












Ancestral Origins and Admixture History of Kazakhs

Chang Lei ^{1,†} Jiaojiao Liu ^{1,†} Rui Zhang ² Yuwen Pan ² Yan Lu ³ Yang Gao ¹
Xixian Ma ² Yajun Yang ¹ Yaqun Guan ⁴ Dolikun Mamatyusupu ⁵ Shuhua Xu ^{1,3,*}

¹State Key Laboratory of Genetic Engineering, Human Phenome Institute, Zhangjiang Fudan International Innovation Center, Center for Evolutionary Biology, School of Life Sciences, Department of Liver Surgery and Transplantation, Liver Cancer Institute, Zhongshan Hospital, Fudan University, Shanghai 200032, China

²Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China

³Ministry of Education Key Laboratory of Contemporary Anthropology, Fudan University, Shanghai 201203, China

⁴Department of Biochemistry and Molecular Biology, Preclinical Medicine College, Xinjiang Medical University, Urumqi 830011, China

⁵College of the Life Sciences and Technology, Xinjiang University, Urumqi 830046, China

[†]These authors contributed equally to this work.

*Corresponding author: E-mail: xushua@fudan.edu.cn.

Associate editor: Maria C. Ávila-Arcos

Abstract

Kazakh people, like many other populations that settled in Central Asia, demonstrate an array of mixed anthropological features of East Eurasian (EEA) and West Eurasian (WEA) populations, indicating a possible scenario of biological admixture between already differentiated EEA and WEA populations. However, their complex biological origin, genomic makeup, and genetic interaction with surrounding populations are not well understood. To decipher their genetic structure and population history, we conducted, to our knowledge, the first whole-genome sequencing study of Kazakhs residing in Xinjiang (KZK). We demonstrated that KZK derived their ancestries from 4 ancestral source populations: East Asian (~39.7%), West Asian (~28.6%), Siberian (~23.6%), and South Asian (~8.1%). The recognizable interactions of EEA and WEA ancestries in Kazakhs were dated back to the 15th century BCE. Kazakhs were genetically distinctive from the Uyghurs in terms of their overall genomic makeup, although the 2 populations were closely related in genetics, and both showed a substantial admixture of western and eastern peoples. Notably, we identified a considerable sex-biased admixture, with an excess of western males and eastern females contributing to the KZK gene pool. We further identified a set of genes that showed remarkable differentiation in KZK from the surrounding populations, including those associated with skin color (*SLC24A5*, *OCA2*), essential hypertension (*HLA-DQB1*), hypertension (*MTHFR*, *SLC35F3*), and neuron development (*CNTNAP2*). These results advance our understanding of the complex history of contacts between Western and Eastern Eurasians, especially those living or along the old Silk Road.

Key words: Kazakhs, Eurasia, population structure, genetic admixture, next-generation sequencing.

Introduction

Recent advances in genotyping and sequencing technologies have facilitated the genome-wide investigations of human genetic variations, providing new insights into population history and genotype–phenotype relationships. International collaborative projects and regional efforts have produced a detailed catalogue of human DNA variation in populations with ancestry from Europe, East Asia, South Asia, West Africa, and the Americas (The-HUGO-Pan-Asian-SNP-Consortium 2009; Tishkoff et al. 2009; Xu et al. 2009; McEvoy et al. 2010; Lachance et al. 2012; Schlebusch et al. 2012). However, although

Central Asia is a vast territory that has been crucial in human history due to its strategic location (Comas et al. 2004), populations in this region have not generally been included and have been largely underrepresented in similar efforts worldwide.

Kazakhs are the second largest Muslim group in Central Asia, and were previously the most influential of the various Central Asian ethnic groups. While the majority of Kazakhs currently reside in Kazakhstan, over 1.25 million are situated in Xinjiang, an area located in far-western China and crossed by the Silk Road—an essential corridor linking East Asia with Central Asia and Europe. Analogous to many human populations that have settled in Central

Received: December 15, 2023. **Revised:** April 29, 2024. **Accepted:** July 02, 2024

© The Author(s) 2024. 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 reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Open Access

Asia, Kazakh people residing in Xinjiang exhibit a diverse range of mixed anthropological features of East Eurasians (EEA) and West Eurasians (WEA) (Xu et al. 2008; Xu and Jin 2008; Feng et al. 2017; Pan et al. 2022; Ning et al. 2023). This suggests a potential scenario of biological admixture between already differentiated EEA and WEA populations. The distinctive Kazakh ethnicity was developed throughout the 15th and 16th centuries. However, Kazakhs' complex biological origin and genomic makeup, migration history, and genetic admixture with surrounding populations over the past millennium are not well understood. A previous study on Altaian Kazakhs revealed that their mtDNA gene pool was comprised of approximately equal proportions of East and West Eurasian haplogroups (Gokcumen et al. 2008). A study based on ancient mtDNA data showed that all samples retrieved from Kazakh archaeological sites before the 13th to 7th century BC belonged to European lineages, with later arrival of East Eurasian sequences that coexisted with the previous West Eurasian genetic substratum (Lalueza-Fox et al. 2004). Previous genetic studies based on blood groups (Fiori et al. 2000), mtDNA (Comas et al. 1998), and Y-chromosome variability (Perez-Lezaun et al. 1999; Wells et al. 2001; Zerjal et al. 2002) have all demonstrated that the present-day populations of Central Asia are genetically extremely heterogeneous, with the presence of well-differentiated EEA and WEA lineages in all populations studied. However, these previous studies, which relied almost exclusively on Y-DNA and mtDNA, could be significantly influenced by genetic drift and sampling biases. Furthermore, individual admixtures could not be investigated and evaluated based on the information of Y-DNA and mtDNA because they are haploid and uniparentally transmitted. A full analysis of the genetic structure of modern Kazakhs using genome-wide data would help decipher the origin of the present-day Kazakh population and shed light on the understanding of complex patterns of admixtures between Westerners and East Asians.

Ninety-nine percent of Kazakhs in China live north of the Tianshan Mountains in Xinjiang, the northern part of China. They are primarily concentrated in the Ili Kazakh Autonomous Prefecture, Balikun Kazakh Autonomous County, and Mubi Kazakh Autonomous County, with a small number living in Gansu and Qinghai. However, little research has been performed on the ancestral origins and admixture history of the Kazakh population in Xinjiang, and the genetic relationship between Kazakh populations in Xinjiang and other Kazakh populations remains ambiguous. Specifically, previous studies based on Y-DNA and mtDNA could not reveal their admixture history at a fine scale.

In this study, we performed the first whole-genome sequencing of the Kazakh population (KZK) residing in northern Xinjiang, China. Using this unprecedented data set, we comprehensively assessed the genomic diversity of KZK and further inferred the genetic origin and admixture history of KZK. Our findings are expected to advance the understanding of the complex admixture history of KZK.

Results

Genetic Affinity of KZK in the Context of Global Populations

To gain insights into the genetic diversity and population history of the Kazakh people, we collected genome-wide single nucleotide polymorphism data of Kazakhs (KZK) from North Xinjiang, including whole-genome microarray data of 213 samples and whole-genome deep sequencing data of 28 samples (see [Materials and Methods](#)). To investigate the general pattern of association between KZK and other populations on a global scale, we analyzed the genome-wide data of KZK and available data of contemporary populations to understand the genomic diversity of KZK and its relationship with other populations. The results of principal component analysis (PCA) showed that the genetic coordinates of Eurasian individuals on the PC diagram were highly correlated with their geographical location ([Fig. 1a](#)). The clusters of KZK populations were located at an intermediate position between the Western European (WE), Central Asian Siberian (SIB), East Asian (EA), and South Asian (SA) populations on the PC diagram. The KZK population was closer to the SIB/EA population and slightly further from the WE/SA population on the PC diagram.

An estimation of the unbiased global F_{ST} of KZK with the global population showed that KZK was genetically closest to the SIB population, followed by EA, SA, and WE ([Fig. 1b](#)). The top 4 populations exhibiting the closest genetic affinity to KZK were Uyghur ($F_{ST} = 0.0038$), Kalmyk ($F_{ST} = 0.0047$), Hazara ($F_{ST} = 0.0047$), and Even ($F_{ST} = 0.0075$) ([Fig. 1c](#), [supplementary fig. S3](#), [Supplementary Material](#) online). Notably, the genetic relationship between KZK and the Chuvash population was closer than that with other WE populations, and the Chuvash population was genetically related to the Central Asian populations ([Callaway 2010](#); [Lazaridis et al. 2014](#); [Yunusbayev et al. 2015](#)). The genetic differences between KZK and other populations measured by F_{ST} were also confirmed with PCA and the outgroup f_3 test ([supplementary fig. S4](#), [Supplementary Material](#) online).

Ancestry Composition of KZK

PCA reveals that the genetic distances between KZK and both EEA and WEA are similar, suggesting that EEA and WEA may have contributed genetic components to KZK. This hypothesis is supported by the results of the f_3 test. $f_3(\text{WE, SIB; KZK})$ showed the smallest f_3 value ([supplementary table S3](#), [Supplementary Material](#) online), indicating that KZK received gene flow from Central Asian and WE populations.

To reveal the ancestral compositions of KZK, we performed an ADMIXTURE analysis of KZK with global populations, varying the number of ancestral components K from 2 to 20. The ancestral composition of KZK revealed by ADMIXTURE became more complex as the K value increased; however, we did not observe a unique ancestral composition for KZK ([supplementary fig. S6](#),

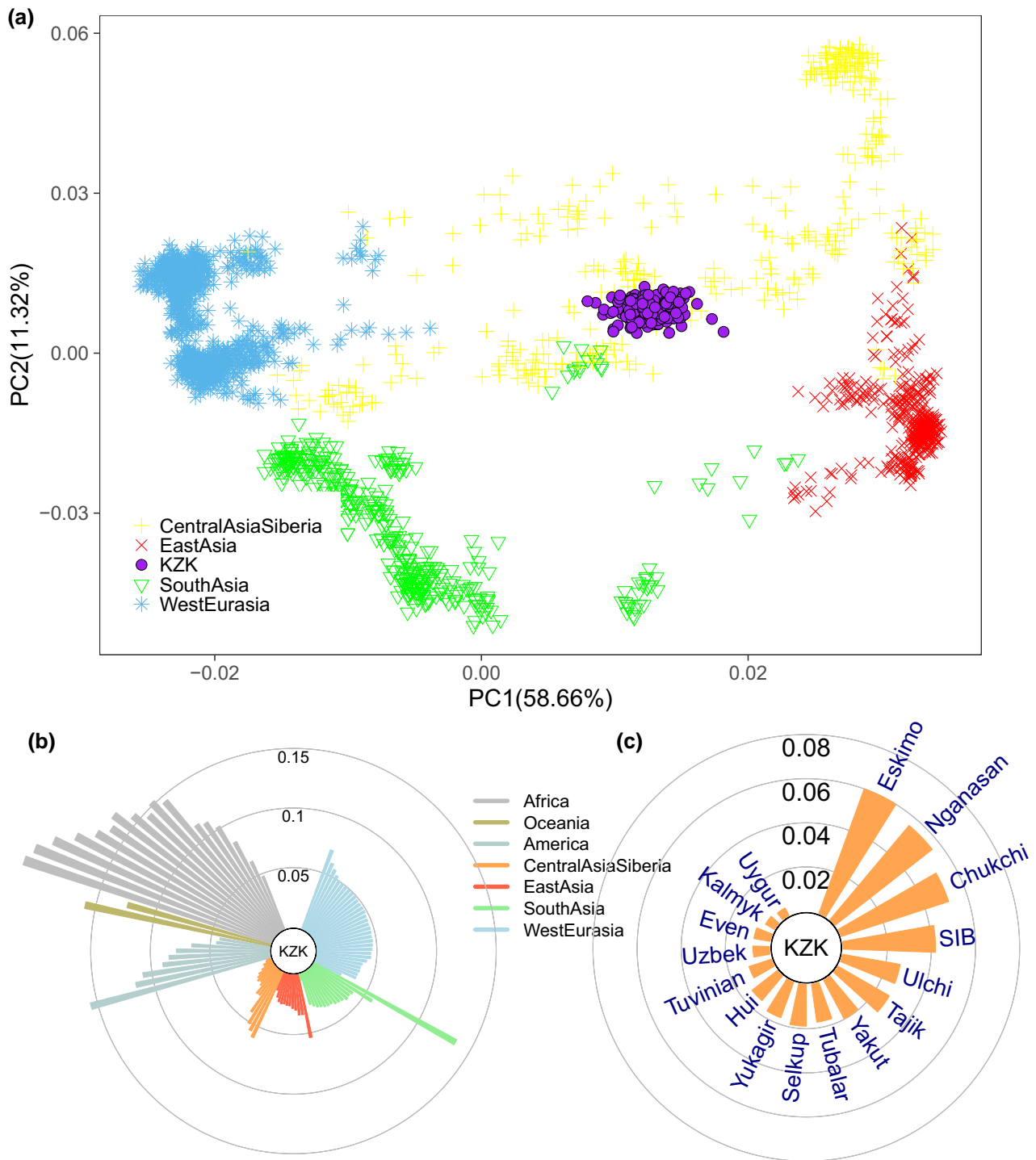


Fig. 1. Genetic affinity of KZK in the context of global populations. a) PCA of 210 KZK samples with other Eurasian populations. Geographical regions from where the samples were collected are labeled on the plot. The number in the bracket represents the variance explained by each PC accounting for the top 10 PCs. b) A fan-like chart shows genetic differences (F_{ST}) between KZK and global populations. Each branch represents a comparison between KZK and 1 of the 210 global populations, and the length is proportional to the F_{ST} value as indicated by gray circles. c) A fan-like chart shows the genetic differences (F_{ST}) between KZK and SIB populations.

Supplementary Material online). Assuming $K = 2$, we found that the ancestry composition of KZK consisted of EEA (~63.5%) and WEA (~36.5%), and these estimations were also confirmed by an f_4 ratio test (supplementary table S4, Supplementary Material online). When the $K = 12$, we got the smallest CV error. However, considering

the increased granularity of ancestral differentiation within continents when the K value exceeds 4, which complicates our subsequent analyses, we have elected to conduct a detailed analysis using a K value of 4 (supplementary fig. S6, Supplementary Material online). At $K = 4$, the main ancestry components of Eurasian populations were summarized

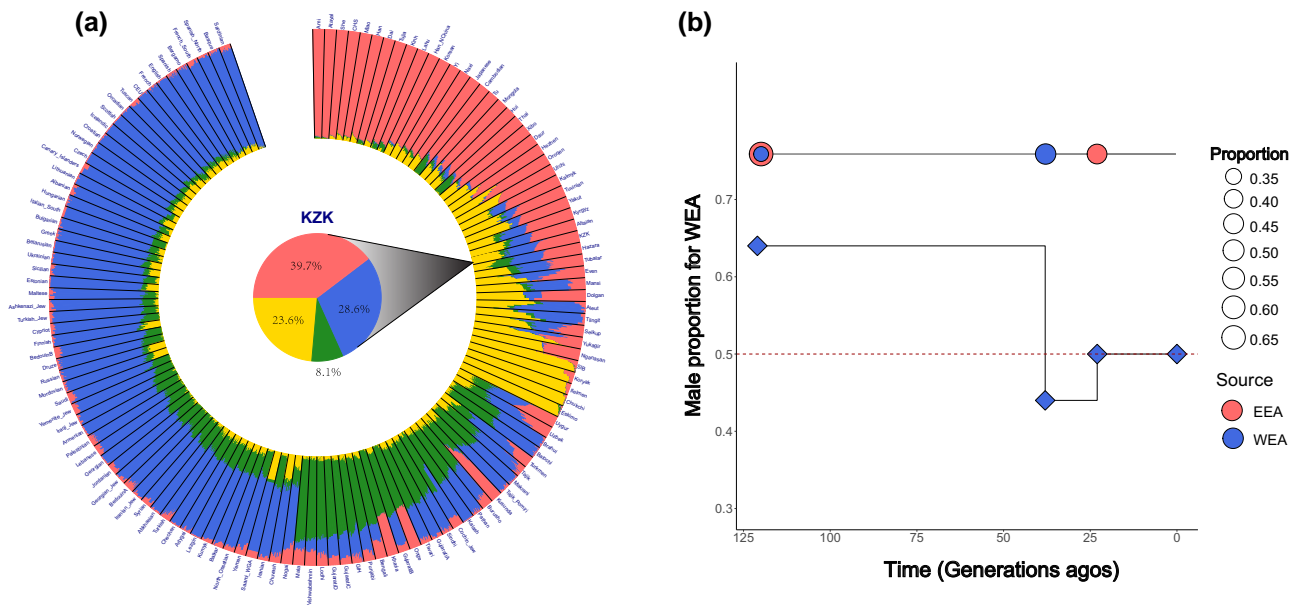


Fig. 2. Ancestral makeup and admixture history of KZK. a) ADMIXTURE result of KZK with other Eurasian populations assuming 4 major ancestral populations ($K = 4$). The result of the population-level admixture of KZK is highlighted and displayed in the pie chart in the center of the circle plot. b) The sex-biased admixture model of KZK inferred by MultiWaverX. Oroqens and Palestinians were used as representatives for Eastern and Western ancestries, respectively. Each circle on the line represents 1 admixture event, in which the color indicates the source ancestry, and the size is proportional to the admixture proportion of the corresponding ancestry. The diamond represents the proportion of male components of West Eurasian.

by 4 sources: EA, WE, SIB, and SA, which was largely consistent with the geographical locations of these populations. In detail, the genetic components of KZK were about 39.7% from EA, 28.6% from WE, 23.6% from SIB, and 8.1% from SA (Fig. 2a). Only a modest variation (5.0%) in admixture proportions was observed among individuals, indicating a lack of recent and continuous gene flow from well-differentiated EEA and WEA populations to present-day Kazakhs.

Furthermore, we applied an admixture history graph (AHG) analysis (Pugach et al. 2016) to infer the admixture events of the 4 ancestral origins of KZK. The analysis conducted by the AHG suggests that it is likely the EA ancestry first contacted with the SIB ancestry, leading to the formation of the EEA ancestry (EA–SIB). Similarly, some clues in the data suggested that the WE ancestry may have mixed with the SA ancestry, resulting in the WEA ancestry (WE–SA; supplementary fig. S7, Supplementary Material online). After that, EA–SIB and WE–SA mixed and formed the gene pool of the present KZK. We also used qpGraph (Patterson et al. 2012) to explore the admixture graph of KZK. The optimal admixture graph inferred by qpGraph (lowest Z score = -0.225) also confirmed the ancestral compositions in KZK and the admixture model of KZK (supplementary fig. S8, Supplementary Material online).

The TreeMix analysis suggested the possibility of gene flow into KZK from multiple sources, including WE, SA, SIB, and EA. When m (migration edge) was 0, the tree structure was divided into 3 clusters, one of which consisted of populations from WEA, another of which consisted of populations from central Eurasia, and the last of which consisted of populations from EEA (supplementary fig. S9a,

Supplementary Material online). KZK were in the eastern Eurasian cluster because of their higher EA ancestry component. When $m = 5$, the TreeMix analysis indicated that KZK received gene flow from WE–SA (supplementary fig. S9b, Supplementary Material online).

Based on the results of the above analysis, we propose that the admixture model of KZK conforms to an “admixture of admixture” scenario, in which SIB and EA first formed the Eastern Eurasian ancestors, whereas WE and SA formed the Western Eurasian ancestor; then, the Western Eurasian ancestor and the Eastern Eurasian ancestor were mixed to form the present-day KZK gene pool.

Sex-Biased Admixture and Admixture Dating

Next, we reconstructed the admixture history of KZK with 4 ancestral components and explored the sex-biased admixture in KZK. We estimated the admixture time of KZK with a sex-biased admixture using MultiWaverX (Zhang et al. 2022). The 2 populations, Oroqen and Palestinian in Human Genome Diversity Project (HGDP; Bergström et al. 2020), were included in the analysis to represent the EEA and WEA ancestries of KZK. The results of MultiWaverX showed that the admixture history of KZK followed a multi-wave admixture model (Fig. 2b). This admixture model was similar to that inferred with qpGraph. We conducted a similar MultiWaverX analysis, employing CHB and CEU from the 1,000 Genomes Project as proxies for EEA and WEA ancestries, respectively. The results paralleled those obtained using Oroqen/Palestinian, reinforcing the multiwave admixture model’s applicability to KZK’s genetic history (supplementary fig. S10, Supplementary Material online).

Our analysis suggested that the eastern and western ancestry of the KZK population experienced at least 3 waves of identifiable admixtures (Fig. 2b). The first admixture occurred ~121 generations ago (95% CI: 118 to 126, SR = 100%), with an east–west ancestry mixing ratio of ~0.64:0.36, with predominantly western males ($P = 0.64$, 95% CI: 0.63 to 0.77, SR = 100%) and eastern females ($P = 0.43$, 95% CI: 0.35 to 0.43, SR = 100%). This admixture in 2000 BP was in agreement with the period reported for mummified humans with European features discovered in Xinjiang (2000 to 1700 BP; Zhang et al. 2021), as well as the Tocharian language (1800 BP), and it might be also associated with the Aryan entry into South Asia (Pathak et al. 2018). They might have brought with them an early version of Sanskrit, proficiency with horses, and a host of new cultural practices such as sacrificial rituals, all of which were the basis of the Hindu/Vedic culture (Narasimhan et al. 2019). This brought a mixture of WEA males as the dominant group.

The second admixture took place ~38 generations ago (95% CI: 36 to 40, SR = 100%) with a mixing ratio of 0.51:0.49 between early KZK and WEA. With a generation of 29 years and considering that the sample was collected in 2014 with the sample age as ~20 years, this admixture took place in 891 (862 to 949) AD, during the Five Dynasties and Ten Kingdoms period. This admixture event may be related to the Turkic migrations to the West. The Turkic migrations brought about the admixture of early KZK males with WEA females ($P = 0.23$, 95% CI: 0.12 to 0.27, SR = 100%). The third admixture event occurred ~23 generations ago (95% CI: 22 to 23, SR = 100%) with a mixing ratio of 0.53:0.47 between early KZK and EEA, 696 (667 to 696) years ago, or AD 1326 (1326 to 1355), when China was in the Yuan dynasty period. This admixture event might be related to the Mongol invasion, in which the early KZK, as “Sephardim,” had a high status during the Yuan Dynasty, bringing an admixture of early KZK males and EEA females ($P = 0.44$, 95% CI: 0.32 to 0.50, SR = 97%). The time of the second and third admixture events was also reconstructed with ALDER (supplementary fig. S11, Supplementary Material online). The ancestry-specific effective population size of KZK showed a decline 30 generations ago, reaching a minimum 17 generations ago, followed by a rapid increase (supplementary fig. S12, Supplementary Material online). These results suggested that the admixture events involving KZK occurred ~30 generations ago, a timeline that aligns with estimates derived from other analytical methods.

Interestingly, after 3 waves of admixtures, sex-biased mating may no longer have been sustained for KZK, potentially aligning with a sex-bias cancellation model as suggested by a previous study using MultiWaverX (Zhang et al. 2022).

The Impact of Admixture on Genome Diversity

Admixtures exert a considerable impact on the genetic diversity of a population. Generally, the introduction of admixture is anticipated to augment the genetic diversity of

a given population. To quantify this, we estimated nucleotide diversity (θ_π), haplotype diversity, number of segregating loci (θ_K), and Tajima's D in KZK (Fig. 3). The results showed that haplotype diversity was significantly higher in KZK compared to the reference populations (Wilcoxon test, $P < 0.001$). Both nucleotide diversity and the number of segregating sites were comparable in KZK and GIH, and significantly higher than those in other reference populations (Wilcoxon test, $P < 0.001$). Tajima's D showed a negatively skewed distribution, and the θ_π in KZK was higher than the θ_K in KZK, suggesting that KZK had specific single-nucleotide variants (SNVs) from diverse ancestral populations. KZK also contained the highest proportion of rare SNVs ($AF < 0.05$), followed by CEU. This observation lends support to the hypothesis that admixture may contribute to the enrichment of rare variants within populations. The effective population size (N_e) of KZK was also larger than that of the reference population (supplementary fig. S13, Supplementary Material online), which also indicated that KZK had higher genetic diversity than the reference population.

Genetic Characteristics and Adaptations in the KZK Population

We expected the allele frequencies of KZK, as an admixed population, to approximate the weighted average of the allele frequencies of their ancestral source populations, where the weights are the percentages inferred from the admixture analysis. We compared the expected allele frequencies of KZK (AF_{exp}) with the observed allele frequency (AF_{obs}) in KZK, which allowed us to obtain loci underlying possible natural selection specific to KZK. We used CHS and CEU as proxy ancestors of KZK and calculated the expected allele frequencies of KZK using the results of ADMIXTURE with $K = 2$ as weights. The site frequency spectrum (SFS) showed that the expected gene frequencies of KZK were aligned with the observed gene frequencies, whereas the gene frequencies of CHS and CEU were significantly different (Fig. 4).

Furthermore, we calculated AFd_e for the whole KZK genome in a 2-way admixture model using CHS and CEU as the reference populations. AFd_e for the whole KZK genome in a 5-way admixture model was estimated using CHS, CEU, GIH, and SIB as reference populations. In the 2-way model, 65% of the SNVs had an $AFd_e < 0.01$, only ~4% of SNVs had an $AFd_e > 0.1$, and ~0.3% of SNVs had an $AFd_e > 0.2$ (supplementary fig. S14, Supplementary Material online). The results of the 4-way model were more conservative relative to the 2-way model, with ~66% of SNVs having an $AFd_e < 0.01$, ~3% of SNVs having an $AFd_e > 0.1$, and ~0.1% of SNVs having an $AFd_e > 0.2$.

We subsequently examined the functional enrichment of SNVs with a significant AFd_e . We expected that SNVs with a significant AFd_e should be less enriched in deleterious mutations, as such mutations are detrimental to inheritance. We determined the deleteriousness of SNVs using the combined annotation-dependent depletion (CADD) score (Kircher et al. 2014), with $CADD \leq 15$ as a modifier,

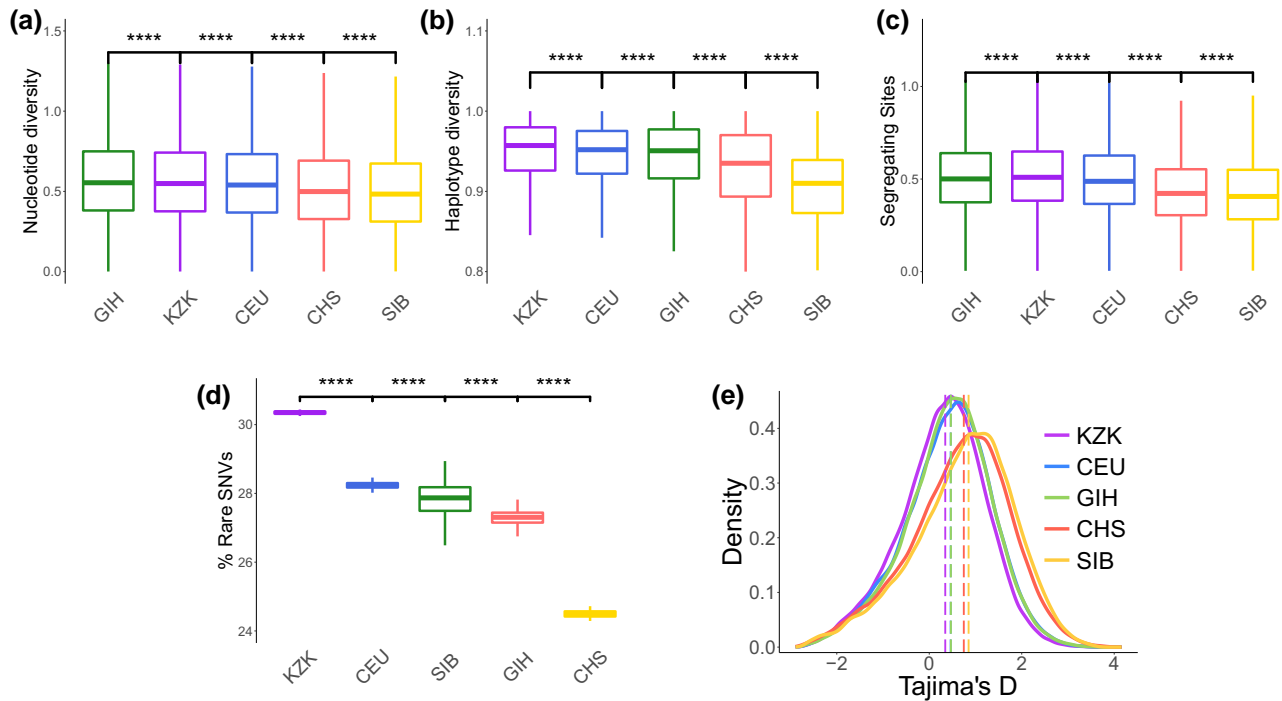


Fig. 3. Genetic diversity of KZK and the reference populations. a) Nucleotide diversity (π /kb) of KZK and the reference populations. b) Haplotype diversity of KZK. c) The number of segregating sites (S /kb). d) Distributions of Tajima's D . All statistics in (a to d) were calculated in the sliding windows with a length of 50 kb and in a step of 25 kb. e) The proportion of rare SNVs ($AF < 0.05$).

$15 < CADD \leq 25$ as a medium, and $CADD > 25$ as a high consequence. The results of both the 2-way model and the 4-way model were consistent with our expectation that the AFd_e was smaller for SNVs with higher functional importance (supplementary fig. S15, Supplementary Material online). The low AFd_e on average of functionally important sites may also be the consequence of their lower derived allele frequency (DAF). We implemented a multinomial logistic regression model as follows:

$$\text{Consequence} \sim \text{DAF} + \text{AFd}_e$$

Next, we employed the variance inflation factor (VIF) in this model. VIF is used to detect the degree of multicollinearity among independent variables in regression analysis. For each explanatory variable X_i , we fit a regression model treating X_i as the dependent variable and the other explanatory variables as independents and calculate the coefficient of determination R_i^2 for that regression. The VIF for each variable is then calculated as

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

where a VIF of 1 indicates no collinearity; a VIF < 5 is generally considered acceptable, and a VIF of 5 or more indicates a severe multicollinearity issue. In our analysis, $\text{VIF}_{\text{DAF}} = 2.93$ and $\text{VIF}_{\text{AFd}_e} = 2.54$, suggesting no severe multicollinearity between these variables. Thus, the low average allele frequency at derived sites (AFd_e) observed in functionally important sites likely resulted from multiple

factors. One contributing factor is their inherently lower DAF. Another possible factor is that mutations at these sites may be detrimental to inheritance, suggesting a negative selective pressure against such mutations, which prevents them from becoming prevalent in the population.

Enrichment analysis was further performed. We further divided the genome into genic, intergenic, transcript, exon, intron, downstream, upstream, 5'-UTR, and 3'-UTR regions according to the GTF file of Ensembl (version 96; Cunningham et al. 2019). The results showed that intergenic regions were enriched with SNVs with a significant AFd_e (supplementary fig. S16, Supplementary Material online).

We estimated the proportion of SNVs with a significant AFd_e in each gene region and ranked these genes accordingly. Modified gene set enrichment analysis (mGSEA; Subramanian et al. 2005; Pan et al. 2022) was used for the pathway search for genes enriched with a significant AFd_e , using the KEGG pathway data set as the reference (Kanehisa et al. 2017). A total of 111 pathways enriched in significant AFd_e genes were identified using the 2-way model. Of these 111 pathways, 61 (55.0%) were associated with "organismal systems" and "human diseases," including pathways related to the immune, digestive, and nervous systems. Another 28 pathways (25.2%) were related to metabolism (supplementary table S5, Supplementary Material online). When the 4-way model was used, a total of 112 pathways rich in significant AFd_e genes were identified, and of these 112 pathways, 62 (55.3%) were related to "organismal systems" and "human diseases," and another 28 pathways (25.0%) were related to metabolism (supplementary table S5, Supplementary Material online).

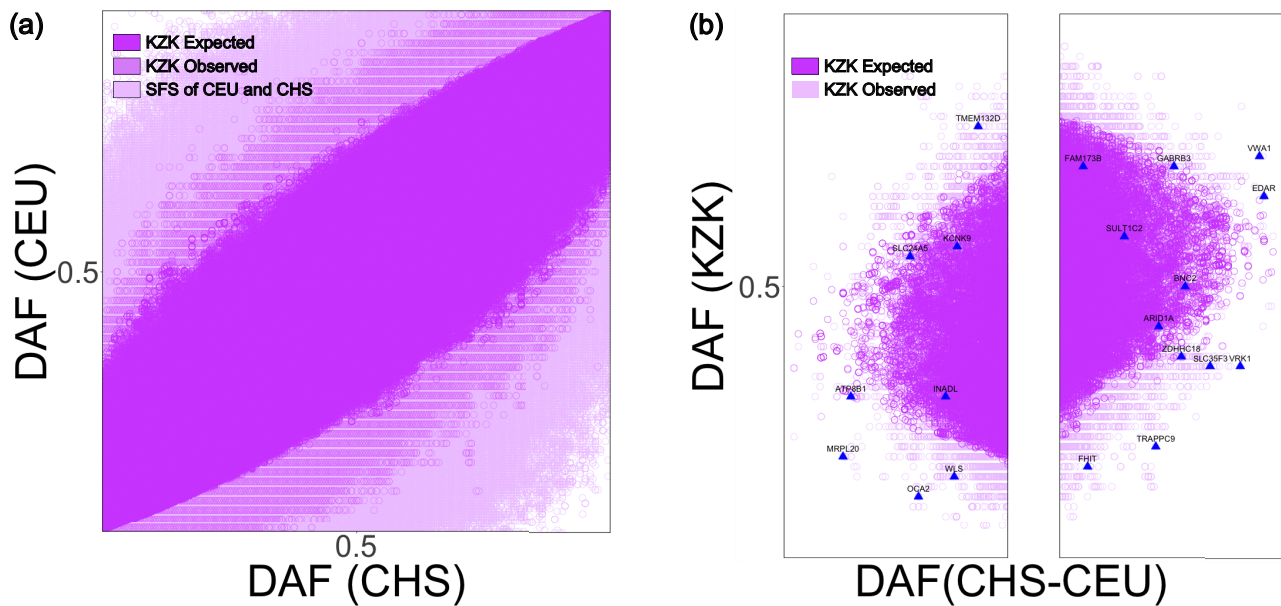


Fig. 4. SFS of KZK. a) The SFS of the frequency profiles of CHS, CEU, KZK, and expected KZK genomes. b) DAF of SNVs in KZK with extreme frequency differences (>0.54 , top 1%) between CHS and CEU.

To broaden our understanding, we also scrutinized the broader evolutionary pressures exerted upon this population. Genomic regions showing strong selection signals were associated with immunity, cancer, neurodevelopment, and cardiovascular diseases, including hypertension and acute myeloid leukemia (AML). The KZK have a homogeneous diet with a low intake of fruits and vegetables and a high intake of smoked meat and wine, thus resulting in a low intake of folic acid, which may lead to a high incidence of hypertension and esophageal cancer (Liu et al. 2010).

We calculated the integrated haplotype score (iHS) of the KZK, Uyghur, CHS, and CEU genomes (Fig. 5a,b, supplementary fig. S17 and table S6, Supplementary Material online). The results of iHS showed that the most significant region of local enrichment signals was in the MHC region of chromosome 6, which is associated with immunity (supplementary fig. S17 and table S6, Supplementary Material online). The gene *HLA-DQB1* is a potential selection gene, and the locus rs28724242 of this gene was annotated in the GWAS catalog (Tragante et al. 2014; Welter et al. 2014) as being associated with essential hypertension. We observed a high concentration of high |iHS| value loci around rs28724242 (supplementary fig. S22, Supplementary Material online), with its allele frequency also being higher in the KZK population compared to CHS and CEU populations (The allele frequency of the G allele at this locus was 0.5482 in KZK, 0.4702 in CHS, and 0.2768 in CEU).

Evidence of selection was obtained from the analysis for the gene *ATXN2*, and its genetically proximate genes, *PTPN11*, *ALDH2*, and *HECTD4*, in both KZK and Uyghur populations with iHS. It turned out to be similar selection patterns for these genes (supplementary fig. S23a, Supplementary Material online). We also performed cross-population extended haplotype homozygosity (XP-EHH)

with Uyghurs, CHS, or CEU as the reference population (Fig. 5c,d, supplementary fig. S18 and table S7, Supplementary Material online). In the $XPEHH_{KZK-CHS}$ and $XPEHH_{KZK-CEU}$, positive selection signals were observed in the *ATXN2* gene region (supplementary fig. S23b,c, Supplementary Material online). However, these signals were not significant in the $XPEHH_{KZK-Uyghur}$ (supplementary fig. S23d, Supplementary Material online), suggesting that there may be a unique selective pressure in this region in Xinjiang, China. Mutations in the *ATXN2* gene were reported to be associated with susceptibility to type I diabetes, obesity, and hypertension (Davalos-Rodriguez et al. 2022). *HECTD4* is involved in the metabolic process of glucose and is also associated with susceptibility to hypertension, with rs2074356 showing significant genetic heterogeneity in the systolic blood pressure (Cho and Jang 2021).

We conducted a population branch statistics (PBS) analysis on the KZK population using the CHS and CEU populations as reference groups. The results revealed strong positive values in the PBS for the *ADAM29* gene located on chromosome 4 (supplementary fig. S19a and table S8, Supplementary Material online). This gene also exhibited strong selection signals in $XPEHH_{KZK-CHS}$, $XPEHH_{KZK-CEU}$, and $XPEHH_{KZK-Uyghur}$. Additionally, we observed a significant enrichment of loci with high allele frequencies in this region (supplementary fig. S24, Supplementary Material online). These findings suggest that this region may be subject to long-term selection in the KZK population. The *ADAM29* gene is potentially associated with fertilization, muscle development, and neurogenesis.

Genetic Adaptations in the KZK Population after Admixture

We focused on the Genetic Adaptations in the Kazakh (KZK) Population after Admixture. Due to the shared

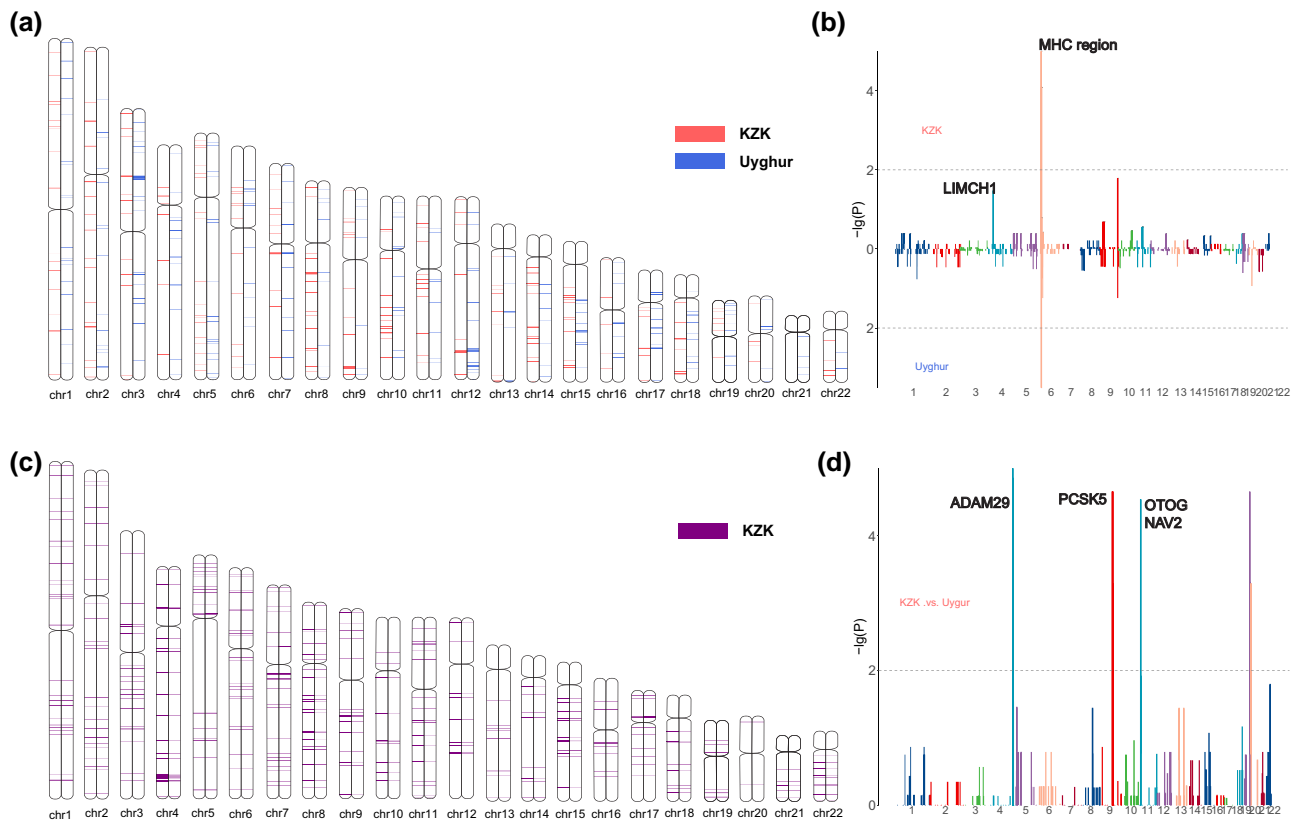


Fig. 5. Selection scans indicated by iHS and XP-EHH signals. a) Chromosome plot for significant components indicated by iHS signals. b) The *P*-value of the local enrichment of sliding windows of 5 Mb and steps of 1 Mb in the KZK and Uyghur genomes with the iHS results. c) Chromosome plot for significant components indicated by XP-EHH signals with Uyghur as the reference population. d) The *P*-value of the local enrichment of sliding windows of 5 Mb and steps of 1 Mb in the KZK genomes with the XP-EHH results.

genetic components between the admixed and ancestral populations, the specific population branch lengths for the admixed population tend to be negative (Pan et al. 2022). Utilizing the negative values of PBS, we can identify genes that exhibit diversity in allele frequencies resulting from admixture. Approximately 86,842 (83.01%) of the windows displayed negative PBS values, supporting the hypothesis that the KZK population is a blend of Eastern and Western ancestries. Regions with $PBS < -0.115$ (the lowest 1%), typically exhibit substantial differences between the reference populations CEU and CHS, and through admixture, these regions contribute to a hybrid gene frequency in the KZK population, referred to as admixture-representative (AR) components.

Some genes indicated that KZK is a CHS-biased population. *SLC24A5* was an AR component in the PBS results (supplementary fig. S19 and table S8, Supplementary Material online), which is associated with skin color. The key SNV, rs1426654 (Stokowski et al. 2007; Adhikari et al. 2016), had a gene frequency of 0.748 in the KZK population, 0.990 in the CHS, and 0.003 in the CEU, showing a strong oriental bias. *LIMS1* and *EDAR* were also AR components of the PBS results, and they are associated with eye thickness, hair morphology, and facial morphology in the oriental population (Fujimoto et al. 2008; Tan et al. 2013; Wu et al. 2018). In the KZK population, strong selection

signals were detected in the *LIMS1* gene upstream of *EDAR*, but no strong selection signals were detected on the *EDAR* gene itself, possibly due to relaxed selection pressures on *EDAR* attributable to the lifestyle and environmental conditions in Western China, similar to the Uyghur population (Pan et al. 2022). *OCA2* was identified as a gene selected away from CEU components (supplementary table S9, Supplementary Material online), and a CHS-biased component in the ancestry-biased (AB) index method with ~78% CHS ancestry (supplementary table S10, Supplementary Material online). It is associated with pigment expression, and in previous reports, *OCA2* also showed selection in eastern populations (Rawofi et al. 2017).

In the analysis of the CEU-biased components using AB index method, we identified 2 regions on chromosome 1 showing significant signatures of selection (supplementary fig. S21 and table S10, Supplementary Material online). One region was associated with the gene *MTHFR*, and the other region was associated with the *SLC35F3* gene. Both *MTHFR* and *SLC35F3* are genes related to metabolism and both are associated with hypertension (Ehret 2010; Zhang et al. 2014; Fan et al. 2016). These results suggested that the hypertension susceptibility gene in KZK may have been introduced from a Western ancestral source and subjected to natural selection locally.

The analysis based on a reconstructed ancestral population, AB_{aCHS} , identified a region of strong natural selection on chromosome 22 (supplementary fig. S20 and table S9, Supplementary Material online), particularly *MRTFA*, which was reported to be associated with megakaryoblast AML (Reed et al. 2021), which is a highly prevalent disease in the Kazakh population (Kulkayeva et al. 2021).

Selection signals were detected in the region encompassing the gene *CNTNAP2* with 5 different methods: PBS , AB_{aCHS} , AB_{aCEU} , $XPEHH_{KZK-CHS}$, and $XPEHH_{KZK-CEU}$. This may be attributed to the extensive length of the *CNTNAP2* gene, which spans ~2.3 Mb across the chromosomal region 7q35 to q36, and to varying selection pressures experienced by different segments of the gene. Specifically, the SNP rs7794745 within this gene has been significantly associated with the development of autism in Bangladeshi children (Uddin et al. 2021). The allele frequency of the T allele at this locus is 0.2857 in the KZK population, 0.4435 in CHS, and 0.3736 in CEU.

Discussion

We have presented a comprehensive characterization of genetic variation in 28 KZK samples, the first whole-genome sequencing study on this ethnic group. With this unprecedented data, we provided new insights into the genetic origin, admixture history, and population structure of KZK. Our results showed that KZK is the most closely related to Central Asian populations compared to other populations globally. Although KZK and Kazakhs of Kazakhstan live in different locations, they cannot be separated into 2 distinct populations in terms of genetic makeup. Part of the reason is the greater genetic heterogeneity of Kazakhs as an admixed population, and part of the reason may be related to the stricter system of endogamy within the KZK.

Four major ancestral components were identified in the KZK population, which were derived from ancestral populations in East Asia, Siberia, Western Eurasia, and South Asia. Modeling admixture histories indicated that these 4 ancestral components came from 2 early admixed populations. The eastern ancestry consisted of ancestral components from East Asia and Siberia. The western ancestry consisted of ancestral components from Western Europe and South Asia. Notably, the admixture among the 4 ancestors of KZK was sex-biased, likely resulting from historical warfare and trade.

Our results suggested a complex scenario of genetic origin, admixture history, and population structure in the Kazakh population. The 3-wave model we proposed here does not necessarily mean there were only 3 admixture events in their history. Rather, it suggested that population admixtures occurred more than once. Moreover, the first wave of admixture was estimated to have occurred 3,150 years ago, which might suggest that the admixture event began to occur 3,150 years ago at the latest. The second wave of admixture was estimated to have occurred 1,083 years ago, which might suggest that the admixture event

started to occur 1,083 years ago at the latest. Furthermore, the admixture time could be underestimated (Leslie et al. 2015). The ancient admixture we identified, dated to over 1,083 years ago, roughly corresponded to the Five Dynasties and Ten Kingdoms period and Turkic migrations. However, the intensive contact between Western and Eastern peoples might have been common in the early Tang Dynasty, according to historical records. Similarly, the time of a recent admixture that occurred nearly 696 years ago as we inferred in this study, corresponding to the Yuan Dynasty, might be also underestimated. According to recorded history, west-east contacts were more frequent during the Mongolian conquests in the 13th and 14th centuries. The concordance would be improved if we adopted a shorter generation time, 25 years, rather than 29 years as a previous study suggested (Fenner 2005). Nonetheless, we believe the history could be more complex than the simplified models we presented in this study.

At the genomic level, population admixtures introduced greater genetic heterogeneity and rare variants to KZK, as 1 ancestral population contributed a certain proportion of rare variants that did not exist in the other ancestral populations. Furthermore, we investigated natural selection in KZK. There was some more significant enrichment of AFd_e in intergenic regions than in genic regions, suggesting that altering gene regulation might also have played an important role in the adaptive evolution of the KZK. Functional enrichment analysis of genes in the KZK genome enriched in SNVs with significant AFd_e suggested that disease risk factors and diet could have shaped the genomic diversity of the KZK population. Genes such as *SLC24A5*, *LIMS1*, and *OCA2* underwent selection in the KZK genome, which was consistent with the eastern-leaning ancestral component of KZK. The results of natural selection analysis indicated the risk of hypertension, AML, and cancer, which could be useful control data for genetic association studies of cardiovascular diseases.

We also explored the genetic relationship between KZK and a Kazakh population in Kazakhstan (KZK.CA; Raghavan et al. 2014). Our analyses indicated that the available data do not support the classification of KZK and KZK.CA as distinct genetic populations. The F_{ST} between the KZK and Kazakhs was 0.0002, smaller than the difference between North and South Han Chinese ($F_{ST} = 0.001$). PCA also did not show significant clustering (supplementary fig. S25a, Supplementary Material online), and the KZK.CA samples were interspersed with the KZK samples. The results of the identity by state (IBS) distribution plot (supplementary fig. S25b, Supplementary Material online) indicated greater genetic heterogeneity in the KZK.CA population compared to KZK. The results of the Gnecci-Ruscone et al. (2021) also indicated higher heterogeneity for the KZK.CA population. The IBS distribution curve for KZK.CA was located between KZK and KZK.CA, indicating that some of the KZK.CA samples highly resembled the KZK samples. Considering that the KZK.CA population is highly heterogeneous, these results can be interpreted to mean that KZK represents a more homogeneous branch of the KZK.CA population. This

interpretation was further supported by the clustering of the KZK.CA and KZK populations inferred by the t-Distributed Stochastic Neighbor Embedding (t-SNE) method (supplementary fig. S25c, Supplementary Material online). The KZK.CA sample enveloped the KZK sample in the t-SNE descending plot, reflecting the higher heterogeneity in KZK.CA that prevented the samples from clustering tightly, whereas the KZK population, being less heterogeneous, formed a cluster. The results of the fine structure also suggested that they could be considered as a single population (supplementary fig. S25d, Supplementary Material online).

In addition, it has been shown that the incidence and mortality of esophageal cancer in the Kazakh population in Xinjiang is very high. Studies on esophageal cancer in the Kazakh population in Xinjiang have found that mutations at the relevant loci of numerous genes were associated with the development of cancer. The incidence rate of hypertension in the Kazakh population in Xinjiang is also the highest among all the Chinese ethnic minorities, and genome-wide data of 241 Kazakhs may serve as useful reference data for genetic association studies of hypertension and other diseases.

Materials and Methods

Populations and Samples

The genomic data of Xinjiang Kazakh (KZK) were sampled from North Xinjiang, including 213 whole-genome microarray samples and 28 whole-genome deep sequencing samples: 2 batches of Illumina Zhonghua microarray (Illu 1, Illu 2) with 52 and 81 samples, respectively; Affymetrix whole-genome SNP microarray (version 6.0) for 47 samples (Affy); Affymetrix genome-wide microarray for 32 samples (CHB2); Illumina genome-wide microarray for 1 sample (1 M) and Illumina HiSeq X10 sequencing platform whole-genome second-generation deep sequencing data (>30×) for 28 Kazakh samples from Xinjiang (NGS), which overlap with 28 Kazakh samples from Illumina Zhonghua (Illu 1; supplementary table S1, Supplementary Material online). Each individual was the offspring of a nonconsanguineous marriage of members of the same nationality within 3 generations. All samples were collected with informed consent and approved by the Biomedical Research Ethics Committee of the Shanghai Institutes for Biological Sciences. Prior to sequencing and analysis, all samples were stripped of personal identifiers (if any existed). The procedures performed in the study were following the ethical standards of the Helsinki Declaration of 1975 (revised in 2000), and approved by the Biomedical Research Ethics Committee of Shanghai Institutes for Biological Sciences (No. ER-SIBS-261408), and further reviewed and approved by the Biomedical Research Ethics Committee of Fudan University (No. FE232771).

Genome Sequencing and Data Processing

Whole-genome sequencing was performed on Illumina HiSeq X Ten with 28 high target coverage (30×) for 150

bp paired-end reads. Each sample for high coverage was run on a unique channel with at least 90 GB of filtered data, while each sample for medium coverage was run on a unique channel with at least 30 GB of filtered data. For high-coverage data, the reads were quality controlled such that 80% of the bases achieved a base quality score of at least 30 (supplementary fig. S1, Supplementary Material online).

The whole-genome second-generation deep sequencing data were generated from the original fastq files, and after merged, and adaptor trimmed, the sequences were aligned with the reference genome (GRCh37) using the Burrows–Wheeler Aligner (Li and Durbin 2010). The indel region re-matching, base quality score recalibration (BQSR), variant calling, and variant quality score recalibration (VQSR) were carried out by Genome Analysis Toolkit (GATK) version 3.8 (McKenna et al. 2010). Only the biallelic SNVs were left for further analysis.

The 28 samples contained a total of 8,389,752 SNVs and 3,163,624 SNVs at the individual level. 0.37% (31,347) of the novel SNVs were obtained based on the dbSNP153 database. 91.6% of the novel SNVs are singleton or doubleton. The variants were annotated using Ensembl VEP (McLaren et al. 2016) with the corresponding VEP-compiled annotation database (v86_GRCh37).

In addition to 28 whole-genome second-generation deep sequencing samples (NGS), 213 whole-genome sequencing microarray samples of KZK were included in the analysis. These 213 samples came from 5 batches (Illu1, Illu2, Affy, CHB2 and 1 M; supplementary table S1, Supplementary Material online). The 28 samples in the NGS overlap with the 28 samples in Illu1. We did not observe any subtle batch effect between high-coverage (30×) sequencing and whole-genome sequencing microarray data (supplementary fig. S2 and table S2, Supplementary Material online).

Phasing was performed using SHAPEIT2 (Delaneau et al. 2014) for the sequencing data with 2,144 whole-genome second-generation samples in the previous study (Ma et al. 2021), and all parameters are default.

We further imputed 5 microarray data sets of KZK. The reference panel for imputation was merged from the 1,000 Genome Project Phase III Panel (Genomes Project et al. 2015) and the phased data set of 2,144 samples using IMPUTE2 (Howie et al. 2009) with the command “-merge_ref_panels.” If the alleles in the microarray data do not match the alleles in the NGS panel, delete the corresponding loci in the microarray data. Totally 29,122,345 SNVs with overlapping physical positions to our sequencing data set were left in the merged reference panel.

Based on both genome-wide random markers and ancestry informative markers (AIMs), analysis was performed using over 8,389,752 genome-wide single-nucleotide polymorphisms (SNPs) and 600 copy-number polymorphisms (CNPs) in 43 unrelated Kazakhs and made comparisons with 3,165 individuals representing 128 worldwide populations.

Public and Published Data

The Affymetrix Human Origin Array (Lazaridis et al. 2014) contains 600,841 SNVs for 2,345 individuals from 203 populations and it was used for admixture analysis under the context of worldwide population.

The 1,000 Genomes Project Phase III (KGP) data contains 2,504 sequenced samples from 26 populations, including 5 populations each in Western Europe, South Asia, and East Asia, 4 in the Americas, and 7 in Africa. We chose CHS, CEU, and GIH in the KGP data set to respectively represent the ancestral EA, WE, and SA populations that contributed to the formation of KZK, according to the results of ADMIXTURE (Alexander et al. 2009).

Seventy-three SIB samples in the data set of the Estonian Biocentre Human Genome Diversity Panel (EGDP; Pagani et al. 2016) were employed in our analyses to represent the ancestral SIB population. There are 108 SIB samples in the EGDP data set sampled from multiple Siberian populations. We filtered 34 “South Siberia” and 1 “West Siberia” samples due to their complex ancestry makeups.

Two population Oroqen and Palestinian in the Human Genome Diversity Project (HGDP; Bergström et al. 2020) data were used to represent the EEA and WEA ancestries of KZK, including 9 Oroqen sequenced samples and 51 Palestinian sequenced samples.

The Illumina Human660W-Quad v1.0 data for the Kazakh population in Kazakhstan (Raghavan et al. 2014) were used for the analysis of the Genetic Affinity of KZK with Kazakh.

Quality Control

First, we removed 28 samples in Illu1 set that were consistent with the whole-genome second-generation deep sequencing samples. Genetic relatedness for all pairs of KZK samples was estimated using PLINK v1.90 (Chang et al. 2015). We preferentially removed 1 individual with a higher genotype missing rate from each pair with a coefficient of relationship larger than 0.125. This step removed 31 individuals, including 1 sample in Illu1 set and 2 samples in Affy set. After quality control, there were 210 individuals left for further analysis.

Principal Component Analysis

PCA was performed at the individual level using smartPCA from the Eigensoft version 6.1.4 (Price et al. 2006; Patterson et al. 2012). To investigate the fine-scale population structure, we carried out a series of PCA by gradually removing “outliers” on the plot of the first 2 principal components and reanalyzing the remaining samples based on the same set of SNV markers. We removed LD by thinning the SNPs to be at least 150 kb apart, resulting in 26,584 SNPs.

F_{ST} Analysis

Genetic distance within Hui and between Hui and other populations was measured with F_{ST} according to Weir

and Cockerham (1984), which accounts for the difference in the sample size of each population. We randomly chose 10 samples from each population to calculate pairwise F_{ST} . For a population with a sample size of <10 , all the samples were used. We repeated 100 times to calculate the pairwise F_{ST} and confidence interval.

f Statistics

All the f statistics were calculated using ADMIXTOOLS 7.0 (Patterson et al. 2012). We calculated the f_3 test in the form of $f_3(X, Y; KZK)$ where X and Y are present-day populations. If the f_3 was more negative, the reference populations were closer to true ancestral populations.

We calculated outgroup f_3 test in the form of $f_3(X, KZK; Ju\ hoan\ North)$, where X is a non-African population (Raghavan et al. 2014). The higher outgroup f_3 value indicates population X shared more genetic drift with KZK derived from African and suggests a closer relationship to KZK population.

We calculated outgroup f_4 ratio in the form of $f_4(\text{Papuan, Ju\ hoan\ North; KZK, Y})/f_4(\text{Papuan, Ju\ hoan\ North; X, Y})$, where X is the population chosen from East Asia and Y is the population chosen from west Eurasian. The value of f_4 ratio indicates the eastern ancestry proportion of KZK.

We also use qpGraph in ADMIXTOOLS to construct the admixture model of KZK. Models with $|Z\ score| < 3$ are considered acceptable. Central African population Mbuti, the West Eurasian population English, the SA population Mala, the Siberian population Chukchi, and the EA population Han are chosen as reference population. Mbuti was used to root the tree.

Detection of Similar Populations

PCA, IBS distribution, t-SNE (Van der Maaten and Hinton 2008), and fineSTRUCTURE (Lawson et al. 2012) are used to find the genetic affinity of KZK with Kazakh.

PLINK v1.90 is used to calculate the IBS matrix of KZK and Kazakh. If Kazakh and KZK are distantly related, then the IBS curve of KZK–Kazakh will be to the left of the IBS curves of KZK–KZK and Kazakh–Kazakh.

t-SNE is a nonlinear dimensionality reduction method that can retain more information than PCA. We do t-SNE analysis by using the R package “Rtsne,” and all the parameters are set to be the default.

FineSTRUCTURE is used to construct the phylogenetic tree of KZK and Kazakh samples based on the haplotype. All the parameters are set to be the default.

Local Ancestry Inference

RFMIX version 2 (Maples et al. 2013), HAPMIX version 1.2 (Price et al. 2009), and Loter (Dias-Alves et al. 2018) are used for local ancestry inference of KZK.

The results of RFMIX are used for further and IBDNe (Browning et al. 2018) analysis. We run local ancestry inference for KZK under 4-way admixture models. The CHS, CEU, and GIH in KGP are used to represent EA, WE, and

SA ancestries, and the 73 SIB samples in EGDG are used to represent SIB ancestry. We set the admixture time to 200G (-G 200) and set the CRF spacing size to 0.5 cM (-c 0.5 -r 0.5). All the SNVs identified among KZK and the reference populations were used.

The results of HAPMIX are used for further admixture model inference. We run local ancestry inference for KZK under 2-way admixture models. The Oroqen and Palestinian in HGDP are used to represent EEA and WEA ancestries. According to the results of ADMIXTURE at $K = 4$, the proportion of ancestries of SA and WE in Oroqen are both $<1\%$, and the ratio of SIB to EA is similar to that of KZK; the proportion of ancestries of EA and SIB in Palestinian are $<5\%$, and the ratio of WE to SA is similar to that of KZK. KZK includes 28 samples in NGS set and 45 microarray data in Affy set. The parameters THETA and LAMBDA are set to 0.633 (based the results of ADMIXTURE at $K = 4$) and 80.

The results of Loter are used for natural selection analysis. Loter is proven to have high accuracy for ancient admixture (>100 generations) and multiway admixture. We run local ancestry inference for KZK both under 2-way and 4-way models. In the 2-way model, the CHS and CEU in KGP are chosen to represent EAS and EUR. In the 4-way model, The CHS, CEU, and GIH in KGP are used to represent EA, WE, and SA ancestries. The 73 SIB samples in EGDG are used to represent SIB ancestries. Parameters were set to the default in our analyses, and all the SNVs identified among KZK and the reference populations were used.

Admixture

We applied ADMIXTURE version 1.3.0 on the merged data set of Human Origins, Hui, Tajik, Uyghur, KZK, CEU, CHS, GIH, and SIB data, which consist of individuals from 211 populations on the same SNPs as PCA. We run ADMIXTURE by assuming the number of ancestries (K) from 2 to 20. For each K , we repeated the analysis 10 times with different random seeds and picked the run with the highest log-likelihood score to avoid the local minimum. To reveal the genetic makeup of the Hui population, for each K , we identified the ancestral component ($>1\%$) in the KZK population. For each ancestral component, we identified the representative reference individuals and populations.

Admixture Model Inference and Admixture Time Estimation

AHG method (Pugach et al. 2016) was used to explore the sequence of the ancestry admixture events of KZK. We first find the sequence of admixture events for any trio (a combination of 3 ancestries, such as “ABC,” “BCA,” and “CAB”) of the 4 ancestries of KZK. For every combination like “ABC,” we obtained the respective percentages of these ancestries in the KZK samples. And the correlation coefficient is calculated as

$$\rho = |\text{Pearson CC}(\ln(A/B), \ln(C))|,$$

where Pearson CC means Pearson correlation coefficient, \ln means logarithm with base e . For each trio, we choose the combination which has the smallest ρ . Then the full graph was reconstructed based on the ordering of likely configurations. ADMIXTURE result for KZK with other Eurasian populations at $K = 4$ was used to estimate the admixture proportion of each ancestry in KZK samples. We take 20 samples from the KZK sample without putting them back for AHG analysis and we repeated the AHG analysis 5,000 times for each trio.

We also use TreeMix (Pickrell and Pritchard 2012) to detect gene flow and infer the admixture graph. Three or four populations from Africa, East Asia, Western Europe, Central Asia, and Siberia were selected as representatives. The number of migrations (migrate edge, m) is set from 0 to 10. The root set to the African Yoruba population (-r Yoruba). The parameter k is set to 500.

MultiWaverX (Zhang et al. 2022) is used to construct the sex-biased admixture model of KZK and infer the time of admixture events. We use the local ancestry inference results by HAPMIX as input. The segments with length <0.007 Morgan will be discarded (-l 0.007). The number of bootstrapping is 1,000 (-bootstrap 1,000).

We also use ALDER (v1.03; Loh et al. 2013) to estimate the time since admixture for all the KZK samples. ALDER measures the decay of admixture LD, and it could be used to identify the gene flow.

Genetic Diversity

We calculated nucleotide diversity θ_π (Nei and Li 1979), Haplotype diversity H (Buntjer et al. 2005), numbers of segregating sites θ_K (Fu 1995), and Tajima's D (Tajima 1989) for KZK, CEU, CHS, SIB, GIH. An equal number of samples were chosen from each population to avoid the bias caused by the sample size. We divided the whole genome into 50 kb windows in steps of 25 kb. We removed the major histocompatibility complex region (chr6:28477797 to 33448356), which has unusually high genetic diversity. The statistics θ_π , H and θ_K were computed by the software “Theta_D_H.Est” (Pan et al. 2022). We also calculated the proportion of rare SNVs ($AF < 0.5$) for each population.

Estimation of Effective Population Size (N_e)

We explored the demographic history of KZK using LD (McEvoy et al. 2011). We merge the KZK sequencing data and the Human Origin data set, resulting in 554,781 SNPs. R^2 was calculated using PLINK v1.90 to measure LD between SNP pairs. The observed pairwise LD was binned into 1 of 240 recombination distance categories from 0.10 cM up to 0.25 cM with incremental upper boundaries of 0.001 cM.

Ancestry-specific effective population size is inferred by IBDNe. All the KZK including the array samples are used to do this analysis. We use RFMIX to do local ancestry inference with CHS, CEU, GIH, and SIB as the reference population. The process refers to the process on <https://>

github.com/hennlab/AS-IBDNe, all parameters are consistent with the process above.

Genome-Wide Allele Frequency Deviation

We calculated the genome-wide allele frequency deviation from expectation (AFd_e) for KZK. SNVs with $AF < 0.01$ or > 0.99 in both KZK and the reference genome were excluded. To eliminate the potential effect of minor allele frequency (MAF), we grouped all variants into bins of size 0.01 according to the expected MAF (MAF_{exp}). Also, the potential effect of allele frequency differences between the 2 reference populations ($AF_{EAS-EUR}$) was controlled. SNVs were classified as $AF_{EAS-EUR}$ [0.0,0.02], [0.02,0.04], [0.04,0.06], [0.06,0.09], [0.09,0.12], [0.12,0.15], [0.15,0.19], [0.19,0.25], [0.25,0.34], and [0.34,1.0], which also ensures that there are a similar and sufficient number of SNVs within each bin. After removing the null set, we obtained a total of 452 bins. The empirical P -value of each SNV was estimated as the rank within the corresponding bin divided by the total number of SNVs. We used an empirical P -value < 0.01 as the threshold for identifying SNVs with significant AFd_e .

Within-Population Signals of Selection Scan (iHS)

iHS (Voight et al. 2006) was used to detect the signatures of selection scan in KZK and reference populations (CHS and CEU). We computed iHS with selscan v1.20 (Szpiech and Hernandez 2014) for each SNVs.

We divided the whole genome into genetic map-based sliding window of 0.05 cM with steps of 0.025 cM, and calculated the maximum |iHS| observed among the SNVs in each window. Windows containing < 20 SNVs will be removed, remaining 93,288 windows. We ranked each window based on its maximum |iHS|. The empirical P -value of each window is the rank value of that window divided by the total number of windows. We chose windows with an empirical P -value $< 0.5\%$ as significant windows.

Cross-Population Signals of Selection Scan

XP-EHH (Sabeti et al. 2007), PBS (Yi et al. 2010), reconstructed ancestral populations (Jin et al. 2012), and AB index (Pan et al. 2022) were used to detect the signatures of selection scan between KZK and reference populations (CHS and CEU).

We computed XP-EHH with selscan v1.20 for each SNVs. Like his analysis, we divided the whole genome into genetic map-based sliding window of 0.05 cM with steps of 0.025 cM, calculated the maximum |XP-EHH| observed among the SNVs in the window, ranked each window based on its maximum |XP-EHH| and finally the empirical P -value of each window is the rank value of that window divided by the total number of windows. We chose windows with an empirical P -value of $< 0.5\%$ as significant windows ($XPEHH_{KZK-CHS}$ or $XPEHH_{KZK-CEU}$).

PBS value was calculated as

$$PBS = \frac{T_{KZK-CHS} + T_{KZK-CEU} - T_{CHS-CEU}}{2}$$

The T statistic was calculated as

$$T = -\log(1 - F_{ST}),$$

which can be interpreted as the measurement of the divergence between 2 populations.

A positive value of PBS indicates the presence of positive selection. Since admixed populations share genetic components with their ancestral populations, the specific population branch length for the admixed population will be negative. By utilizing PBS, we are able to identify genomic segments that are equally contributed by highly divergent ancestries (Pan et al. 2022). Based on this analysis, we can infer genes that exhibit diversity in allele frequencies resulting from admixture.

We calculated the genome-wide PBS of KZK using a sliding window of length 50 kb and a step size of 25 kb, where we used the average F_{ST} of each SNV in the sliding window when calculating the T statistic. We selected the windows with $PBS > 0.046$ (top 1%) as the potential selection components and the windows with $PBS < -0.115$ (top 1%) as the potential AR components.

Loter was used to infer the local ancestry component for each SNV of the whole KZK genome. Based on the local ancestry inference results for each SNV, we partitioned the KZK genome to reconstruct the ancestral population. By comparing the differences between the reconstructed ancestry population and the reference ancestry population, we can identify the ancestry components of KZK that were subject to selection after admixture. We calculate the AF_{anc} of the reconstructed ancestral population and the AF_{ref} of the reference population, and then calculate the difference between them:

$$AFd_e = |AF_{anc} - AF_{ref}|$$

SNVs with $AF < 0.01$ or $AF > 0.99$ in KZK, CEU, and CHS populations were removed. A total of 3,907,914 SNVs remained. The significance of SNV was estimated by calculating the ranking of AFd_e . To eliminate the effect from the MAF of the reconstructed ancestral population, we assigned all SNVs to bins in steps of 0.01 according to the MAF of the reconstructed ancestral population (MAF_{anc}), and then calculated the AFd_e rank of each SNV in the corresponding bin as its empirical P -value. We selected the SNVs in the top 1% of empirical P -values in each bin as significant SNV. In the CHS and reconstructed ancestral CHS (aCHS) populations, a total of 39,075 SNVs had an empirical P -value < 0.01 ; in the CEU and reconstructed ancestral CEU (aCHS) populations, a total of 39,079 SNVs had an empirical P -value < 0.01 . Like iHS analysis, we divided the whole genome into sliding windows with 50 kb length in steps of 25 kb, calculated the proportion of significant SNVs in the window, then grouped windows to bins by the size of 10 to eliminate the effect of the number of SNVs, and finally calculated the empirical P -value for each window. We chose windows with an empirical P -value $< 0.5\%$ as AB components (AB_{aCEU} and AB_{aCHS}).

AB index was applied to detect AB local components. We divided the whole genome into windows using a sliding window of length 50 kb with a step size of 25 kb, and the AB index of each window was calculated as the geometric mean of the quantiles of AFd_e and local ancestry deviation. We take the highest 0.5% of AB index as the significant window (CEU-biased AB components or CHS-biased AB components).

Mapping Selected Regions to the Gene

Following the identification of genomic windows potentially subject to selection using a sliding window approach, we first amalgamate regions exhibiting overlap. Subsequently, if these potentially selected windows overlap with gene loci as delineated by Ensembl96, they are considered to be associated with the respective genes. For loci under selection, their positioning onto genes is refined based on information from the dbSNP database.

Local Enrichment of the Signals of Selection Scan

Because natural selection signals may have a cascading effect, we performed local enrichment of selection signals. We divided the genome into 5 Mb long sliding windows with a step size of 1 Mb, and then analyzed the size of local enrichment signals on each chromosome. Assuming a Poisson distribution of signal across the chromosome, local enrichment of signal was analyzed by counting the number of segments with significant natural selection signal within each sliding window. The Poisson distribution test was used to calculate the P -value magnitude for each sliding window. We picked windows with BH-corrected P -values < 0.001 , ($\lg P > 4$) as candidates.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank all the participants of this project.

Author Contributions

S.X. conceived and designed the study and supervised the project. D.M., Y.Y., and Y.G. contributed to the sample collection. Y.L. managed data generation. C.L., J.L., R.Z., Y.P., Y.G., and X.M. analyzed the data and drafted the manuscript with the contribution of other authors. S.X. revised the manuscript. All authors discussed the results and implications and commented on the manuscript.

Funding

This work was supported by the National Key Research and Development Program of China (no. 2023YFC2605400), the National Natural Science Foundation of China

(NSFC) grants (32030020, 32288101), the Shanghai Science and Technology Commission Program (23JS1410100), the Office of Global Partnerships (Key Projects Development Fund). The computational work in this study was supported by the CFFF Computing Platform and the Human Phenome Data Center of Fudan University. The funders had no role in the study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest

None declared.

Data Availability

The release of the variants of 241 Kazakh samples by this work is permitted by The Ministry of Science and Technology of the People's Republic of China (permission no. 2023BAT0255) at the National Genomics Data Center (<https://ngdc.cncb.ac.cn>).

References

- Adhikari K, Fontanil T, Cal S, Mendoza-Revilla J, Fuentes-Guajardo M, Chacon-Duque JC, Al-Saadi F, Johansson JA, Quinto-Sanchez M, Acuna-Alonzo V, et al. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nat Commun*. 2016;**7**(1):10815. <https://doi.org/10.1038/ncomms10815>.
- Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;**19**(9):1655–1664. <https://doi.org/10.1101/gr.094052.109>.
- Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. A global reference for human genetic variation. *Nature*. 2015;**526**(7571):68–74. <https://doi.org/10.1038/nature15393>.
- Bergström A, McCarthy SA, Hui RY, Almarri MA, Ayub Q, Danecsek P, Chen Y, Felkel S, Hallast P, Kamm J, et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science*. 2020;**367**(6484):1339. <https://doi.org/10.1126/science.aay5012>.
- Browning SR, Browning BL, Daviglus ML, Durazo-Arvizu RA, Schneiderman N, Kaplan RC, Laurie CC. Ancestry-specific recent effective population size in the Americas. *PLoS Genet*. 2018;**14**(5):e1007385. <https://doi.org/10.1371/journal.pgen.1007385>.
- Buntjer JB, Sorensen AP, Peleman JD. Haplotype diversity: the link between statistical and biological association. *Trends Plant Sci*. 2005;**10**(10):466–471. <https://doi.org/10.1016/j.tplants.2005.08.007>.
- Callaway E. The rise of the genome bloggers. *Nature*. 2010;**468**(7326):880–881. <https://doi.org/10.1038/468880a>.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;**4**(1):7. <https://doi.org/10.1186/s13742-015-0047-8>.
- Cho SB, Jang J. A genome-wide association study of a Korean population identifies genetic susceptibility to hypertension based on sex-specific differences. *Genes (Basel)*. 2021;**12**(11):231–237. <https://doi.org/10.3390/genes12111804>.
- Comas D, Calafell F, Mateu E, Perez-Lezaun A, Bosch E, Martinez-Arias R, Clarimon J, Facchini F, Fiori G, Luiselli D, et al. Trading genes along the silk road: mtDNA sequences and

- the origin of Central Asian populations. *Am J Hum Genet.* 1998;**63**(6):1824–1838. <https://doi.org/10.1086/302133>.
- Comas D, Plaza S, Wells RS, Yuldaseva N, Lao O, Calafell F, Bertranpetit J. Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. *Eur J Hum Genet.* 2004;**12**(6):495–504. <https://doi.org/10.1038/sj.ejhg.5201160>.
- Cunningham F, Achuthan P, Akanni W, Allen J, Amodè MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, et al. Ensembl 2019. *Nucleic Acids Res.* 2019;**47**(D1):D745–D751. <https://doi.org/10.1093/nar/gky1113>.
- Davalos-Rodriguez NO, Rincon-Sanchez AR, Madrigal Ruiz PM, Flores-Alvarado LJ, Lopez-Toledo S, Villafan-Bernal JR, Castro-Juarez CJ, Guzman-Lopez R, Siliceo-Murrieta JL, Ramirez-Garcia SA. VNTR (CAG)_n polymorphism of the ATXN2 gene and metabolic parameters of cardiovascular risk associated with the degree of obesity in the Amerindian population of Oaxaca. *Endocrinol Diabetes Nutr (Engl Ed).* 2022;**69**(1): 15–24. <https://doi.org/10.1016/j.endien.2021.04.001>.
- Delaneau O, Marchini J, Genomes Project C. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun.* 2014;**5**(1):3934. <https://doi.org/10.1038/ncomms4934>.
- Dias-Alves T, Mairal J, Blum MGB. Loter: a software package to infer local ancestry for a wide range of species. *Mol Biol Evol.* 2018;**35**(9):2318–2326. <https://doi.org/10.1093/molbev/msy126>.
- Ehret GB. Genome-wide association studies: contribution of genomics to understanding blood pressure and essential hypertension. *Curr Hypertens Rep.* 2010;**12**(1):17–25. <https://doi.org/10.1007/s11906-009-0086-6>.
- Fan S, Yang B, Zhi X, Wang Y, Wei J, Zheng Q, Sun G. Interactions of methylenetetrahydrofolate reductase C677T polymorphism with environmental factors on hypertension susceptibility. *Int J Environ Res Public Health.* 2016;**13**(6):601. <https://doi.org/10.3390/ijerph13060601>.
- Feng Q, Lu Y, Ni X, Yuan K, Yang Y, Yang X, Liu C, Lou H, Ning Z, Wang Y. Genetic history of Xinjiang's Uyghurs suggests Bronze age multiple-way contacts in Eurasia. *Mol Biol Evol.* 2017;**34**(10):2572–2582. <https://doi.org/10.1093/molbev/msx177>.
- Fenner JN. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol.* 2005;**128**(2):415–423. <https://doi.org/10.1002/ajpa.20188>.
- Fiori G, Facchini F, Ismagulov O, Ismagulova A, Tarazona-Santos E, Pettener D. Lung volume, chest size, and hematological variation in low-, medium-, and high-altitude Central Asian populations. *Am J Phys Anthropol.* 2000;**113**(1):47–59. [https://doi.org/10.1002/1096-8644\(200009\)113:1<47::AID-AJPA5>3.0.CO;2-K](https://doi.org/10.1002/1096-8644(200009)113:1<47::AID-AJPA5>3.0.CO;2-K).
- Fu YX. Statistical properties of segregating sites. *Theor Popul Biol.* 1995;**48**(2):172–197. <https://doi.org/10.1006/tpbi.1995.1025>.
- Fujimoto A, Ohashi J, Nishida N, Miyagawa T, Morishita Y, Tsunoda T, Kimura R, Tokunaga K. A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. *Hum Genet.* 2008;**124**(2): 179–185. <https://doi.org/10.1007/s00439-008-0537-1>.
- Gnecchi-Ruscone GA, Khussainova E, Kahbatkyzy N, Musralina L, Spyrou MA, Bianco RA, Radzeviciute R, Martins NFG, Freund C, Iksan O, et al. Ancient genomic time transect from the Central Asian Steppe unravels the history of the Scythians. *Sci Adv.* 2021;**7**(13):eabe4414. <https://doi.org/10.1126/sciadv.abe4414>.
- Gokcumen O, Dulik MC, Pai AA, Zhadanov SI, Rubinstein S, Osipova LP, Andreenkov OV, Tabikhanova LE, Gubina MA, Labuda D, et al. Genetic variation in the enigmatic Altaian Kazakhs of South-Central Russia: insights into Turkic population history. *Am J Phys Anthropol.* 2008;**136**(3):278–293. <https://doi.org/10.1002/ajpa.20802>.
- Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 2009;**5**(6):e1000529. <https://doi.org/10.1371/journal.pgen.1000529>.
- Jin W, Xu S, Wang H, Yu Y, Shen Y, Wu B, Jin L. Genome-wide detection of natural selection in African Americans pre- and post-admixture. *Genome Res.* 2012;**22**(3):519–527. <https://doi.org/10.1101/gr.124784.111>.
- Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2017;**45**(D1):D353–D361. <https://doi.org/10.1093/nar/gkw1092>.
- Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet.* 2014;**46**(3):310–315. <https://doi.org/10.1038/ng.2892>.
- Kulkayeva GU, Kemaykin VM, Kuttymuratov AM, Burlaka ZI, Saparbay JZ, Zhakhina GT, Adusheva AA, Dosayeva SD. First report from a single center retrospective study in Kazakhstan on acute myeloid leukemia treatment outcomes. *Sci Rep.* 2021;**11**(1):24001. <https://doi.org/10.1038/s41598-021-03559-3>.
- Lachance J, Vernot B, Elbers CC, Ferwerda B, Froment A, Bodo JM, Lema G, Fu W, Nyambo TB, Rebbeck TR, et al. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell.* 2012;**150**(3): 457–469. <https://doi.org/10.1016/j.cell.2012.07.009>.
- Lalueza-Fox C, Sampietro ML, Gilbert MTP, Castri L, Facchini F, Pettener D, Bertranpetit J. Unravelling migrations in the steppe: mitochondrial DNA sequences from ancient Central Asians. *Proc R Soc Lond Ser B-Biol Sci.* 2004;**271**(1542):941–947. <https://doi.org/10.1098/rspb.2004.2698>.
- Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. *PLoS Genet.* 2012;**8**(1): e1002453. <https://doi.org/10.1371/journal.pgen.1002453>.
- Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature.* 2014;**513**(7518):409–413. <https://doi.org/10.1038/nature13673>.
- Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, Hutnik K, Royrvik EC, Cunliffe B, Lawson DJ, et al. The fine-scale genetic structure of the British population. *Nature.* 2015;**519**(7543):309–314. <https://doi.org/10.1038/nature14230>.
- Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2010;**26**(5): 589–595. <https://doi.org/10.1093/bioinformatics/btp698>.
- Liu F, Ma YT, Yang YN, Xie X, Li XM, Huang Y, Ma X, Chen BD, Gao X, Du L. Current status of primary hypertension in Xinjiang: an epidemiological study of Han, Uyghur and Hazakh populations. *Zhonghua Yi Xue Za Zhi.* 2010;**90**(46):3259–3263. <https://pubmed.ncbi.nlm.nih.gov/21223782/>.
- Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics.* 2013;**193**(4):1233–1254. <https://doi.org/10.1534/genetics.112.147330>.
- Ma X, Yang W, Gao Y, Pan Y, Lu Y, Chen H, Lu D, Xu S. Genetic origins and sex-biased admixture of the Huis. *Mol Biol Evol.* 2021;**38**(9): 3804–3819. <https://doi.org/10.1093/molbev/msab158>.
- Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet.* 2013;**93**(2):278–288. <https://doi.org/10.1016/j.ajhg.2013.06.020>.
- McEvoy BP, Lind JM, Wang ET, Moyzis RK, Visscher PM, van Holst Pellekaan SM, Wilton AN. Whole-genome genetic diversity in a sample of Australians with deep Aboriginal ancestry. *Am J Hum Genet.* 2010;**87**(2):297–305. <https://doi.org/10.1016/j.ajhg.2010.07.008>.
- McEvoy BP, Powell JE, Goddard ME, Visscher PM. Human population dispersal “Out of Africa” estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res.* 2011;**21**(6): 821–829. <https://doi.org/10.1101/gr.119636.110>.

- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;**20**(9):1297–1303. <https://doi.org/10.1101/gr.107524.110>.
- McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The ensemble variant effect predictor. *Genome Biol.* 2016;**17**(1):122. <https://doi.org/10.1186/s13059-016-0974-4>.
- Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, Lazaridis I, Nakatsuka N, Olalde I, Lipson M, et al. The formation of human populations in South and Central Asia. *Science.* 2019;**365**(6457):eaat7487. <https://doi.org/10.1126/science.aat7487>.
- Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A.* 1979;**76**(10):5269–5273. <https://doi.org/10.1073/pnas.76.10.5269>.
- Ning Z, Tan X, Yuan Y, Huang K, Pan Y, Tian L, Lu Y, Wang X, Qi R, Lu D. Expression profiles of east–west highly differentiated genes in Uyghur genomes. *Natl Sci Rev.* 2023;**10**(4). <https://doi.org/10.1093/nsr/nwad077>.
- Pagani L, Lawson DJ, Jagoda E, Morseburg A, Eriksson A, Mitt M, Clemente F, Hudjashov G, DeGiorgio M, Saag L, et al. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature.* 2016;**538**(7624):238–242. <https://doi.org/10.1038/nature19792>.
- Pan Y, Zhang C, Lu Y, Ning Z, Lu D, Gao Y, Zhao X, Yang Y, Guan Y, Mamatyusupu D, et al. Genomic diversity and post-admixture adaptation in the Uyghurs. *Natl Sci Rev.* 2022;**9**(3):nwab124. <https://doi.org/10.1093/nsr/nwab124>.
- Pathak AK, Kadian A, Kushniarevich A, Montinaro F, Mondal M, Ongaro L, Singh M, Kumar P, Rai N, Parik J, et al. The genetic ancestry of modern Indus valley populations from Northwest India. *Am J Hum Genet.* 2018;**103**(6):918–929. <https://doi.org/10.1016/j.ajhg.2018.10.022>.
- Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. Ancient admixture in human history. *Genetics.* 2012;**192**(3):1065–1093. <https://doi.org/10.1534/genetics.112.145037>.
- Perez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, Martinez-Arias R, Clarimon J, Fiori G, Luiselli D, Facchini F, et al. Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. *Am J Hum Genet.* 1999;**65**(1):208–219. <https://doi.org/10.1086/302451>.
- Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 2012;**8**(11):e1002967. <https://doi.org/10.1371/journal.pgen.1002967>.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006;**38**(8):904–909. <https://doi.org/10.1038/ng1847>.
- Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* 2009;**5**(6):e1000519. <https://doi.org/10.1371/journal.pgen.1000519>.
- Pugach I, Matveev R, Spitsyn V, Makarov S, Novgorodov I, Osakovsky V, Stoneking M, Pakendorf B. The complex admixture history and recent southern origins of Siberian populations. *Mol Biol Evol.* 2016;**33**(7):1777–1795. <https://doi.org/10.1093/molbev/msw055>.
- Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW, Orlando L, Metspalu E, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature.* 2014;**505**(7481):87–91. <https://doi.org/10.1038/nature12736>.
- Rawoif L, Edwards M, Krithika S, Le P, Cha D, Yang Z, Ma Y, Wang J, Su B, Jin L, et al. Genome-wide association study of pigmented traits (skin and iris color) in individuals of East Asian ancestry. *PeerJ.* 2017;**5**:e3951. <https://doi.org/10.7717/peerj.3951>.
- Reed F, Larsuel ST, Mayday MY, Scanlon V, Krause DS. MRTFA: a critical protein in normal and malignant hematopoiesis and beyond. *J Biol Chem.* 2021;**296**:100543. <https://doi.org/10.1016/j.jbc.2021.100543>.
- Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. Genome-wide detection and characterization of positive selection in human populations. *Nature.* 2007;**449**(7164):913–918. <https://doi.org/10.1038/nature06250>.
- Schlebusch CM, Skoglund P, Sjodin P, Gattepaille LM, Hernandez D, Jay F, Li S, Jongh D, Singleton M, Blum A, et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science.* 2012;**338**(6105):374–379. <https://doi.org/10.1126/science.1227721>.
- Stokowski RP, Pant PV, Dadd T, Fereday A, Hinds DA, Jarman C, Filsell W, Ginger RS, Green MR, van der Ouderaa FJ, et al. A genome-wide association study of skin pigmentation in a South Asian population. *Am J Hum Genet.* 2007;**81**(6):1119–1132. <https://doi.org/10.1086/522235>.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 2005;**102**(43):15545–15550. <https://doi.org/10.1073/pnas.0506580102>.
- Szpiech ZA, Hernandez RD. Selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol.* 2014;**31**(10):2824–2827. <https://doi.org/10.1093/molbev/msu211>.
- Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics.* 1989;**123**(3):585–595. <https://doi.org/10.1093/genetics/123.3.585>.
- Tan J, Yang Y, Tang K, Sabeti PC, Jin L, Wang S. The adaptive variant EDARV370A is associated with straight hair in East Asians. *Hum Genet.* 2013;**132**(10):1187–1191. <https://doi.org/10.1007/s00439-013-1324-1>.
- The-HUGO-Pan-Asian-SNP-Consortium. Mapping human genetic diversity in Asia. *Science.* 2009;**326**(5959):1541–1545. <https://doi.org/10.1126/science.1177074>.
- Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, et al. The genetic structure and history of Africans and African Americans. *Science.* 2009;**324**(5930):1035–1044. <https://doi.org/10.1126/science.1172257>.
- Tragante V, Barnes MR, Ganesh SK, Lanktree MB, Guo W, Franceschini N, Smith EN, Johnson T, Holmes MV, Padmanabhan S, et al. Gene-centric meta-analysis in 87,736 individuals of European ancestry identifies multiple blood-pressure-related loci. *Am J Hum Genet.* 2014;**94**(3):349–360. <https://doi.org/10.1016/j.ajhg.2013.12.016>.
- Uddin MS, Azima A, Aziz MA, Aka TD, Jafrin S, Millat MS, Siddiqui SA, Uddin MG, Hussain MS, Islam MS. CNTNAP2 gene polymorphisms in autism spectrum disorder and language impairment among Bangladeshi children: a case–control study combined with a meta-analysis. *Human Cell.* 2021;**34**(5):1410–1423. <https://doi.org/10.1007/s13577-021-00546-8>.
- Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res.* 2008;**9**(86):2579–2605. <http://jmlr.org/papers/v9/vandermaaten08a.html>.
- Voight BF, Kudravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. *PLoS Biol.* 2006;**4**(3):e72. <https://doi.org/10.1371/journal.pbio.0040072>.
- Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. *Evolution.* 1984;**38**(6):1358–1370. <https://doi.org/10.1111/j.1558-5646.1984.tb05657.x>.
- Wells RS, Yuldashewa N, Ruzibakiev R, Underhill PA, Evseeva I, Blue-Smith J, Jin L, Su B, Pitchappan R, Shanmugalakshmi S,

- et al.* The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc Natl Acad Sci U S A*. 2001;**98**(18): 10244–10249. <https://doi.org/10.1073/pnas.171305098>.
- Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*. 2014;**42**(D1):D1001–D1006. <https://doi.org/10.1093/nar/gkt1229>.
- Wu S, Zhang M, Yang X, Peng F, Zhang J, Tan J, Yang Y, Wang L, Hu Y, Peng Q, *et al.* Genome-wide association studies and CRISPR/Cas9-mediated gene editing identify regulatory variants influencing eyebrow thickness in humans. *PLoS Genet*. 2018;**14**(9): e1007640. <https://doi.org/10.1371/journal.pgen.1007640>.
- Xu S, Huang W, Qian J, Jin L. Analysis of genomic admixture in Uyghur and its implication in mapping strategy. *Am J Hum Genet*. 2008;**82**(4):883–894. <https://doi.org/10.1016/j.ajhg.2008.01.017>.
- Xu S, Jin L. A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. *Am J Hum Genet*. 2008;**83**(3):322–336. <https://doi.org/10.1016/j.ajhg.2008.08.001>.
- Xu S, Yin X, Li S, Jin W, Lou H, Yang L, Gong X, Wang H, Shen Y, Pan X, *et al.* Genomic dissection of population substructure of Han Chinese and its implication in association studies. *Am J Hum Genet*. 2009;**85**(6):762–774. <https://doi.org/10.1016/j.ajhg.2009.10.015>.
- Yi X, Liang Y, Huerta-Sanchez E, Cuo JX, Pool ZX, Xu JE, Jiang X, Vinckenbosch H, Korneliussen N, Korneliussen TS, *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*. 2010;**329**(5987):75–78. <https://doi.org/10.1126/science.1190371>.
- Yunusbayev B, Metspalu M, Metspalu E, Valeev A, Litvinov S, Valiev R, Akhmetova V, Balanovska E, Balanovsky O, Turdikulova S, *et al.* The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genet*. 2015;**11**(4):e1005068. <https://doi.org/10.1371/journal.pgen.1005068>.
- Zerjal T, Wells RS, Yuldasheva N, Ruzibakiev R, Tyler-Smith C. A genetic landscape reshaped by recent events: Y-chromosomal insights into Central Asia. *Am J Hum Genet*. 2002;**71**(3):466–482. <https://doi.org/10.1086/342096>.
- Zhang F, Ning C, Scott A, Fu Q, Bjorn R, Li W, Wei D, Wang W, Fan L, Abuduresule I, *et al.* The genomic origins of the Bronze Age Tarim Basin mummies. *Nature*. 2021;**599**(7884):256–261. <https://doi.org/10.1038/s41586-021-04052-7>.
- Zhang K, Huentelman MJ, Rao F, Sun EI, Corneveaux JJ, Schork AJ, Wei Z, Waalen J, Miramontes-Gonzalez JP, Hightower CM, *et al.* Genetic implication of a novel thiamine transporter in human hypertension. *J Am Coll Cardiol*. 2014;**63**(15):1542–1555. <https://doi.org/10.1016/j.jacc.2014.01.007>.
- Zhang R, Ni X, Yuan K, Pan Y, Xu S. MultiWaverX: modeling latent sex-biased admixture history. *Brief Bioinform*. 2022;**23**(5): bbac179. <https://doi.org/10.1093/bib/bbac179>.