populations show a greater affinity to the Iron Age and Historical Era Central Asian populations (e.g., Saka_TianShan_600BCE,

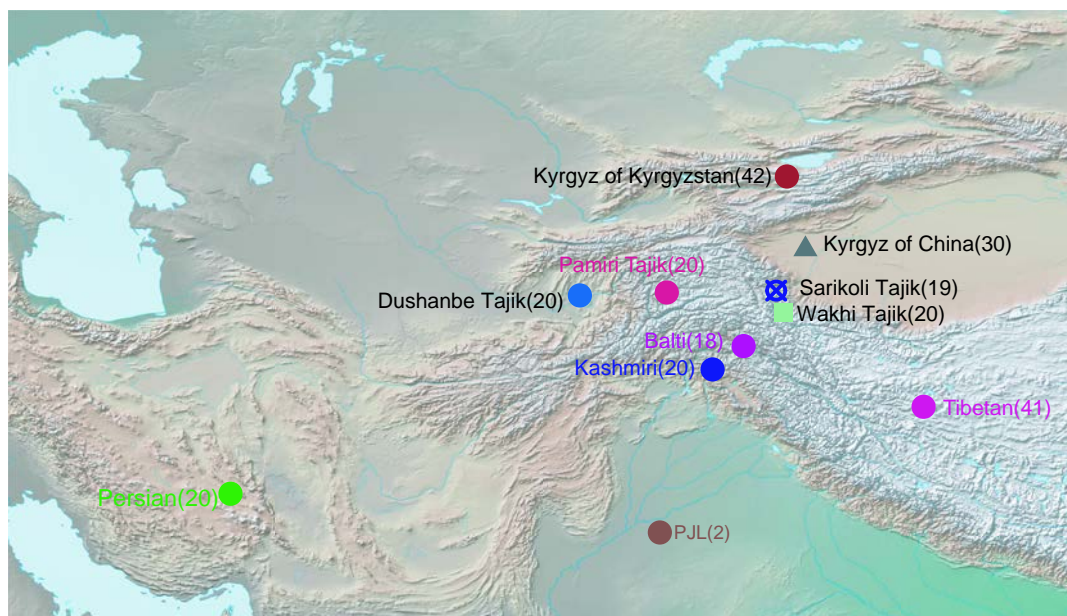


Fig. 1. Sample locations of the 252 modern samples used for joint SNP genotyping. A total of 252 individuals are included. The populations sequenced in this study are including: Kyrgyz of Kyrgyzstan, Kyrgyz of China, Sarikoli Tajik, Wakhi Tajik, and Dushanbe Tajik. The Wakhi Tajiks are immigrants from Wakhan Corridor since late 19th and early 20th Century. The detailed information is described in Table 1. The map was obtained from the Natural Earth public domain map data set (<https://www.naturalearthdata.com/downloads/10m-raster-data/10m-cross-blend-hypso/>).

China_Xinjiang_IA, Xinj_HE7, and Xinj_IA2_aEA) rather than those of Bronze Age populations (supplementary figs S8–S11 and table S2, Supplementary Material online).

We calculated the shared genetic drift using the outgroup f_3 statistic (Patterson et al. 2012) in the form of $f_3(\text{Mbuti}; \text{Kyrgyz/Tajik}, X)$, where Mbuti was used as the outgroup, and X presented a set of Eurasian populations. The highest genetic affinities between four Tajik populations and Bronze Age populations (e.g., Tarim_EMBA1, Central_Steppe_EMBA, and Russia_Samara_EBA_Yamna ya) are verified (supplementary fig. S12 and table S3, Supplementary Material online). Especially, compared with other modern Central Asian populations, the three Tajik from the Pamirs (i.e., Sarikoli Tajik, Wakhi Tajik, and Pamiri Tajik) share a higher level of genetic drift with Tarim_EMBA1 (supplementary fig. S13, Supplementary Material online). The Kyrgyz populations share great genetic drift with Neolithic, Bronze, and Iron Age populations from northern China and Mongolia (e.g., China_Wuzhuangguoliang_LN.EC, China_WLR_BA_o, China_AR_Xianbei_IA, and Mongolia_EIA_8), suggesting the genetic continuity in East Eurasia since the Neolithic Age (supplementary fig. S14 and table S3, Supplementary Material online).

Inference of Admixture Scenarios

We adopted qpAdm (Patterson et al. 2012; Haak et al. 2015) to infer the admixture models including the ancestral sources and their related genetic proportions in the Kyrgyz and Tajik populations, respectively. We first used

the distal modeling, with pre-Copper Age populations or genetically isolated populations as sources of admixture. The distal models with the best fit of Kyrgyz and Tajik can be modeled with five sources (supplementary table S4, Supplementary Material online). The major ancestry components in the Kyrgyz are from Baikal hunter-gatherer (i.e., Russia_Shamanka_Eneolithic; 59.3–69.8%) and Iranian farmer-related ancestries (16–23.8%). The remaining minor ancestry components are from Anatolian farmers (5.1–5.6%), Western European hunter-gatherers (5.3–6.6%) and ANE-related Tarim_EMBA1 (3.2–5.3%). The ancestry profiles of Tajik populations can be dissected into five components from related ancestries of Iranian farmer (43.8–52.8%), ANE (13.3–15.8%), Western European hunter-gatherer (9.5–11.8%), Baikal hunter-gatherer (7.7–17.1%), and Anatolian farmer (9.7–15.6%).

We then conducted proximal modeling with Bronze Age, Iron Age, and Historical Era populations as sources for modern Central Asians. The Sarikoli Tajik and Pamiri Tajik can be modeled as a mixture of Russia_Andronovo, BMAC, Tarim_EMBA1, and Mongolia_Xiongnu_o1 (supplementary table S5, Supplementary Material online and fig. 4A). The results are supported by the qpWave (Reich et al. 2012) analysis that at least three separate sources are present in the Tajik populations (supplementary table S6, Supplementary Material online). The Tarim_EMBA1 is required under the four-way models in the admixture inference for the Sarikoli Tajik and Pamiri Tajik. When removing the Tarim_EMBA1, the admixture modeling failed. For the Sarikoli Tajik and Pamiri Tajik, the admixture models unanimously failed when using any group of Russia_MLBA_Sintashta,

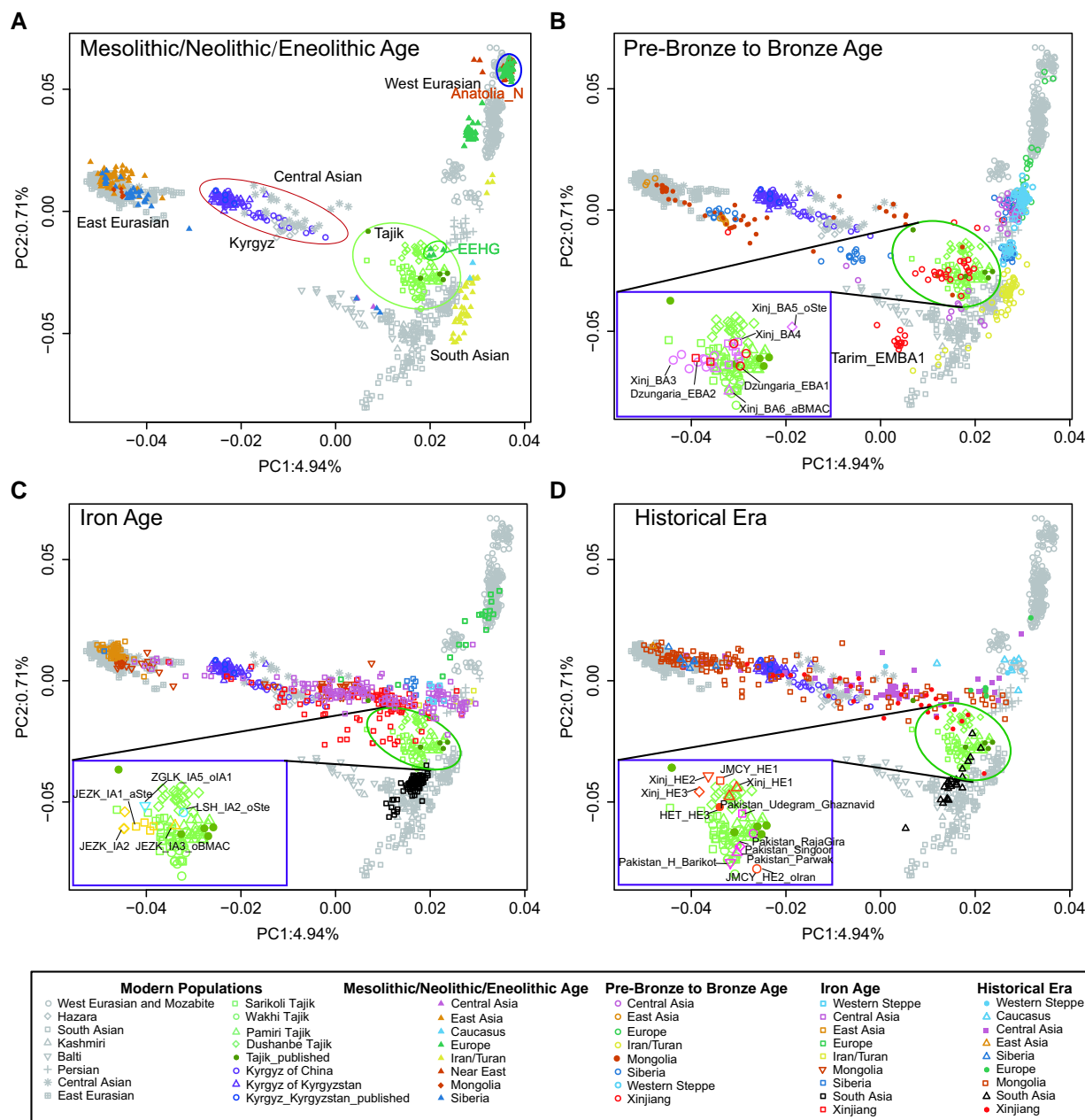


Fig. 2. Principal component analysis of the modern and ancient Eurasians. A total of 1,372 ancient individuals are projected onto the first two principal components defined by modern Eurasian populations. (A) Mesolithic to Eneolithic Age, (B) Pre-Bronze to Bronze Age, (C) Iron Age and (D) Historical Era individuals were shown.

Central_Steppe_MLBA, Russia_Afnasievo, and Russia_Samara_EBA_Yamnaya as a Steppe source. For three highland Tajik (i.e., Sarikoli Tajik, Wakhi Tajik, and Pamiri Tajik), when we replaced Russia_Andronovo or Russia_Andronovo and BMAC with Turkmenistan_IA, the models worked well (supplementary table S5, Supplementary Material online), as Turkmenistan_IA was an admixture of BMAC and Andronovo (Guarino-Vignon et al. 2022). The Dushanbe Tajik can be modeled as a mixture of Turkmenistan_IA and Mongolia_Xiongnu_o1. All the results suggest Russia_Andronovo as the proxy for the Steppe ancestry of Tajiks. Consequently, the major ancestry of Tajiks can be traced back to the Bronze Age populations admixed

with BMAC and Andronovo, and then, the highland Tajik received additional gene flow from Tarim_EMBA1. For comparison, the Turkic-speaking populations present different patterns. The major ancestry of Kyrgyz and Kazakh populations is from Xinj_HE3 (44.8–58.9%) and Mongolia_Xiongnu_o1 (41.1–55.2%; supplementary table S5, Supplementary Material online and fig. 4A). The Uyghur, Uzbek, and Turkmen populations are modeled as a mixture of Turkmenistan_IA (48.8–65.1%) and Mongolia_Xiongnu_o1 (34.9–51.2%; supplementary table S5, Supplementary Material online and fig. 4A), which is consistent with the population history of Uyghur (Feng et al. 2017).

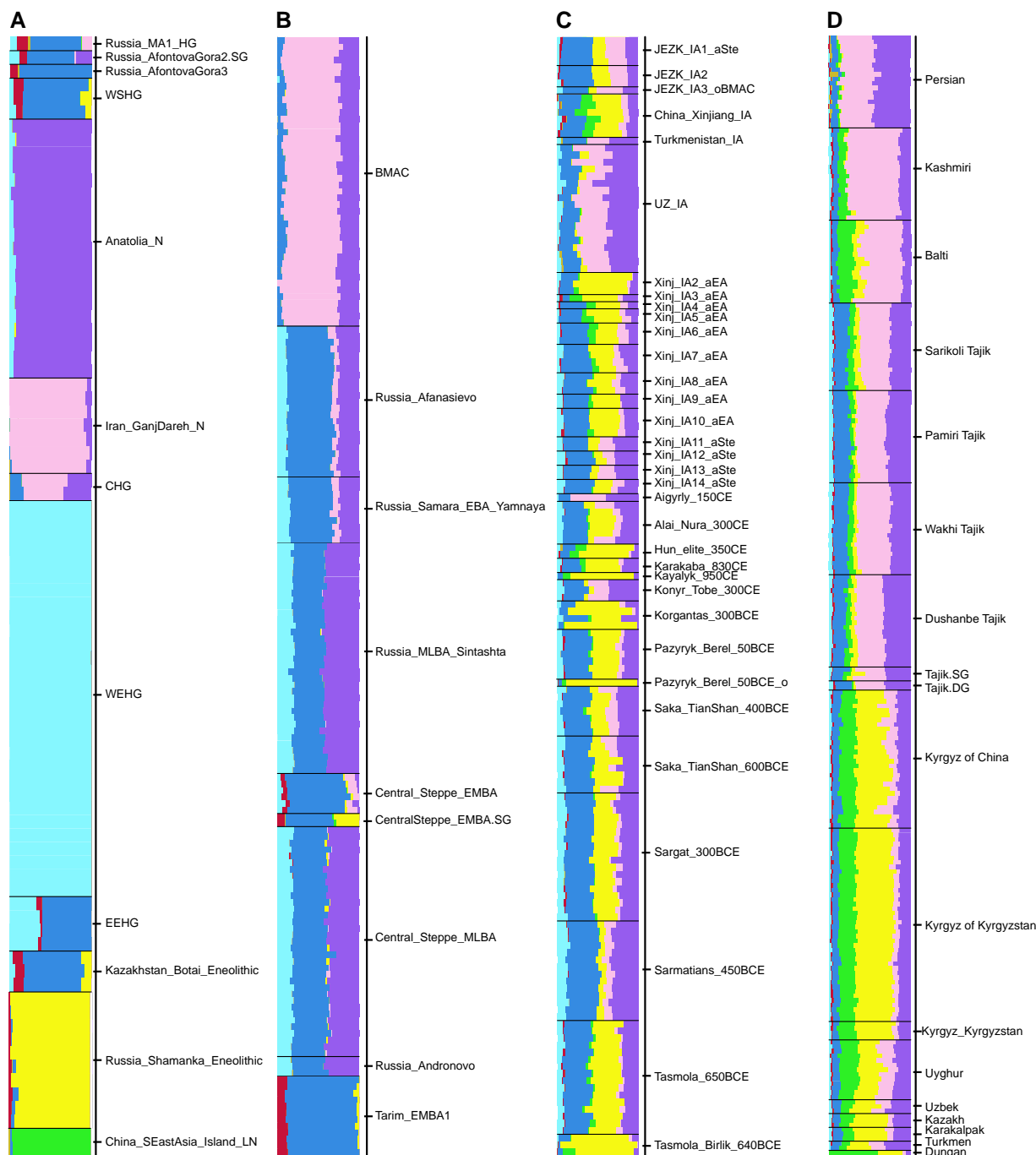


FIG. 3. Admixture ancestry components ($K = 8$) of Kyrgyz, Tajik, other modern Central Asians, and ancient Eurasian populations. (A) Paleolithic to Eneolithic Age populations. (B) Bronze Age populations. The Anatolian_N (purple)- and the Iran_GanjDareh_N (pink)-related ancestry components coexist in the BMAC individuals, reflecting the admixture origin of BMAC (Narasimhan et al. 2019). Similarly, the ancient Steppe populations are indicated by the admixture of the EEHG (dark blue and sky blue)- and CHG (pink, purple, and dark blue)-related ancestry components (de Barros Damgaard et al. 2018). The Tarim_EMBA1 had minor component (red) prevalent in the Native American populations (supplementary fig. S5, Supplementary Material online), which is in agreement with the ANE ancestry of the Tarim_EMBA1 (Zhang et al. 2021). (C) Iron Age populations. (D) Modern populations.

Finally, we employed DATES (Narasimhan et al. 2019) to date the admixture events involved in the Tajik and Kyrgyz populations. For the Sarikoli Tajik and Pamiri Tajik, the admixture of BMAC and Andronovo occurred 5,970 (3,951–7,988) and 3,211 (2,550–3,871) years ago, respectively, which was compatible with the time of the Steppe-related ancestry appearing in Central Asia

(Narasimhan et al. 2019). The broad range will require sampling additional Andronovo individuals to refine. Using BMAC as the dominant source, the gene flows derived from Tarim_EMBA1 into the Sarikoli Tajik and Pamiri Tajik populations were dated to 2,957 (2,326–3,587) and 3,616 (2,818–4,414) years ago, respectively. The admixture in the four Tajik populations, with the recent westward

dispersal of eastern Eurasian represented by Mongolia_Xiongnu_o1 was dated to 805–1,418 years ago (supplementary table S7 and fig. S15, Supplementary Material online). For the two Kyrgyz populations, the major admixture events involved in Mongolia_Xiongnu_o1 and Xinj_HE3 were dated to 493 (417–570) and 784 (629–940) years ago, respectively (supplementary table S7 and fig. S16, Supplementary Material online).

mtDNA and Y-Chromosome Markers

To investigate the impact of the ancient ancestry on the maternal and paternal gene pools of Kyrgyz and Tajik populations, we analyzed the variation of mtDNA and Y chromosome retrieved from the WGS data. The haplogroup profiles are shown in supplementary table S8, Supplementary Material online. The mtDNA haplogroups U and H which were proposed as the ancient Iran/Turan and Steppe-related connections (Haak et al. 2015; Sahakyan et al. 2017; Narasimhan et al. 2019) exist in the Kyrgyz and Tajik populations. The mtDNA haplogroup C4 characterized in Tarim_EMBA1 (Zhang et al. 2021) is also found in the Kyrgyz and Tajik. The most prevalent paternal lineage in the Kyrgyz (26/44) and Tajik (16/33) populations was haplogroup R1a1 (supplementary table S8, Supplementary Material online), which has been reported in Steppe-related populations, such as Corded Ware, Andronovo, and Sintashta (Mathieson et al. 2015; Krzewińska et al. 2018; Shriner 2018). The Y haplogroup R1b1 characterized in Yamnaya and Afanasievo (Allentoft et al. 2015; Mathieson et al. 2015) is also found in the Tajik (2/33). The Y haplogroups J and R2 existing widely in the Bronze Age Iran/Turan (Narasimhan et al. 2019) are also found in modern Central Asians. These results indicate that both females and males with the Steppe and BMAC-related ancestries have contributed to the gene pools of Kyrgyz and Tajik populations.

Estimation of Endogamy

The cultural impact on genetic diversity is a hot topic in Central Asia (Chaix et al. 2007). In the context of admixture history, we identified the runs of homozygosity (ROHs; Ringbauer et al. 2021) in the ancient populations and the Tajik and Kyrgyz populations to explore the history of endogamous and exogamous marriages in Central Asia. The Tajiks have a high proportion of individuals above the long ROH threshold (the total length of ROHs longer than 20 cM is over 50 cM; 18 out of 79 in total). By contrast, the Kyrgyz populations present low levels of ROHs (supplementary fig. S17, Supplementary Material online). The patterns are in accordance with the endogamous and exogamous marriages in Tajik and Kyrgyz, respectively (Chaix et al. 2007). Interestingly, both BMAC and Russia_Andronovo populations, that is, the sources for the Tajiks, have a low level of ROHs, implying that they might not adopt endogamous marriages. The Tarim_EMBA1 individuals present higher levels of ROHs than BMAC and Russia_Andronovo (supplementary fig.

S18, Supplementary Material online). We proposed a parsimonious scenario for endogamy practiced in the Tajiks likely after the admixture of BMAC and Andronovo populations around 3,211–5,970 years ago (fig. 4B).

Discussion

Given Central Asia is endowed with complicated terrain including steppe, oasis, valley, desert, and highland, the genetic diversity as well as population structure are still underrepresented in available Central Asian genomes (Mallick et al. 2016; Bergström et al. 2020). In this study, we conducted the largest WGS for Central Asians. The newly generated 131 high-depth genomes from three Tajik and two Kyrgyz populations expand the catalog of genetic variation to underrepresented Central Asian populations. Our results indicate the population structure in the modern Tajik and Kyrgyz ethnic groups, mainly corresponds to the geographic factors (supplementary fig. S2, Supplementary Material online), and suggest that the genome data of ethnic populations from different geographical areas of Central Asians are essential to studying their population history.

Despite that a series of genetic studies have been done to investigate the admixture history of the Central Asians (Martínez-Cruz et al. 2011; Palstra et al. 2015; Feng et al. 2017); however, due to the limited resolution of genetic markers and lacking source panels of ancient populations, these studies dissected the ancestry components derived from modern Eurasians, that is, the indirect representatives of genetic sources. Leveraging the advances of high-depth sequenced genomes from Central Asia and its surrounding regions (Lu et al. 2016; Yang et al. 2018; Charati et al. 2019; Bergström et al. 2020; Yang et al. 2021), we re-appraised the origin of admixture history for the Tajik and Kyrgyz populations in the context of ancient Eurasian populations across a broad time span (Mathieson et al. 2015; Damgaard et al. 2018; Narasimhan et al. 2019; Ning et al. 2019; Jeong et al. 2020; Gneccchi-Ruscione et al. 2021; Kumar et al. 2021, 2022; Zhang et al. 2021). Our results revealed that the Tajik populations present high genetic affinity with the Bronze Age Central Asians, especially from Xinjiang of China (fig. 2B and supplementary figs S8, S9, S11, and S12, Supplementary Material online). The major ancestry components in the four Tajik populations could be traced back to the admixture of BMAC and Andronovo (fig. 4A and 4B). Given the Steppe-related ancestry (e.g., Andronovo) existing in Scythians (i.e., Saka; Unterländer et al. 2017; Damgaard et al. 2018; Guarino-Vignon et al. 2022), the proposed linguistic and physical anthropological links between the Tajiks and Scythians (Han 1993; Kuz'mina and Mallory 2007) may be ascribed to their shared Steppe-related ancestry. By contrast, the Kyrgyz, together with other Turkic-speaking populations, originated from the admixture since the Iron Age. The Historical Era gene flow derived from the Eastern Steppe with the representative of Mongolia_Xiongnu_o1 made a

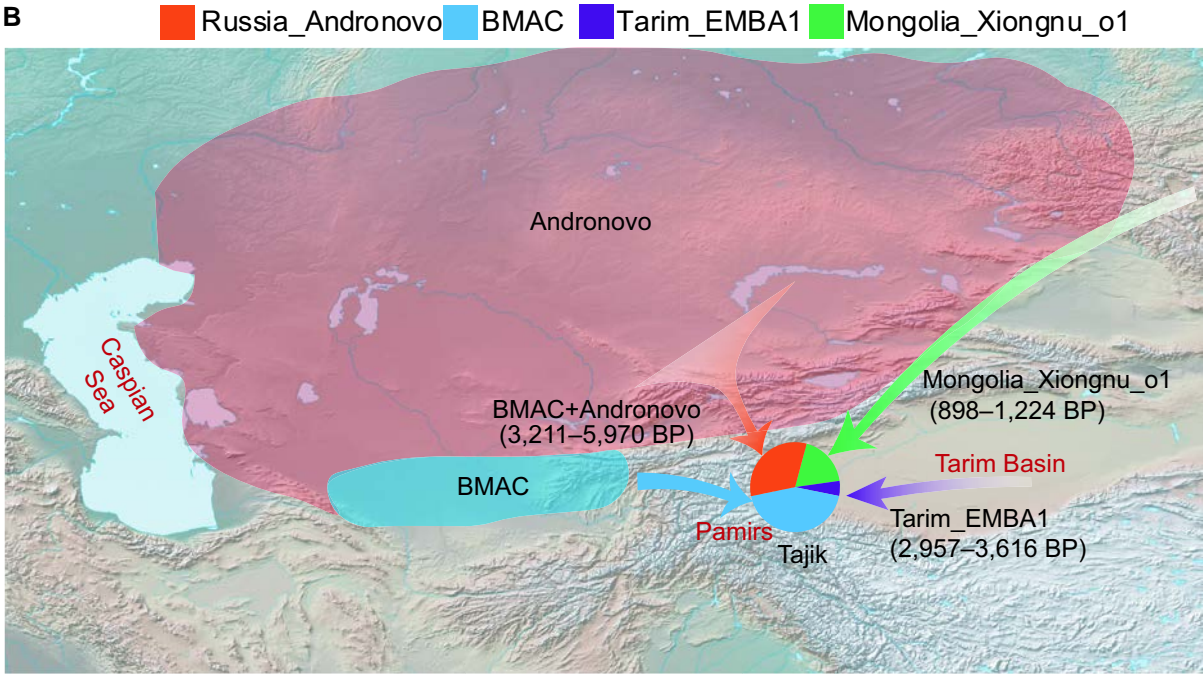
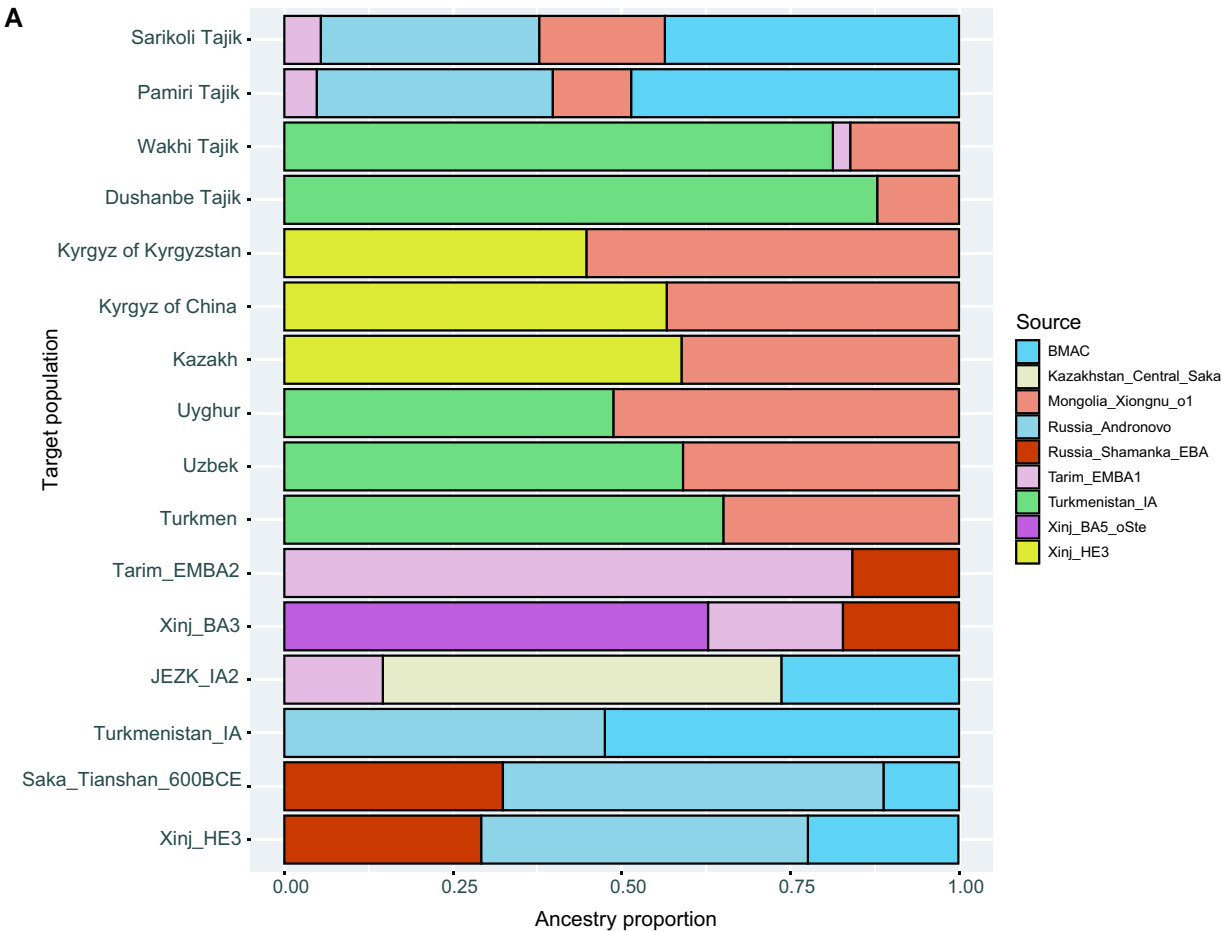


FIG. 4. Inferred qpAdm models and admixture history of the Sarikoli Tajik and Pamiri Tajik. (A) Feasible qpAdm proximal models for four Tajik, Kyrgyz, and other modern and ancient populations. The Iron Age Turkmenistan_IA can be modeled as a mixture of Andronovo and BMAC. The Wakhi Tajik and Dushanbe Tajik modeled with Turkmenistan_IA and other populations reflect their common ancestries from Andronovo and BMAC. (B) Schematic map showing the possible admixture model of Sarikoli Tajik and Pamiri Tajik. The time in parentheses represent a range. Arrows in different colors indicate ancestral sources and directions of the gene flows. The map was obtained from the Natural Earth public domain map data set (<https://www.naturalearthdata.com/downloads/10m-raster-data/10m-cross-blend-hyps/>).

more substantial contribution to Kyrgyz and other Turkic-speaking populations (i.e., Kazakh, Uyghur, Turkmen, and Uzbek; 34.9–55.2%) higher than that to the Tajik populations (11.6–18.6%; [fig. 4A](#)), suggesting Tajiks suffer fewer impacts of the recent admixtures ([Martínez-Cruz et al. 2011](#)). Consequently, the Tajik populations generally present patterns of genetic continuity of Central Asians since the Bronze Age. Our results are consistent with linguistic and genetic evidence that the spreading of Indo-European speakers into Central Asia was earlier than the expansion of Turkic speakers ([Kuz'mina and Mallory 2007](#); [Yunusbayev et al. 2015](#)).

More importantly, we identified the newly characterized ancestry component represented by Tarim_EMBA1 as a genetically isolated ANE-related population ([Zhang et al. 2021](#)), which left a genetic legacy in the modern Central Asian gene pool. Although Tarim_EMBA1 ancestry existed in the Iron Age and Historical Era Xinjiang populations ([Kumar et al. 2022](#)), among modern Central Asian populations, the Tarim_EMBA1 was only detectable in the Sarikoli Tajik, Wakhi Tajik, and Pamiri Tajik from the Pamirs neighboring Tarim Basin ([fig. 1](#)). In the Dushanbe Tajik west of the Pamirs, as well as other Turkic-speaking populations, we failed to detect the signature of Tarim_EMBA1 ancestry. It is expected that western Tajiks in Uzbekistan from the previous study ([Guarino-Vignon et al. 2022](#)) may have no Tarim_EMBA1 ancestry. By integrating evidence from archaeology and genetic studies ([Mallory and Mair 2000](#); [Zhang et al. 2021](#); [Kumar et al. 2022](#)), we propose an intriguing scenario that, the isolated Early-Middle Bronze Age Tarim populations (language is unclear) have not vanished completely. After abandoning the settlements in Tarim Basin, they likely migrated into the Pamirs and then admixed with the Indo-European speakers ~3,286 years ago. Tarim_EMBA1 ancestry is maintained in the Iron Age populations (i.e., JEZK_IA2 from Taxkorgan of Xinjiang; [Kumar et al. 2022](#)) and modern Tajik populations from the Pamirs. The interaction was also indicated by the archaeological evidence of wheat and barley imported from West Asia appeared in both the Pamirs and Tarim Basin during the Bronze Age ([Zhang et al. 2016](#); [Chen et al. 2017](#)). It supports that the Pamirs serving not only as geographic corridors for West and East Eurasian cultural interaction ([Frachetti 2012](#); [Li 2021](#)) but also as a refugium for the isolated Early-Middle Bronze Age Tarim populations.

Taken together, we unveiled a more delicate scenario of ancestral origins, population structure, and admixture history of Central Asians than previously reported. However, there is a long way to go in the future study. First, the multiple-wave admixture dating method based on ancient DNA is required. Second, more ancient DNA data from Central Asia and the neighboring regions ([Wang, Yeh, et al. 2021](#)) across time and space, especially with high quality ([Orlando et al. 2021](#); [Marchi et al. 2022](#)) are essential to refining the details about the demographic dynamics, local adaptation, and phenotypic evolution in the heartland of Eurasia.

Materials and Methods

Sample Collection

A total of 131 peripheral blood samples (including 72 Kyrgyz and 59 Tajik individuals) were collected ([fig. 1](#) and [table 1](#)). The Kyrgyz individuals are from Xinjiang Uyghur Autonomous Region of northwestern China ($n = 30$) and Kyrgyzstan ($n = 42$). The Tajik individuals are from Xinjiang Uyghur Autonomous Region of northwestern China (i.e., $n = 39$; 19 Sarikoli Tajiks and 20 Wakhi Tajiks) and from Dushanbe of Tajikistan ($n = 20$) ([table 1](#)). The study has been reviewed and approved by the Life Sciences Ethics Committee of Kunming Institute of Zoology, Chinese Academy of Sciences (SMKX-20160102-02). The protocol and data release policy have been reviewed and approved by the Human Genetic Resources Information, Backup Platform (*BF2021062906855), and Ministry of Science and Technology of the People's Republic of China (2021BAT3236). The sample collection was conducted by Xinjiang Medical University, Institute of Molecular Biology and Medicine, and E.N. Pavlovsky Institute of Zoology and Parasitology, Academy of Sciences of Republic of Tajikistan and for China, Kyrgyzstan, and Tajikistan, respectively. Before sample collection, the project was explained to the community leaders and participants and then got their permissions. Written informed consents were obtained. The adult participants were recruited without referring to any healthy information. The ethnic information for individual was self-declared. All the collected samples are anonymous. The private information such as detailed place of residence is masked.

Whole-Genome Sequencing and SNP Calling

A total of 131 genomes were sequenced with a sequencing depth $>30\times$ using Illumina HiSeq X Ten and Illumina HiSeq 2000 platform. The published high-depth WGS data for 121 individuals including: 20 Persians from Kerman in Iran ([Charati et al. 2019](#)), 20 Kashmiris, 18 Baltis, two Punjabis from Pakistan, 20 Pamiri Tajiks from Tajikistan ([Yang et al. 2021](#)), and 41 Tibetans from Tibet of China ([Lu et al. 2016](#); [Yang et al. 2018](#)) were included in joint SNP genotyping. Access to the 20 Persian WGS is permitted by the Data Access Committee (Contact Prof. Ali Esmailizadeh, aliesmaili@uk.ac.ir). SNP calling was performed following GATK best practices ([McKenna et al. 2010](#); [Van der Auwera et al. 2013](#)). BAM files were generated by mapping the reads to the reference genome hg19 with BWA-MEM (v0.7.12; [Li and Durbin 2009](#)) with default parameters. Picard tools (<https://broadinstitute.github.io/picard/>, v 1.119) were used to mask duplications. We performed local realignment around indels and base quality score recalibration with GATK by using dbSNP ([Sherry et al. 2001](#)) and the 1000 Genomes Project (1000 Genomes Project Consortium 2010; [Mills et al. 2006](#)) as known sites file. GATK HaplotypeCaller module was used to joint genotype SNPs and indels via local *de novo* assembly of haplotypes in an active region for 252

genomes. Then, we used GenotypeGVCFs with -allSites option to get the vcf file containing all site presented in the reference genome. Finally, we performed GATK variant quality score recalibration (VQSR) to get high-quality variants sites by using HapMap (The International HapMap 3 Consortium 2010), dbSNP (Sherry et al. 2001), the 1000 Genomes Project, and Omni (1000 Genomes Project Consortium 2010) as training resource. Annotations considered during VQSR were: Coverage, QualByDepth, FisherStrand, MappingQualityRankSumTest, and ReadPos RankSumTest. A ‘tranche’ level of 99.9 during VQSR was used to mark SNPs for filtering.

SNP Annotation

We first extracted high-quality variants from our 131 new genomes. Then we used ANNOVAR (Wang et al. 2010) to annotate the SNPs by examining their functional consequence on genes, and to compare with the variants reported in the 1000 Genomes Project (1000 Genomes Project Consortium 2015), and dbSNP (Sherry et al. 2001).

Data Merging

We also merged 252 genomes with 1,372 ancient individuals and 832 modern individuals (see [supplementary table S1, Supplementary Material](#) online for additional details). The genotype data of 832 modern and 1,085 ancient individuals were downloaded from Allen Ancient DNA Resource <https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data> (version 50.0). And the genotype data of 287 ancient genomes were obtained from the published articles (Gneccchi-Ruscone et al. 2021; Kumar et al. 2021, 2022; Zhang et al. 2021). PLINK v2.0 (Purcell et al. 2007; Chang et al. 2015) was used to merge and filter the data. We only kept the SNPs which exist in 1,240K panel and the merged data set was filtered with criteria: the SNPs with genotyping success rate <70% were removed and ancient individuals with <15,000 SNPs were excluded.

PCA and ADMIXTURE

Based on the merged data set containing ancient DNA, we carried out PCA using smartpca package of EIGENSOFT (v6.1.4) with the option (lsqproject: YES; Patterson et al. 2006) to project the ancient individuals onto the present-day Eurasian genomic variations. ADMIXTURE (Alexander et al. 2009) was used to infer ancestral components of the individuals with K from 2 to 10. For each K, we repeated the analysis ten times with different random seeds and picked the run with the lowest cross-validation error.

f Statistics

In order to test the shared genetic drift between Central Asian (Kyrgyz and Tajik) and other modern or ancient populations, we used qp3pop (v410) in ADMIXTOOLS (v5.1; Patterson et al. 2012) to calculate outgroup f_3 statistics (Reich et al. 2009) in the form of f_3 (Mbuti; Kyrgyz/Tajik,

X), where the outgroup was the central African Mbuti and X stood for ancient or modern Eurasian population. We also computed outgroup f_3 (Mbuti; Tarim_EMBA1, X) to examine for the relatedness between the Tarim_EMBA1 and modern non-African populations. We used qpDstat (v755) in ADMIXTOOLS (v5.1; Patterson et al. 2012) to calculate f_4 -statistics (Reich et al. 2009) in the form of $f_4(X, Y; \text{Tajik/Kyrgyz, Mbuti})$, where X and Y represented ancient or modern Eurasian populations.

Modeling Population Admixture History

We referred to the strategies described before (Kumar et al. 2021) to estimate mixture proportions and P -value for a target population (i.e., Kyrgyz/Tajik and other ancient/modern Central Asian populations) as a combination of sources populations by exploiting shared genetic drift with a set of outgroups with qpAdm (v810; Haak et al. 2015) as implemented in ADMIXTOOLS (Patterson et al. 2012). First, the distal modeling was conducted with the pre-Copper age or relatively genetically isolated populations as potential sources. The outgroups included Mbuti, Russia_Ust_Ishim_HG, Russia_Kostenki14, EEHG, Russia_MA1_HG, Belgium_UP_GoyetQ116_1, Czech_Vestonice16, Israel_Natufian_published, Italy_North_Villabruna_HG, Spain_ElMiron, and China_Tianyuan. The probable sources were referred to Anatolia_N, Iran_GanjDareh_N, Tarim_EMBA1, WEHG, and Russia_Shamanka_Eneolithic. Then, we used more recent populations of Bronze Age, Iron Age, and Historical Era as potential sources in the proximal modeling. We followed the recommend strategy (Patterson et al. 2022) to adjust the left source and right outgroup populations in the modeling. The outgroups included Mbuti, Israel_Natufian_published, Russia_Ust_Ishim_HG, Russia_MA1_HG, Russia_Kostenki14, Italy_North_Villabruna_HG, China_Tianyuan, CHG, Anatolia_N, WEHG, EEHG, and Mixe. According to the results of f statistics, we set the probable source populations including Russia_Samara_EBA_Yamnaya, Russia_Afanasiovo, Russia_Andronovo, BMAC, Russia_MLBA_Sintashta, Central_Steppe_MLBA, Russia_Afanasiovo, Turkmenistan_IA, Tarim_EMBA1, and Mongolia_Xiongnu_o1 for the Tajiks. The probable source populations for the Kyrgyz, Uyghur, Uzbek, Kazakh, and Turkmen populations were Russia_Shamanka_EBA, Mongolia_Xiongnu_o1, LateMed_Khitans, Xinj_HE2, Xinj_HE3, Xinj_HE4, Xinj_HE6, Kyrgyzstan_TianShan_Saka, Saka_TianShan_600BCE, Kazakhstan_Central_Saka, and Turkmenistan_IA. We searched mixture models with all possible combinations of source populations by fixing the ‘right’ populations (allsnps: YES). For the JEZK_IA2, Tarim_EMBA2, Turkmenistan_IA, Saka_TianShan_600BCE, Xinj_HE3, and Xinj_BA3, we used the source populations as described previously (Damgaard et al. 2018; Narasimhan et al. 2019; Gneccchi-Ruscone et al. 2021; Zhang et al. 2021; Kumar et al. 2022). We considered the qpAdm model with $P > 0.05$ to be acceptable and models

with $0.01 < P < 0.05$ to be marginally acceptable (Kumar et al. 2022). We applied qpWave (Patterson et al. 2012; Reich et al. 2012; v410) to check the minimum number of streams of ancestry required in modeling four Tajik and two Kyrgyz and other modern Central Asian populations (i.e., Kazakh, Uyghur, Uzbek, and Turkmen).

Admixture Dating

We dated the admixture events in the Tajik and Kyrgyz populations with DATES v.753 (Narasimhan et al. 2019). The parameters were set with the options binsize: 0.001, maxdis: 1.0, runmode: 1, qbin: 10, and lovalfit: 0.45. For the Sarikoli Tajik and Pamiri Tajik, we used three references: BMAC and Andronovo, BMAC and Tarim_EMBA1, BMAC and Mongolia_Xiongnu_o1, to test the admixture events. In order to reduce the effect of sample size on admixture dating, we grouped Russia_Andronovo and Kazakhstan_Andronovo together as group Andronovo in DATES analysis. For the Wakhi Tajik and Dushanbe Tajik, we set BMAC and Mongolia_Xiongnu_o1 as the reference. For Kyrgyz, we assigned Mongolia_Xiongnu_o1 and Xinj_HE3 as the reference. We assume a generation time of 29 years (Fenner 2005).

Runs of Homozygosity

We applied the hapROH (Ringbauer et al. 2021) methods using the Python library hapROH v. 0.4a1 with default parameters to characterize ROH of the ancient and modern populations.

mtDNA and Y-Chromosomal Analysis

mtDNA sequences of Kyrgyz and Tajik individuals were extracted from the mtDNA SNP calling results with the cutoff value of heteroplasmy as 0.2 (Peng et al. 2018). By using HaploGrep 2 (Weissensteiner et al. 2016), a total of 131 mtDNA sequences were assigned into specific haplogroups, which were further checked by using MitoTool (<http://mitotool.kiz.ac.cn>; Fan and Yao 2013). The mtDNA haplogroup nomenclature was referred to PhyloTree (<http://phylotree.org/>; Build 17; van Oven and Kayser 2009).

We performed quality control (Peng et al. 2014) for the Y-chromosomal SNPs of 77 males extracted from the WGS data. The Y-chromosomal haplogrouping was conducted by using yHaplo (Poznik 2016), which was also checked with HaploGrouper (Jagadeesan et al. 2021). The Y-chromosomal haplogroup nomenclature was referred to ISOGG Y-DNA Haplogroup Tree (v2016; https://isogg.org/tree/2016/ISOGG_YDNA_Version_History16.html).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors are grateful to all volunteers involving in samplings. They thank He-Qun Liu and Yao-Ming Li for their

assistance in sample collection. This study was supported by grants from the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA20040100 and XDA20090000), the National Natural Science Foundation of China (31301026), Science and Technology Department of Xinjiang Uygur Autonomous Region (201491188), the Bureau of Science and Technology of Yunnan Province, and Spring City Plan: the high-level talent promotion and training project of Kunming. M.S.P. appreciates the support from the Youth Innovation Promotion Association of the Chinese Academy of Sciences. This work was supported by Research Center for Ecology and Environment of Central Asia and the Animal Branch of the Germplasm Bank of Wild Species, Chinese Academy of Sciences (the Large Research Infrastructure Funding).

Author Contributions

Y.P.Z., M.S.P., L.F.C., and A.A.A. designed research; X.S., J.L., W.F.X., A.A.A., N.T.A., M.E.A., K.B.I., and W.K.Y. collected the samples; S.S.D. and X.S. conducted experiments; S.S.D., M.S.P., and X.S. analyzed the data; S.S.D. and M.S.P. wrote the paper; Y.P.Z., X.S., J.L., X.C., W.K.Y., M.S.W., A.S., W.C., and Y.G.Y. revised the paper; Q.K.S. and X.Y.Y. provided technical assistance. All authors read and approved the final manuscript.

Data Availability

The 131 new-genome sequences reported in this study are deposited in the Genome Sequence Archive for human (<https://ngdc.cncb.ac.cn/gsa-human>) in National Genomics Data Center, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences, under accession number HRA001331 and HRA001332.

References

- The International HapMap 3 Consortium. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**(7311):52–58.
- 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**(7319):1061–1073.
- 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**(7571):68–74.
- Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**(9):1655–1664.
- Allentoft ME, Sikora M, Sjögren KG, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlström T, Vinner L, et al. 2015. Population genomics of Bronze Age Eurasia. *Nature* **522**(7555):167–172.
- Almarri MA, Haber M, Lootah RA, Hallast P, Al Turki S, Martin HC, Xue Y, Tyler-Smith C. 2021. The genomic history of the Middle East. *Cell* **184**(18):4612–4625.e14.
- Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J, et al. 2020. Insights into

- human genetic variation and population history from 929 diverse genomes. *Science* **367**(6484):eaay5012.
- Bregel Y. 2003. *An historical atlas of Central Asia*. Leiden: Brill Academic Publishers.
- Chaix R, Quintana-Murci L, Hegay T, Hammer MF, Mobasher Z, Austerlitz F, Heyer E. 2007. From social to genetic structures in Central Asia. *Curr Biol*. **17**(1):43–48.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**(1):7.
- Charati H, Peng MS, Chen W, Yang XY, Jabbari Ori R, Aghajanzpour-Mir M, Esmailzadeh A, Zhang YP. 2019. The evolutionary genetics of lactase persistence in seven ethnic groups across the Iranian plateau. *Hum Genomics*. **13**(1):7.
- Chen FH, An CB, Dong GH, Zhang DJ. 2017. Human activities, environmental changes, and rise and decline of silk road civilization in Pan-Third Pole region. *Bull Chin Acad Sci*. **32**(9): 967–975.
- Comas D, Calafell F, Mateu E, Pérez-Lezaun A, Bosch E, Martínez-Arias R, Clarimon J, Facchini F, Fiori G, Luiselli D, et al. 1998. Trading genes along the silk road: mtDNA sequences and the origin of Central Asian populations. *Am J Hum Genet*. **63**(6):1824–1838.
- Damgaard PB, Marchi N, Rasmussen S, Peyrot M, Renaud G, Korneliusen T, Moreno-Mayar JV, Pedersen MW, Goldberg A, Usmanova E, et al. 2018. 137 ancient human genomes from across the Eurasian steppes. *Nature* **557**(7705):369–374.
- Dani AH, Masson VM. 1992. *History of civilizations of Central Asia Volume I: the dawn of civilization: earliest times to 700 B.C.* Paris: UNESCO Publishing.
- de Barros Damgaard P, Martiniano R, Kamm J, Moreno-Mayar JV, Kroonen G, Peyrot M, Barjamovic G, Rasmussen S, Zacho C, Baimukhanov N, et al. 2018. The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* **360**(6396):eaar7711.
- Fan L, Yao YG. 2013. An update to MitoTool: using a new scoring system for faster mtDNA haplogroup determination. *Mitochondrion* **13**(4):360–363.
- Feng Q, Lu Y, Ni X, Yuan K, Yang Y, Yang X, Liu C, Lou H, Ning Z, Wang Y, et al. 2017. Genetic history of Xinjiang's Uyghurs suggests Bronze Age multiple-way contacts in Eurasia. *Mol Biol Evol*. **34**(10):2572–2582.
- Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol*. **128**(2):415–423.
- Findley CV. 2004. *The Turks in world history*. New York: Oxford University Press.
- Frachetti MD. 2012. Multiregional emergence of mobile pastoralism and nonuniform institutional complexity across Eurasia. *Curr Anthropol*. **53**(1):2–38.
- Gnecchi-Ruscone GA, Khussainova E, Kahbatkyzy N, Musralina L, Spyrou MA, Bianco RA, Radzeviciute R, Martins NFG, Freund C, Iksan O, et al. 2021. Ancient genomic time transect from the Central Asian Steppe unravels the history of the Scythians. *Sci Adv*. **7**(13):eabe4414.
- Guarino-Vignon P, Marchi N, Bendezu-Sarmiento J, Heyer E, Bon C. 2022. Genetic continuity of Indo-Iranian speakers since the Iron Age in southern Central Asia. *Sci Rep*. **12**(1):733.
- Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**(7555):207–211.
- Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, Zegura SL. 2001. Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol*. **18**(7):1189–1203.
- Han KX. 1993. *The collected papers about the racial anthropological study of the ancient Silk Road inhabitants*. Urumqi: Xinjiang People's Publishing House.
- Harmatta J, Puri BN, Etemadi GF. 1994. *History of civilizations of Central Asia volume II: the development of sedentary and nomadic civilizations: 700 B.C. to A.D. 250*. Paris: UNESCO Publishing.
- Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A genetic atlas of human admixture history. *Science* **343**(6172):747–751.
- Ioannidis AG, Blanco-Portillo J, Sandoval K, Hagelberg E, Miquel-Poblete JF, Moreno-Mayar JV, Rodríguez-Rodríguez JE, Quinto-Cortés CD, Auckland K, Parks T, et al. 2020. Native American gene flow into Polynesia predating Easter Island settlement. *Nature* **583**(7817):572–577.
- Jagadeesan A, Ebenesersdóttir SS, Guðmundsdóttir VB, Thordardóttir EL, Moore KHS, Helgason A. 2021. HaploGroup: a generalized approach to haplogroup classification. *Bioinformatics* **37**(4):570–572.
- Järve M, Saag L, Scheib CL, Pathak AK, Montinaro F, Pagani L, Flores R, Guellil M, Saag L, Tambets K, et al. 2019. Shifts in the genetic landscape of the western Eurasian Steppe associated with the beginning and end of the Scythian dominance. *Curr Biol*. **29**(14): 2430–2441.e10.
- Jeong C, Balanovsky O, Lukianova E, Kahbatkyzy N, Flegontov P, Zaporozhchenko V, Immel A, Wang CC, Ixan O, Khussainova E, et al. 2019. The genetic history of admixture across inner Eurasia. *Nat Ecol Evol*. **3**(6):966–976.
- Jeong C, Wang K, Wilkin S, Taylor WTT, Miller BK, Bemmman JH, Stahl R, Chiorelli C, Knolle F, Ulziibayar S, et al. 2020. A dynamic 6,000-year genetic history of Eurasia's Eastern Steppe. *Cell* **183**(4): 890–904.e29.
- Kivisild T, Saag L, Hui R, Biagini SA, Pankratov V, D'Atanasio E, Pagani L, Saag L, Rootsi S, Mägi R, et al. 2021. Patterns of genetic connectedness between modern and medieval Estonian genomes reveal the origins of a major ancestry component of the Finnish population. *Am J Hum Genet*. **108**(9):1792–1806.
- Krzewińska M, Kilińc GM, Juras A, Koptekin D, Chyleński M, Nikitin AG, Shcherbakov N, Shuteleva I, Leonova T, Kraeva L, et al. 2018. Ancient genomes suggest the eastern Pontic-Caspian steppe as the source of western Iron Age nomads. *Sci Adv*. **4**(10):eaat4457.
- Kumar V, Bennett EA, Zhao D, Liang Y, Tang Y, Ren M, Dai Q, Feng X, Cao P, Yang R, et al. 2021. Genetic continuity of Bronze Age ancestry with increased steppe-related ancestry in late Iron Age Uzbekistan. *Mol Biol Evol*. **38**(11):4908–4917.
- Kumar V, Wang W, Zhang J, Wang Y, Ruan Q, Yu J, Wu X, Hu X, Wu X, Guo W, et al. 2022. Bronze and Iron Age population movements underlie Xinjiang population history. *Science* **376**(6588):62–69.
- Kuz'mina EE, Mallory JP. 2007. *The origin of the Indo-Iranians*. Leiden: Brill.
- Li Y. 2021. Agriculture and palaeoeconomy in prehistoric Xinjiang, China (3000–200 BC). *Veg Hist Archaeobot*. **30**(2):287–303.
- Li H, Cho K, Kidd JR, Kidd KK. 2009. Genetic landscape of Eurasia and "admixture" in Uyghurs. *Am J Hum Genet*. **85**(6):934–937.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754–1760.
- Lu D, Lou H, Yuan K, Wang X, Wang Y, Zhang C, Lu Y, Yang X, Deng L, Zhou Y, et al. 2016. Ancestral origins and genetic history of Tibetan highlanders. *Am J Hum Genet*. **99**(3):580–594.
- Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**(7624):201–206.
- Mallory JP, Mair VH. 2000. *The Tarim mummies: ancient China and the mystery of the earliest peoples from the West*. London: Thames & Hudson.
- Marchi N, Winkelbach L, Schulz I, Bami M, Hofmanová Z, Blöcher J, Reyna-Blanco CS, Diekmann Y, Thiéry A, Kapopoulou A, et al. 2022. The genomic origins of the world's first farmers. *Cell* **185**(11):1842–1859.e18.
- Martínez-Cruz B, Vitalis R, Ségurel L, Austerlitz F, Georges M, Thiéry S, Quintana-Murci L, Hegay T, Aldashev A, Nasyrova F, et al. 2011.

- In the heartland of Eurasia: the multilocus genetic landscape of Central Asian populations. *Eur J Hum Genet.* **19**(2):216–223.
- Massilani D, Skov L, Hajdinjak M, Gunchinsuren B, Tseveendorj D, Yi S, Lee J, Nagel S, Nickel B, Deviese T, *et al.* 2020. Denisovan ancestry and population history of early East Asians. *Science* **370**(6516):579–583.
- Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, *et al.* 2015. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**(7583):499–503.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al.* 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**(9):1297–1303.
- Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE. 2006. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16**(9):1182–1190.
- Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, Lazaridis I, Nakatsuka N, Olalde I, Lipson M, *et al.* 2019. The formation of human populations in South and Central Asia. *Science* **365**(6457):eaat7487.
- Nei M, Roychoudhury AK. 1993. Evolutionary relationships of human populations on a global scale. *Mol Biol Evol.* **10**(5):927–943.
- Ning C, Wang CC, Gao S, Yang Y, Zhang X, Wu X, Zhang F, Nie Z, Tang Y, Robbeets M, *et al.* 2019. Ancient genomes reveal Yamnaya-related ancestry and a potential source of Indo-European speakers in Iron Age Tianshan. *Curr Biol.* **29**(15):2526–2532.e4.
- Orlando L, Allaby R, Skoglund P, Der Sarkissian C, Stockhammer PW, Ávila-Arcos MC, Fu Q, Krause J, Willerslev E, Stone AC, *et al.* 2021. Ancient DNA analysis. *Nat Rev Methods Primers.* **1**(1):1–26.
- Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, Clemente F, Hudjashov G, DeGiorgio M, Saag L, *et al.* 2016. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* **538**(7624):238–242.
- Palstra FP, Heyer E, Austerlitz F. 2015. Statistical inference on genetic data reveals the complex demographic history of human populations in Central Asia. *Mol Biol Evol.* **32**(6):1411–1424.
- Patterson N, Isakov M, Booth T, Büster L, Fischer CE, Olalde I, Ringbauer H, Akbari A, Cheronet O, Bleasdale M, *et al.* 2022. Large-scale migration into Britain during the Middle to Late Bronze Age. *Nature* **601**(7894):588–594.
- 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**(3):1065–1093.
- Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet.* **2**(12):e190.
- Peng MS, He JD, Fan L, Liu J, Adeola AC, Wu SF, Murphy RW, Yao YG, Zhang YP. 2014. Retrieving Y chromosomal haplogroup trees using GWAS data. *Eur J Hum Genet.* **22**(8):1046–1050.
- Peng MS, Xu W, Song JJ, Chen X, Sulaiman X, Cai L, Liu HQ, Wu SF, Gao Y, Abdulloevich NT, *et al.* 2018. Mitochondrial genomes uncover the maternal history of the Pamir populations. *Eur J Hum Genet.* **26**(1):124–136.
- Pérez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, Martínez-Arias R, Clarimón J, Fiori G, Luiselli D, Facchini F, *et al.* 1999. Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. *Am J Hum Genet.* **65**(1):208–219.
- Poznik GD. 2016. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. *bioRxiv*, 088716.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, *et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* **81**(3):559–575.
- Quintana-Murci L, Chaix R, Wells RS, Behar DM, Sayar H, Scozzari R, Rengo C, Al-Zahery N, Semino O, Santachiara-Benerecetti AS, *et al.* 2004. Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. *Am J Hum Genet.* **74**(5):827–845.
- Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW, Jr, Orlando L, Metspalu E, *et al.* 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**(7481):87–91.
- Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N, *et al.* 2012. Reconstructing Native American population history. *Nature* **488**(7411):370–374.
- Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. *Nature* **461**(7263):489–494.
- Ringbauer H, Novembre J, Steinrücken M. 2021. Parental relatedness through time revealed by runs of homozygosity in ancient DNA. *Nat Commun.* **12**(1):5425.
- Sahakyan H, Hooshiar Kashani B, Tamang R, Kushniarevich A, Francis A, Costa MD, Pathak AK, Khachatryan Z, Sharma I, van Oven M, *et al.* 2017. Origin and spread of human mitochondrial DNA haplogroup U7. *Sci Rep.* **7**:46044.
- Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**(1):308–311.
- Shriner D. 2018. Re-analysis of whole genome sequence data from 279 ancient Eurasians reveals substantial ancestral heterogeneity. *Front Genet.* **9**:268.
- Unterländer M, Palstra F, Lazaridis I, Pilipenko A, Hofmanová Z, Gross M, Sell C, Blöcher J, Kirsanow K, Rohland N, *et al.* 2017. Ancestry and demography of descendants of Iron Age nomads of the Eurasian Steppe. *Nat Commun.* **8**:14615.
- Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, *et al.* 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics.* **43**(1110):11.10.11–11.10.33.
- van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat.* **30**(2):E386–394.
- Wang W, Ding M, Gardner JD, Wang Y, Miao B, Guo W, Wu X, Ruan Q, Yu J, Hu X, *et al.* 2021. Ancient Xinjiang mitogenomes reveal intense admixture with high genetic diversity. *Sci Adv.* **7**(14):eabd6690.
- Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**(16):e164.
- Wang CC, Yeh HY, Popov AN, Zhang HQ, Matsumura H, Sirak K, Cheronet O, Kovalev A, Rohland N, Kim AM, *et al.* 2021. Genomic insights into the formation of human populations in East Asia. *Nature* **591**(7850):413–419.
- Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt HJ, Kronenberg F, Salas A, Schönherr S. 2016. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**(W1):W58–63.
- Wells RS, Yuldasheva N, Ruzibakiev R, Underhill PA, Evseeva I, Blue-Smith J, Jin L, Su B, Pitchappan R, Shanmugalakshmi S, *et al.* 2001. The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc Natl Acad Sci U S A.* **98**(18):10244–10249.
- Yang XY, Dai SS, Liu HQ, Peng MS, Zhang YP. 2018. The uncertainty of population relationship and divergence time inferred by the multiple sequentially Markovian coalescent model. *J Hum Genet.* **63**(6):775–777.
- Yang XY, Rakha A, Chen W, Hou J, Qi XB, Shen QK, Dai SS, Sulaiman X, Abdulloevich NT, Afanasevna ME, *et al.* 2021. Tracing the genetic legacy of the Tibetan Empire in the Balti. *Mol Biol Evol.* **38**(4):1529–1536.
- Yao YG, Kong QP, Wang CY, Zhu CL, Zhang YP. 2004. Different matrilineal contributions to genetic structure of ethnic groups

- in the Silk Road region in China. *Mol Biol Evol.* **21**(12): 2265–2280.
- Yunusbayev B, Metspalu M, Metspalu E, Valeev A, Litvinov S, Valiev R, Akhmetova V, Balanovska E, Balanovsky O, Turdikulova S, *et al.* 2015. The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genet.* **11**(4):e1005068.
- Zerjal T, Wells RS, Yuldasheva N, Ruzibakiev R, Tyler-Smith C. 2002. A genetic landscape reshaped by recent events: Y-chromosomal insights into Central Asia. *Am J Hum Genet.* **71**(3):466–482.
- Zhang F, Ning C, Scott A, Fu Q, Bjørn R, Li W, Wei D, Wang W, Fan L, Abuduresule I, *et al.* 2021. The genomic origins of the Bronze Age Tarim Basin mummies. *Nature* **599**(7884): 256–261.
- Zhang XY, Wei D, Wu Y, Nie Y, Hu YW. 2016. Carbon and nitrogen stable isotope ratio analysis of Bronze Age humans from the Xiabandi cemetery, Xinjiang, China: implications for cultural interactions between the East and West. *Chin Sci Bull.* **61**(32): 3509–3519.