Genetic History of Xinjiang's Uyghurs Suggests Bronze Age Multiple-Way Contacts in Eurasia

Qidi Feng,[†,1,2] Yan Lu,[†,1] Xumin Ni,[†,3] Kai Yuan,[†,1,2] Yajun Yang,[†,4] Xiong Yang,[1,2] Chang Liu,[1,2] Haiyi Lou,[1] Zhilin Ning,[1,2] Yuchen Wang,[1,2] Dongsheng Lu,[1,2] Chao Zhang,[1,2] Ying Zhou,[1,2] Meng Shi,[1,2] Lei Tian,[1,2] Xiaoji Wang,[1,2,5] Xi Zhang,[1,2,5] Jing Li,[1] Asifullah Khan,[1] Yaqun Guan,[6] Kun Tang,[*,1] Sijia Wang,[*,1,7] and Shuhua Xu[*,1,2,5,7]

[1]Chinese Academy of Sciences Key Laboratory of Computational Biology, Max Planck Independent Research Group on Population Genomics, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, CAS, Shanghai, China

[2]University of Chinese Academy of Sciences, Beijing, China

[3]Department of Mathematics School of Science, Beijing Jiaotong University, Beijing, China

[4]School of Life Sciences, Fudan University, Shanghai, China

[5]School of Life Science and Technology, ShanghaiTech University, Shanghai, China

[6]Department of Biochemistry and Molecular Biology, Preclinical Medicine College Xinjiang Medical University, Urumqi, China

[7]Collaborative Innovation Center of Genetics and Development, Shanghai, China

[†]These authors contributed equally to this work.

*__Corresponding authors:__ E-mails: xushua@picb.ac.cn; wangsijia@picb.ac.cn; tangkun@picb.ac.cn.

Associate editor: Daniel Falush

Abstract

The Uyghur people residing in Xinjiang, a territory located in the far west of China and crossed by the Silk Road, are a key ethnic group for understanding the history of human dispersion in Eurasia. Here we assessed the genetic structure and ancestry of 951 Xinjiang's Uyghurs (XJU) representing 14 geographical subpopulations. We observed a southwest and northeast differentiation within XJU, which was likely shaped jointly by the Tianshan Mountains, which traverses from east to west as a natural barrier, and gene flow from both east and west directions. In XJU, we identified four major ancestral components that were potentially derived from two earlier admixed groups: one from the West, harboring European (25–37%) and South Asian ancestries (12–20%), and the other from the East, with Siberian (15–17%) and East Asian (29–47%) ancestries. By using a newly developed method, *MultiWaver*, the complex admixture history of XJU was modeled as a two-wave admixture. An ancient wave was dated back to ~3,750 years ago (ya), which is much earlier than that estimated by previous studies, but fits within the range of dating of mummies that exhibited European features that were discovered in the Tarim basin, which is situated in southern Xinjiang (4,000–2,000 ya); a more recent wave occurred around 750 ya, which is in agreement with the estimate from a recent study using other methods. We unveiled a more complex scenario of ancestral origins and admixture history in XJU than previously reported, which further suggests Bronze Age massive migrations in Eurasia and East-West contacts across the Silk Road.

Key words: **Uyghurs, Xinjiang, Eurasia, population structure, genetic admixture, SNP.**

Introduction

Xinjiang, previously known as "Xiyu" or the "Western Region", is a vast territory located in northwestern China, spanning over 1.6 million square kilometers. Xinjiang has been crucial in human history due to its strategic location. It is crossed by the well-known route of the historical Silk Road (Mair 1995) and borders the countries of Afghanistan, India, Kazakhstan, Kyrgyzstan, Tajikistan, Pakistan, Russia, and Mongolia. However, investigations on human genetic diversity in these regions are limited. With a population size of >10 million, the Uyghur people used to be one of the most influential ethnic groups in Xinjiang. They are believed to be descendants of the most ancient Turkic tribes with mixed Caucasian and East Asian ancestries (Balfour 1985). Therefore, the Uyghur people are a key ethnic group for understanding both the history of recent genetic exchanges between Eastern and Western Eurasian people, and the impact of genetic admixture on population genetic diversity in Eurasia. Nonetheless, the origins and history of the Uyghur people remain poorly understood and thus have been the topic of intense debates. Uyghur historians view the Uyghurs as the original inhabitants of Xinjiang, having occupied the area for 6,400–9,000 years (Tursun 2008). Well-preserved Tarim mummies exhibiting European features were discovered and dated 4,000–2,000 years ago (ya), indicating the migration of people of European ancestry into Xinjiang at the beginning of the Bronze Age (Li et al. 2015). According to ancient Chinese historical texts (e.g., *Book of Wei*), the Uyghurs in Xinjiang

doi:10.1093/molbev/msx177 Advance Access publication June 8, 2017

originated from the Tiele tribes, a confederation of Turkic people that was established after the disintegration of the Xiongnu confedeacy, and migrated to Xinjiang from Mongolia after the collapse of the Uyghur Khaganate during the ninth century. Numerous contemporary Western scholars, however, do not consider modern Uyghurs to be of direct linear descent from the old Uyghur Khaganate of Mongolia. Rather, they consider them to be descendants of a number of people, one of which are the ancient Uyghurs (Henders 2006; Millward 2007; Millward and Perdue 2004).

Genetic studies based on mtDNA (Yao et al. 2004), Y chromosome (Wells et al. 2001), and autosomal data (Hui Li and Kidd 2009; Xu et al. 2008; Xu and Jin 2008; Xu and Jin 2009; Xu et al. 2009) have shown that modern Uyghurs are an admixed population with ancestries that were mainly derived from Eastern and Western Eurasian people. However, previous studies either focused on a single geographical population, or a limited number of markers, and thus the global population structure and admixture of the Xinjiang's Uyghurs (XJU) remain unclear. In the present study, we genotyped 951 Uyghur samples from 14 geographical regions (prefectures) on the Illumina OminiZhongHua and Affymetrix Genome-Wide Human SNP 6.0 array (fig. 1A). Our data well represented all regions where the Uyghur people reside. With this unprecedented data set, we attempted to comprehensively characterize the ancestral makeup of XJU, uncover their origins, and reconstruct their admixture history.

Results

Genetic Affinity and Population Structure of XJU

To understand the general patterns of relatedness between XJU and worldwide populations, we analyzed genome-wide data of XJU together with 203 worldwide populations from the Human Origins data set (Lazaridis et al. 2014). Reference populations were classified into five groups representing major geographical regions: Africa, America, Central Asia/Siberia (SIB), East Asia (EA), Oceania, South Asia (SA), and West Eurasia (WE) (see Materials and Methods). Principal component analysis (PCA) showed that XJU lay along the axis between groups from WE and EA (supplementary fig. S1, Supplementary Material online). After removing populations outside of Eurasia from PCA, the XJU samples were surrounded by populations from SIB, EA, SA, and WE (fig. 1B). Among these neighboring populations, XJU was most closely related to the Central/South Asian populations, followed by the EA/WE populations. Interestingly, the Turkish were not the group with the closest relationship with XJU ($F_{ST} = 0.0180$), whereas the Hazara ($F_{ST} = 0.0098$) and Uzbek ($F_{ST} = 0.0130$) groups showed significantly less differentiation from XJU (supplementary fig. S2, Supplementary Material online).

We further explored the substructures of XJU using PCA and observed that individuals from the same region tended to cluster together (supplementary fig. S3, Supplementary Material online). In particular, samples from Southwest Xinjiang, e.g., Hotan, Kizilsu, Kaxgar, and Bortala, could be separated from those from Northeast Xinjiang, e.g., Turpan, Changji, and Kumul. On the other hand, samples from some regions distributed sparsely, such as those from Aksu and Bayingolin1 exhibited higher diversity in these regions. The largest pairwise genetic distance (as measured by F_{ST}) was between Kumul and Bortala, which are located in East and West Xinjiang, respectively, although these are not the most distant geographic pair (supplementary fig. S4, Supplementary Material online). Despite the high diversity, the extent of genetic differentiation between different regional populations was still significantly smaller than that between XJU and the nonXJU populations (supplementary table S1, Supplementary Material online).

Our analysis showed that PC1 had a significant correlation with longitude ($P = 8.58 \times 10^{-4}$) but not latitude ($P = 0.992$) (supplementary fig. S5, Supplementary Material online). Longitude and latitude could jointly explain 68.0% of the total variance of genetic differentiation on PC1 among regional populations. XJU samples were thus classified into "Northeast" and "Southwest" clusters, which correspond to their geographical distribution (fig. 1C). Significant correlation between genetic distance (F_{ST}) and geographical distance (great circle distance) was observed in XJU (supplementary fig. S6, Supplementary Material online, $R^2 = 0.178$, $P = 1.22 \times 10^{-4}$). We further applied EEMS analysis and identified a distinct genetic barrier that roughly runs from north to south (fig. 1D). The barrier starts from the Altai Mountains, moving around Urumqi on its way to south, and coincides with the Tian Shan Mountains for a distance; then, it turns south, extending into the Taklamakan Desert, and ending at the Kunlun Mountains (supplementary fig. S7, Supplementary Material online). This pattern coincides with the observed stratification of XJU samples in PCA but was in contrast to the north-south divergence assuming the east–west chain of the Tianshan Mountains to be a natural barrier. A possible explanation could be that gene flow from the eastern and western neighboring populations has some considerable influence on the genetic makeups of XJU, which will be discussed in the next sections.

Ancestral Makeup of XJU

To unveil the ancestry makeup of XJU and investigate the genetic influence of the surrounding populations, we performed ADMIXTURE analysis of XJU combined with global populations, assuming that K ranged from 2 to 20 (supple mentary fig. S8 and table S2, Supplementary Material online). At $K = 4$, the ancestral makeup of XJU could be explained by two major ancestral components that were represented by EA and WE populations, respectively (supplementary fig. S8, Supplementary Material online), which is in agreement with the findings of previous studies (Li et al. 2009; Xu et al. 2008). Southwest XJU showed a larger proportion of West Eurasian ancestry, particularly Kizilsu XJU (49.9%). The proportion gradually decreased towards the northeast XJU, and was lowest in Kumul XJU (33.8%), whereas the distribution of East Asian ancestry in XJU was in the opposite direction (supplementary fig. S9, Supplementary Material online).

From $K = 8$ to $K = 20$ in ADMIXTURE analysis, different XJU regional populations shared the majority of their

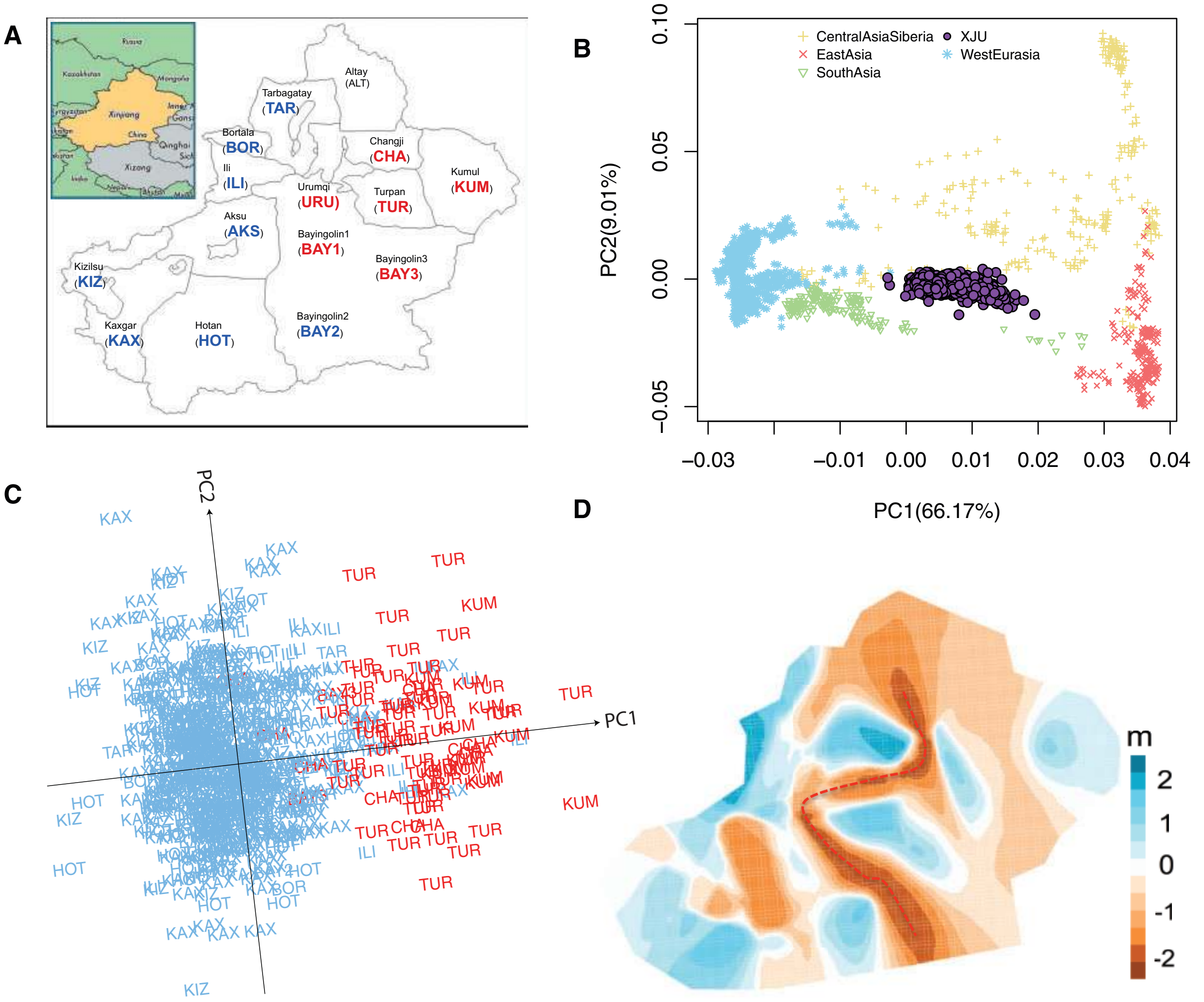

Fig. 1. Genetic affinity and population structure of Xinjiang's Uyghur (XJU). (*A*) Distribution map of XJU samples. The abbreviated name of each region was colored according to figure 1C. No samples were collected from Altay. (*B*) PCA of 951 XJU with reference populations from Central Asia Siberia, East Asia, South Asia, and West Eurasia. (*C*) PCA of 679 XJU individuals from 11 regions (samples from Aksu (AKS), Bayingolin1 (BAY1), and Urumqi (URU) were excluded). (*D*) Effective migration rates of XJU individuals based on EEMS analysis with 700 demes. Dark blue color indicates higher migration rate, whereas brown color indicates lower migration rate. Red dashed line indicates the putative genetic barrier.

ancestral makeup with populations from EA, SIB, WE, and SA (supplementary fig. S8, Supplementary Material online). These four major ancestries of XJU were confirmed by running ADMIXTURE for XJU together with representative populations of EA, SIB, WE, and SA ancestries ($K = 3$–8, see Materials and Methods, supplementary fig. S10, Supplementary Material online). The estimated ancestral proportions, as confirmed by reduced data sets ($K = 4$), are as follows: EA (28.8%–46.5%), SIB (15.2%–16.8%), WE (24.9%–36.6%), and SA (12.0%–19.9%) (fig. 2*A* and *B*). In contrast, the Turkish populations share the majority of their ancestral makeup with populations from WE (74.8%) and SA (16.5%), whereas those from the East is considerably lower (4.03% EA and 4.66% SIB) (fig. 2*A*). A significant difference in admixture proportions was observed between Northeast and Southwest XJU (supplementary fig. S11 and S12, Supplementary Material online). WE ancestry proportions were positively correlated with SA ancestry proportions in XJU across different regions, and both were negatively correlated with longitudes from Southwest to Northeast, whereas EA and SIB ancestry in XJU were positively correlated with longitudes (fig. 2*C*). However, none of the ancestries were significantly correlated with latitude (fig. 2*C* and supplementary fig. S13, Supplementary Material online), indicating that the gene flow in the east–west direction was more frequent than in the north–south direction. Therefore, the observed southwest–northeast differentiation within XJU likely resulted from a joint effect of the barrier of the Tianshan Mountains and gene flow from eastern and western neighboring populations.

Population Admixture and Admixture History

We further applied admixture history graph (AHG) analysis (Pugach et al. 2016) to determine admixture chronology, i.e., the chronology of introduction of each ancestry into XJU's gene pool (see Methods). The results indicated that the WE

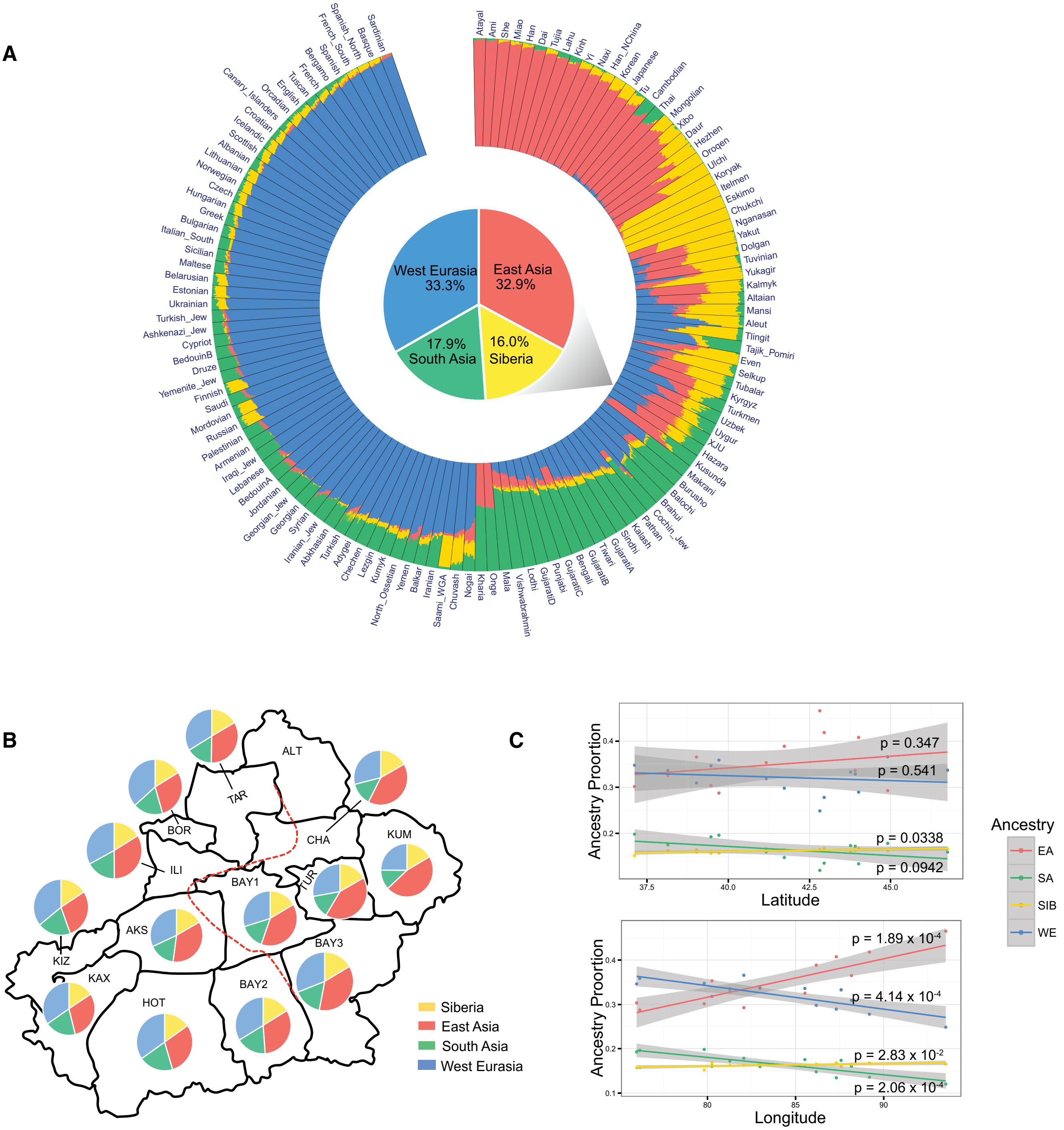

Fig. 2. Ancestry makeup and variations in admixture proportion within XJU. (*A*) ADMIXTURE results of XJU with West Eurasia, South Asia, East Asia, and Central Asia Siberia populations at $K = 4$ (in order to show a full picture of Eurasian populations, all populations of the four major ancestries (including representative populations) were included.). Proportion of each ancestry in XJU was highlighted with a pie chart. Height of the bar is proportional to admixture proportion. (*B*) Admixture proportions of the four major ancestries in regional XJU based on ADMIXTURE analysis in figure 2*A*. Red dashed line indicates the genetic barrier from EEMS analysis in figure 1*D*. (*C*) Correlations between proportions of four major ancestries and latitudes as well as longitudes of XJU across regions. Grey region indicate 95% confidence region for each regression fit. EA = East Asia; SA = South Asia; SIB = Siberia; WE = West Eurasia.

ancestry first admixed with the SA ancestry in the West, whereas the SIB ancestry admixed with the EA ancestry in the East. Next, the mixed Western ancestries (WE-SA) and the mixed Eastern ancestries (EA-SIB) joined together to form the gene pool of XJU (fig. 3*A* and supplementary fig. S14, Supplementary Material online). The AHG results were highly consistent across ADMIXTURE replicates (supplementary figs. S15 and S16, Supplementary Material online). This configuration supported the inference generated by Globetrotter (Hellenthal et al. 2014), in which Iranians and Mongolians are the best representative ancestral origins of the Uyghur admixture. Among other Central Asian populations, our analysis showed that Uzbek and Turkmen own the same admixture configuration as XJU.

If the inferred configuration for XJU is true, then we would expect some other existing (or once existing) populations that show signals of EA-SIB admixture or WE-SA admixture. Therefore, we conducted a systematic analysis detecting each Eurasian population in the Human Origins data set for such admixture signals, including the EA-SIB, SIB-WE, WE-SA, and SA-EA admixtures (supplementary fig. S17 and table S3, Supplementary Material online). Strong signals were observed for the EA-SIB, SIB-WE, and WE-SA admixtures in Eurasia, but not for the SA-EA admixture. The limited SA-EA admixture signal could have resulted from geographic barriers such as the Himalaya Mountains situated between South Asia and East Asia. In contrast, the vast Eurasia steppe stretching from Eastern Europe, through Central Asia, Siberia, to the northern part of East Asia facilitated gene flow between the neighboring populations living on the steppe.

Given the ([WE, SA], [EA, SIB]) admixture model we established in XJU data, we expect that the WE-SA and EA-SIB admixtures have occurred earlier than the West (harboring WE and SA ancestries)—East (harboring EA and SIB ancestries) admixture. We estimated the EA-SIB admixture time based on Eastern admixed populations (Daur, Hezhen, Mongolian, Oroqen, Ulchi, and Xibo), whose major components (proportion >10%) were limited to EA and SIB (fig. 2A), and inferred WE-SA admixture time based on Western admixed populations (Brahui, Burusho, Pathan, Sindhi, and Tajik_Pomiri), whose major components were limited to WE and SA (fig. 2A). We estimated the time for West–East contacts based on the data of XJU by using *AdmixInfer* (Ni et al. 2016). The results showed that the WE-SA admixture occurred 120∼100 generations ago under a continuous gene flow (CGF) model, EA-SIB occurred 140∼100 generations ago under a CGF model, and the West–East admixture occurred 100∼90 generations ago under a gradual admixture model (supplementary fig. S18, Supplementary Material online). The admixture dates for most Eastern admixed populations were earlier than the West-East admixture, thereby providing additional support for the admixture topology earlier described. However, this particular pattern was not significant in Western admixed populations, which might due to differences in true source populations from those analyzed.

Because *AdmixInfer* analysis indicated multiple waves of admixture in the history of XJU, we used *MultiWaver* (Ni unpublished data) to infer multiple waves from multiple sources in admixed populations (fig. 3B). The first wave of WE-SA admixture was estimated to have occurred around 180∼130 generations ago (variations exist across populations), followed by subsequent gene flow. In the East, the first wave of EA-SIB admixture was estimated to have occurred 225∼180 generations ago, and was also followed by subsequent gene flow. Two waves were detected in most regional populations of XJU, with the first admixture being estimated to have occurred around 150 generations ago, followed by a second wave at around 30 generations ago (fig. 3B). The 95% confidence intervals of the admixture dates for each population were obtained from 1,000 bootstrapping repeats, and the intervals were generally small with a range of 4–38% of each point estimation (supplementary table S4, Supplementary Material online). Simulations were performed to specifically evaluate the performance of *MultiWaver* in detecting two-wave admixtures (see supplementary text S1, Supplementary Material online). Our results confirmed the ability of *MultiWaver* to determine the correct admixture model and detect ancient admixture events. Nevertheless, *MultiWaver* tends to underestimate the time of the first wave in cases involving very ancient admixtures (>100 generations). Despite this slight underestimation, the time for the first wave admixture, i.e., ∼3750 ya, assuming a generation time of 25 years (McEvoy et al. 2011) was much earlier than that reported by previous studies (Hellenthal et al. 2014; Patterson et al. 2012; Loh et al. 2013; Xu and Jin 2008). Interestingly, this early wave of admixture largely overlapped with the range of the dating of mummies exhibiting European features that were discovered in the Tarim basin, which is situated in southern Xinjiang (4,000–2,000 ya) (Hemphill and Mallory 2004; Mair 1995). Our estimation of time for the second wave was around 30 generations ago (fig. 3B), which coincides with the estimate of a recent study using a different method (Patterson et al. 2012). A summary of the admixture history of the Uyghurs based on our analyses is depicted in figure 3C.

Further anlaysis using *MultiWaver* indicated that some regional XJU groups such as from Kumul, Bayingolin2, and Changji showed slightly larger time estimations for the first admixture wave, as well as an additional admixture wave from the east at around 100∼80 generations ago (fig. 3B). On the one hand, these results suggested a more complex admixture history for XJU; on the other hand, it was possible that *MultiWaver* underestimated the number of admixture waves, which could occur in case of long-term isolation and recent gene flow. To find evidence of recent gene flow in XJU, we categorized the local ancestry tracts in XJU into the East origin and West origin, and binned the tracts based on their lengths. After correction for admixture proportion, the results showed that long tracts (>0.2 Morgan) of the West origin were significantly enriched in Southwest XJU groups such as Hotan and Kizilsu, whereas long tracts of East origin were significantly enriched in Northeast XJU groups such as Kumul and Turpan (supplementary fig. S19, Supplementary Material online). These results thus suggest recent gene flow (<5 generations) into XJU from west and east neighboring populations.

Discussion

Chronological estimations of the introduction of various ancestries into the XJU's gene pool are quite complex, and thus we should point out that the model we provided in the present study is much simplified. For example, our analysis also indicated signatures of direct SIB-WE admixture during the population history (supplementary fig. S17, Supplementary Material online). A recent study has suggested that the European ancestry could be one of the

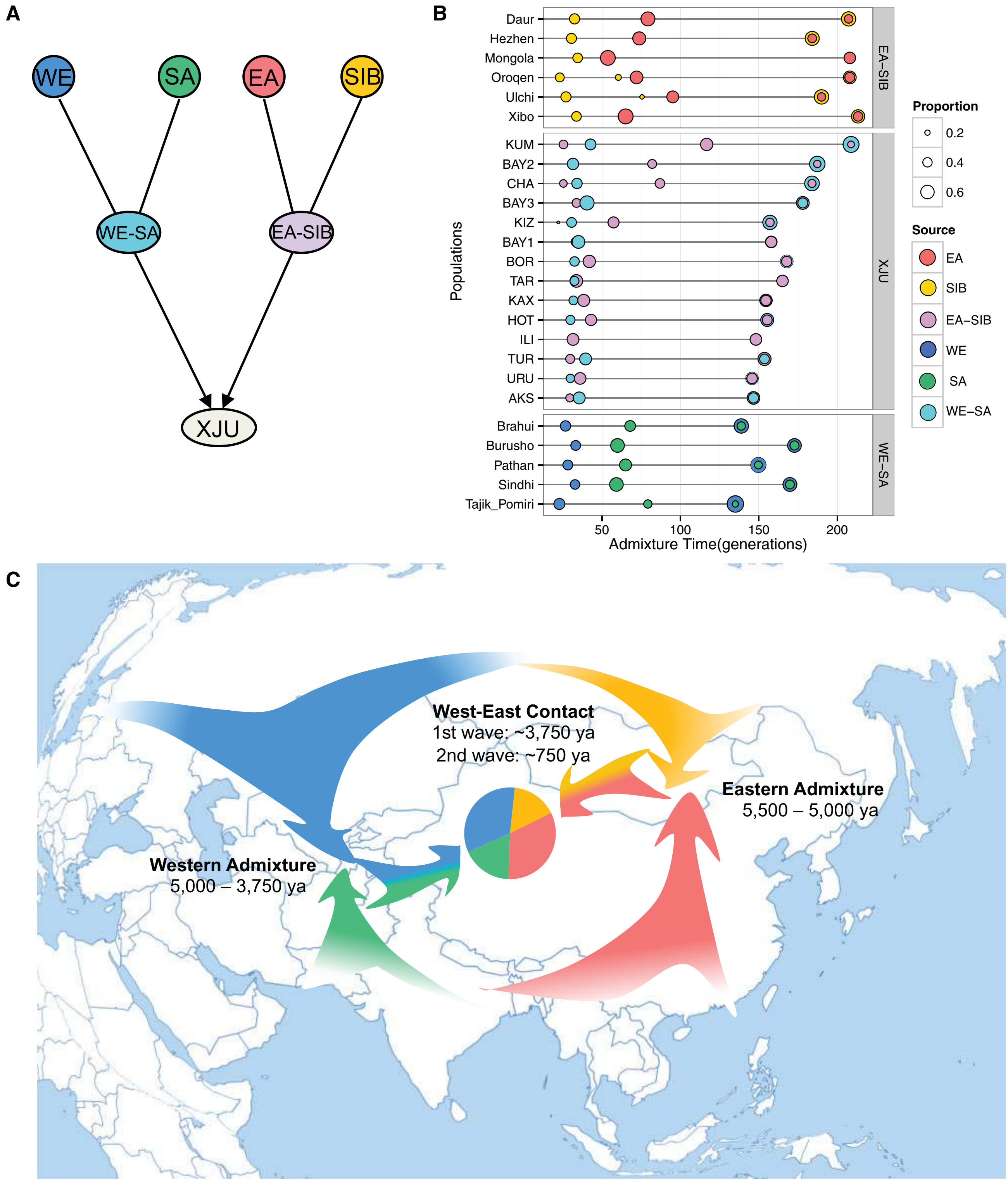

FIG. 3. Population admixture and admixture history of XJU. (*A*) The most likely admixture topology of the four major ancestries in XJU based on Admixture History Graph (AHG) results. Ancestries represented as colored balls were as follows: EA (Red); SIB (Gold); EA-SIB (Plum); WE (Blue); SA (Green); WE-SA (Cyan). (*B*) Admixture time of EA-SIB admixture, WE-SA admixture, and West–East admixture in XJU by *MultiWaver*. Representative populations (represented as colored balls) used in *MultiWaver* were as follows: EA = Atayal; SIB = Nganasan; EA-SIB = Hezhen; WE = Sardinian; SA = Mala; WE-SA = Tajik_pomiri. Representative populations were selected according to ADMIXTURE results with their major components (>10%) limited to representative ancestries. XJU from different regions in Xinjiang were inferred separately. (*C*) Schematic map showing the possible admixture model of XJU. Arrows in different colors indicate ancestral sources and directions of the gene flow.

most ancient components of Siberian populations (Pugach et al. 2016). Therefore, it is possible that some Siberian-like component was also introduced into the Uyghur by the West Eurasian populations, and vice versa. In addition, multiple entries of the Siberia ancestry into the Uyghur population could have resulted in weaker correlations between Siberian ancestry and other ancestries, which might explain the signals of topologies ([WE, EA], SIB) and ([EA, SA], SIB) in some northeast XJU samples. On the other hand, recent gene flow from neighboring East Asians or Siberians into Northeast XJU could be another factor (supplementary fig. S19, Supplementary Material online).

Several previous studies (supplementary fig. S20, Supplementary Material online) (Ni et al. 2016; Patterson et al. 2012; Loh et al. 2013; Xu et al. 2008) that utilized different methods have estimated the admixture time in Uyghurs to range from ∼120 to ∼20 generations ago. However, previous methods assumed only a single admixture event, whereas multiple waves of admixtures have likely occurred during XJU history, considering the very complex ancestral composition of their gene pool. In the present study, we used *MultiWaver* to detect multiple admixture events and showed that at least two admixture events involving XJU has occurred, with the ancient wave of admixture occurring ∼150 generations ago, and the second wave occurring ∼30 generations ago. The time of the more recent wave was close to the estimates using Rolloff (Patterson et al. 2012), ALDER (Loh et al. 2013), and Globetrotter (Hellenthal et al. 2014) (supplementary fig. S20, Supplementary Material online), whereas the time of the older admixture wave overlapped with the age range of the mummies with European features discovered in Xinjiang (4,000–2,000 ya) (Hemphill and Mallory 2004; Mair 1995), in which both western and eastern lineages have been identified (Li et al. 2015). Analysis of ancient human mtDNA also suggested a West–East admixture in Tarim Basin in the early Bronze Age (Li et al. 2010).

We further estimated the admixture times of EA-SIB and WE-SA ancestral source populations using Southern Siberian and South Asian populations as references. The time of the first admixture in Southern Siberian populations such as Hezhen (∼180 generations ago) and Oroqen (∼200 generations ago, fig. 3B), which could still be underestimated according to our simulation results, was older than previous estimates based on StepPCO (both at ∼70 generations ago) (Pugach et al. 2016). StepPCO assumes a single wave admixture, and thus it is also likely to underestimate the time for the ancient admixture events given multiple admixture waves actually occurred in the history of the populations studied. A recent study has also revealed that both prehistoric SIB and EA lineages in the ancient populations that were located in northern Xinjiang underwent admixture between Siberians and East Asians dating back to the early Bronze Age (3900–3300 ya) (Gao et al. 2015).

Taken together, we unveiled a more complex scenario of ancestral origins, population structure, and admixture history for XJU than previously reported. The two-wave model we proposed here does not necessarily mean there were only two admixture events that had occurred in the history of XJU. Rather, it suggests that population admixture occurred more than once during human dispersal in Xinjiang and its surrounding areas. The Silk Road was not a single route, but rather a series of routes (Wood 2002), and Xinjiang is a territory with a documented history of at least 2,500 years and populated by a succession of peoples and empires during the course of its history. Therefore, the vast territory is expected to have experienced a complex scenario of genetic admixture between populations from all around Eurasia. Further studies of over ten other ethnic groups residing in the area such as the Han Chinese, Hui, Kazakhs, Kyrgyz, Mongols, Russians, and Tajiks, together with information of archaeological remains and ancient DNA, are expected to provide further insights into the complete picture of human dispersal in Xinjiang and all of Eurasia.

Materials and Methods

Populations and Samples

Peripheral blood samples of 968 unrelated Uyghur individuals were collected from 12 prefectures (Kaxgar, Hotan, Kizilsu, Aksu, Bayingolin, Turpan, Kumul, Changji, Ili, Bortala, Tarbagatay, and Altay) and 1 prefecture-level city (Urumqi) in Xinjiang Uyghur Autonomous Region, China. The samples enrolled in this study were collected randomly but roughly proportional to the total Uyghur population size of each geographical region (supplementary fig. S21, Supplementary Material online). An estimated 80% of Uyghur people are living in the Southwest portion of Xinjiang Tarim Basin. Correspondingly, the majority of the samples were from southwest Xinjiang (supplementary fig. S22, Supplementary Material online). Each individual was the offspring of a non-consanguineous marriage of members of the same nationality within three generations. Informed consent was acquired from the participants. All procedures performed were in accordance with the ethical standards of the Responsible Committee on Human Experimentation (approved by the Biomedical Research Ethics Committee of Shanghai Institutes for Biological Sciences, No. ER-SIBS-261408) and the Helsinki Declaration of 1975, as revised in 2000.

Genotyping, SNP Calling, and Quality Control

We genotyped 968 Uyghur samples on the Illumina HumanOmniZhongHua-8 chips ($n = 727$) and Affymetrix Genome-Wide Human SNP Array 6.0 ($n = 241$). The SNP genotypes from the Illumina OmniZhongHua array were called with GenomeStudio V2011.1 (Illumina, Inc). SNP calling yielding a GenCall Score < 0.15 were treated as missing genotypes (as recommended by Illumina). SNP genotypes from the Affymetrix Genome-Wide Human SNP array were called with "apt-probeset-genotype" from Affymetrix Power Tools 1.10.2 (Affymetrix, Inc). SNP calling yielding a confidence value > 0.1 was considered as missing data (as recommended by Affymetrix). In total, 212,890 common SNPs with concordant allele states were shared between the two platforms. Initial quality control left us with 951 Uyghur individuals for further analyses, of which 225 were assayed on Affymetrix 6.0, and 726 on Illumina OmniZhongHua chips. The missing rate of each sample was <10% (0.00–5.01%). The 212,890 common SNPs with concordant allele states were further filtered to 183,070 loci by removing SNPs showing a missing rate > 0.1 or Hardy–Weinberg disequilibrium (P value $< 10^{-6}$) with PLINK (v1.07) (Purcell et al. 2007). We examined the potential batch effects using principal component analysis (PCA) (Patterson et al. 2006; Price et al. 2006) and ADMIXTURE (Alexander et al. 2009) analysis, but did not observe any substantial batch effects (supplementary figs. S23 and S24, Supplementary Material online). In addition, because the samples genotyped using both platforms encompassed all geographical regions and were distributed evenly, the risk for potential batch effects, if existing, were considerably reduced.

Determination of Geographical Location of Each Sample

The region of each sample was determined based on the participant's grandparents' region information. The region of this sample could be determined only when the region information of all the participant's grandparents was consistent. Otherwise, the sample was not classified into any region. In total, 902 samples were assigned locations, whereas 49 samples remained ambiguous with regards to specific region information. Because we divided the Bayingolin into three subregions, ten more samples without specific subregion information were excluded. Finally, 892 samples had specific region information. The number of samples for each region is listed in supplementary figure S22, Supplementary Material online.

Public and Published Data

The Affymetrix Human Origins genotyping data set (Lazaridis et al. 2014) for 2,367 human samples was obtained with a signed letter permitting full data access, and was used for comparison with XJU under a global context. We used the "Simple population ID" instead of "Verbose population ID" to assign the population identification for each individual. Finally, 49 populations from Africa, 60 populations from West Eurasia, 22 populations from South Asia, 23 populations from Central Asia/Siberia, 22 populations from East Asia, 3 populations from Oceania, and 24 populations from America, resulted in a total of 2,345 individuals from 203 populations (supplementary fig. S25, Supplementary Material online) for use in the following analyses: e.g., ADMIXTURE analyses, f-statistics analyses, PCA. For the purposes of this study, only SNPs with reference sequence numbers and vendor-specified strands were used in combining data, with 66,410 SNPs left. On the other hand, for analysis that required higher density data sets, e.g., analysis of local ancestry, we imputed Uyghur data from the two platforms separately by IMPUTE2 (Howie et al. 2009), with the 1000 Genomes phase III data sets as references. The haplotypes underlying the Uyghur genotype were prephased with SHAPEIT2 prior to imputation (Delaneau et al. 2013; Howie et al. 2012). We next combined the imputed Uyghur data with the Human Origins data set, which left us with 557,093 loci for all samples. Again, we examined the potential batch effects of imputed data between two platforms using PCA and ADMIXTURE, but did not observe any substantial batch effects (supplementary fig. S26 and S27, Supplementary Material online).

Calculation for F_{ST}

Genetic difference between populations was measured using F_{ST} following Weir and Cockerham (1984), which accounts for differences in sample sizes between populations. SNP-specific F_{ST} was also calculated. Confidence intervals of the F_{ST} over loci were calculated by bootstrap resampling with 1,000 replications. To reduce the influence of sample size differences between regions, we randomly selected 11 samples (the smallest sample size among regions, except for Bayingolin3, which only had three samples) for each region to calculate the pairwise F_{ST} matrix.

PCA

PCA was performed at the individual level using EIGENSOFT v4.2 (Patterson et al. 2006). To investigate fine-scale population structure and individual genetic affinities, we performed a series of PCA by gradually removing outliers based on a plot of the first and second principle components (PCs), and reanalyzing the remaining samples based on the same set of SNPs.

Linear Regression Analysis

To investigate the relationship between genetic differentiation of regional XJU subpopulations and their geographic locations, we applied simple linear regression to analyze the correlation between genetic and geographic distance using the "lm" function in the "R" package. Pairwise F_{ST} between regional subpopulations were taken as genetic distances. Great circle distance was calculated by using the "Math::Trig" module in Perl, which served as the geographic distance between regions. We also applied multiple linear regression to analyze the correlation between the geographic location (longitude and altitude) and PC1 coordinates of regional XJU. The PC1 was based on the analysis of XJU samples without reference populations. The plot was drawn by using the "scatterplotMatrix" function in the "car" package.

Quantification and Visualization of Migration Rates

We applied an estimated effective migration surface (EEMS) (Petkova et al. 2014) algorithm to quantify migration rate within the XJU population based on both genetic and geographic information. This method uses a population genetic model to relate underlying migration rates to expected pairwise genetic dissimilarities, and estimates migration rates by matching these expectations to the observed dissimilarities.

For a total of 951 Uyghur samples, 49 samples without region information and 5 outliers in PCA analysis were excluded. Finally, 897 samples remained for EEMS analysis. A geographic outline of Xinjiang was assigned by 27 coordinates using the "polyline" function in Google Maps. The number of demes within the outline was set to be 100, 400, 700, and 1000, respectively. We ran three repeats for each deme setting to ensure that the Markov Chain Monte Carlo converged. The main information transmitted by different deme settings was similar, and "R" package rEEMSplots were used to visualize the results of 700 demes.

Inferring Global Ancestry by Using ADMIXTURE

ADMIXTURE (Alexander et al. 2009) was applied on the merged data set of Human Origins and XJU data, which consisted of 3,296 (2,345 + 951) samples representing 204 populations. We used PLINK 1.07 (Purcell et al. 2007) to prune the original data set with dense SNPs, after assigning an r^2 threshold of 0.4 in every continuous window of 200 SNPs advanced by 25 SNPs (–indep-pairwise 200 25 0.4); 57,480 SNPs remained for ADMIXTURE analysis. We ran ADMIXTURE with random seeds for the merged data set assuming the number of ancestral clusters (K) ranged from 2 to 20.

Because the clustering algorithm implemented in ADMIXTURE may incorporate stochastic simulation as part of the inference, independent analyses of the same data may result in slightly different results (Jakobsson and Rosenberg 2007). To obtain more reliable results, we replicated six times with different seeds for each run of ADMIXTURE assuming the same K and used CLUMPP 1.1.2 (Jakobsson and Rosenberg 2007) to obtain the optimal alignment of the six replicates. We chose the replicate that was most similar to the optimal alignment by using CLUMPP for further analysis (supplementary fig. S8, Supplementary Material online). We also assessed the cross validation error among the six replicates, $K = 16$ to $K = 19$ was the optimal K range that best explained Human Origins + XJU data sets, as shown in supplementary figure S28, Supplementary Material online.

To resolve the admixture ancestry of XJU, for each K, we identified the major clusters (with proportion > 5%) consisting of XJU. For each major cluster, we identified its representative reference populations. To confirm the major clusters in XJU at each K, we ran ADMIXTURE with the subset of the data (XJU and its representative reference populations of major clusters in XJU identified in the main run). The major clusters were validated with different data sets (supplementary table S2, Supplementary Material online). In addition, we specifically ran ADMIXTURE for the subset of data (XJU and the representative reference populations of EA, SIB, WE, and SA), assuming $K = 3$–8 with ten replicates to confirm the four major ancestries identified in XJU (supplementary fig. S10, Supplementary Material online). ADMIXTURE analysis was performed under unsupervised mode.

Analysis of Admixture History Graph (AHG)

To disentangle the admixture chronology of the four major ancestries represented by East Asia (EA), Siberia (SIB), West Eurasia (WE), and South Asia (SA) in XJU, we applied AHG analysis to XJU samples of each region (Pugach et al. 2016). The analysis was based on the idea that for an admixed population with ancestries A, B, and C, if the admixture topology is ([A, B], C), then the admixture proportion ratio of A and B would be independent of the admixture proportion of C. Thus, the covariance of recent ancestry C and the ratio of the two older ancestries A and B should be zero. To determine the sequence of admixture events in XJU, we compared all possible combinations of any three ancestries from EA, SIB, WE, and SA. For each trio (a group of any three ancestries), we chose the topology that produced the lowest absolute value of covariance. Then, we reconstructed the full graph of the four ancestries based on the likely topologies of each trio.

The admixture proportion of each ancestry in XJU was obtained by ADMIXTURE analysis of global populations and assuming eight ancestral clusters ($K = 8$), where XJU's ancestral makeup was best explained by four major ancestral components (supplementary table S2, Supplementary Material online). Pearson correlation coefficient instead of covariance was used in choosing the topology because the former adjusted the bias caused by admixture proportion differences among ancestries. Topology analysis of each trio was repeated 100 times, with ten individuals randomly sampled with replacement for each repeat. The topologies that produced the lowest absolute value of Pearson correlation coefficient were chosen.

Consistency of AHG results were further examined by conducting AHG analysis across 100 ADMIXTURE runs in the context of global populations ($K = 8$) as well as reduced data sets (XJU and representative reference populations of EA, SIB, WE, SA, $K = 4$), separately. AHG results showed high concordance across ADMIXTURE repeats (supplementary figs. S15 and S16, Supplementary Material online).

Detecting Admixture Signal Using 3-Population Test

For each population in the Human Origins data set, we detected gene flow from ancestral populations as represented by neighboring regions by f_3 (X; ref1, ref2) with admixture-tool 1.1 (Patterson et al. 2012). Atayal, Mala, Sardinian, and Nganasan were chosen as representative populations of EA, SA, WE, and SIB, respectively, based on ADMIXTURE analysis using global populations and assuming $K = 8$. A significantly negative f_3 value indicates that X is an admixed population of ancestries, similar to ref1 and ref2 (supplementary table S3, Supplementary Material online).

Local Ancestry Inference

HAPMIX (Price et al. 2009) was employed to infer local ancestry assuming a two-way admixture. Atayal (proxy of EA) and Nganasan (proxy of SIB) were used as reference populations for EA-SIB admixed populations such as Hezhen. Sardinian (proxy of WE) and Mala (proxy of SA) were used as reference populations for WE-SA admixed populations such as Tajik_Pomiri. Hezhen and Tajik_Pomiri were used as reference populations for XJU based on their similarities in EA/SIB and WE/SA admixture ratio relative to that of XJU. The admixture proportion parameter "theta" in HAPMIX was set according to the results of ADMIXTURE analysis results using global populations at $K = 8$. The admixture time parameter "lambda" was set to 80 because "lambda 80" gave one of the largest log likelihoods across XJU, EA-SIB admixed populations, and WE-SA admixed populations (supplementary fig. S29, Supplementary Material online).

Estimation of Admixture Time Based on Admixture Linkage Disequilibrium (LD)

ALDER (Loh et al. 2013) based on admixture LD information was applied to XJU to infer admixture time using a single reference. In the first scenario, we randomly sampled 100 XJU. A list of reference populations representative of WE, SA, EA, or SIB ancestries were used. The reference populations were chosen according to the ADMIXTURE results that contained representative ancestry of >90%. In the second scenario, we dated XJU from each region with Korean, Itelmen, Sardinian and Kalash as proxies of EA, SIB, WE, and SA, respectively, which were chosen according to amplitude score in the first scenario. Both scenarios gave an estimation of 22∼17 generations, which is in agreement with that reported by previous studies using ALDER (Loh et al. 2013) (supplementary figs. S30 and S31, Supplementary Material online).

Estimation of Admixture Time Based on Length Distribution of Ancestral Tracks

AdmixInfer (Ni et al. 2016) and *MultiWaver* (Ni unpublished data) were applied to infer admixture time and admixture model, assuming a two-way admixture. Both methods are based on the length distribution of ancestral tracks inferred from admixed genomes. Ancestral tracks were inferred using HAPMIX, where Tajik_Pomiri and Hezhen were selected as representatives of ancestral source populations for XJU, Sardinian (proxy of WE) and Mala (proxy of SA) for WE-SA populations, and Atayal (proxy of EA) and Nganasan (proxy of SIB) for EA-SIB populations. Default parameters were used for both algorithms, and 1 million iterations were performed to ensure convergence of key parameters (Ni et al. 2016).

AdmixInfer infers admixture history under three typical two-way admixture models, namely, the hybrid isolation (HI) model, gradual admixture (GA, continuous gene flow from both sources) model, and continuous gene flow (CGF, continuous gene flow from only one source) model. *MultiWaver* is a new method we had developed for inferring the optimal model without prior model assumptions or estimate parameters. It can infer a multiple-wave admixture model with multiple-ancestral populations based on the information of ancestral tracks in admixed genomes (Ni unpublished data). The program proceeds in two steps: using a likelihood ratio test and an exhaustion method to select an optimal admixture model based on the length distribution of ancestral tracks, and then applying an EM-algorithm to estimate the corresponding parameters (admixture times and proportions) under this optimal model.

The performance of *MultiWaver* was evaluated with simulations under the two-wave model of XJU, EA-SIB, and WE-SA admixed populations that were inferred in our study. Forward-time simulator *AdmixSim* (Yang et al. 2016) generated simulation data with the same parental populations for time estimation (see supplementary figs. S32–S34, Supplementary Material online). The results showed that *MultiWaver* could correctly identify the admixture model and estimate admixture time.

Web Resources

The URLs for data presented herein are as follows:

Database of Genomic Structural Variation (dbVar), http://www.ncbi.nlm.nih.gov/dbvar/; last accessed June 16, 2017.

Database of Genomic Variants, http://dgv.tcag.ca/dgv/app/home; last accessed June 16, 2017.

UCSC Genome Browser, http://genome.ucsc.edu/; last accessed June 16, 2017.

AdmixInfer and *MultiWaver*, http://www.picb.ac.cn/PGG/resource.php; last accessed June 16, 2017.

AdmixSim, http://www.picb.ac.cn/PGG/resource.php; last accessed June 16, 2017.

The 1000 Genomes Project, http://www.1000genomes.org, last accessed June 16, 2017.

Google Map, http://www.birdtheme.org/useful/v3tool.html; last accessed June 16, 2017.

Accession Numbers

The accession number for the SNP data reported in this paper is National Omics Data Encyclopedia (NODE, http://www.biosino.org/node/): ND00000038EP.

Supplementary Material

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

Acknowledgments

S.X. acknowledges financial support from the National Natural Science Foundation of China (NSFC) grant (91331204 and 31711530221), the Strategic Priority Research Program (XDB13040100) and Key Research Program of Frontier Sciences (QYZDJ-SSW-SYS009) of the Chinese Academy of Sciences (CAS), the National Science Fund for Distinguished Young Scholars (31525014), and the Program of Shanghai Academic Research Leader (16XD1404700); Y.L. acknowledges support from NSFC grant (31501011) and Science and Technology Commission of Shanghai Municipality (STCSM) (14YF1406800); H.L. acknowledges support from STCSM grant (16YF1413900); S.W. acknowledges supports from NSFC grant (91631307), the Strategic Priority Research Program (XDB13041000), STCSM grant (16JC1400504). S.W. also acknowledges the National Thousand Young Talents Award, and a Max Planck-CAS Paul Gerson Unna Independent Research Group Leadership Award. Y.G. acknowledges support from NSFC grant (31260263). This work is also supported by the National Key Research and Development Program (2016YFC0906403 and 2016YFF0202301). S.X. is Max-Planck Independent Research Group Leader and member of CAS Youth Innovation Promotion Association. S.X. also gratefully acknowledges the support of the National Program for Top-notch Young Innovative Talents of The “Wanren Jihua” Project. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

References

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

Balfour E. 1985. The cyclopedia of India and of Eastern and Southern Asia: commercial, industrial and scientific, products of the mineral, vegetable, and animal kingdoms, useful arts and manufactures, 3rd edn. London: B. Quaritch.

Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*. 5: e1000529

Delaneau O, Zagury JF, Marchini J. 2013. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 10:5–6.

Gao SZ, Zhang Y, Wei D, Li HJ, Zhao YB, Cui YQ, Zhou H. 2015. Ancient DNA reveals a migration of the ancient Di-qiang populations into Xinjiang as early as the early Bronze Age. *Am J Phys Anthropol*. 157:71–80.

Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A genetic atlas of human admixture history. *Science* 343:747–751.

Hemphill BE, Mallory JP. 2004. Horse-mounted invaders from the Russo-Kazakh steppe or agricultural colonists from western Central Asia? A craniometric investigation of the Bronze Age settlement of Xinjiang. *Am J Phys Anthropol*. 124:199–222.

Henders SJ. 2006. Democratization and identity: regimes and ethnicity in east and southeast Asia. Lanham: Lexington Books.

Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*. 44:955–959.

Li H, Cho K, Kidd JR, Kidd KK. 2009. Genetic landscape of Eurasia and "admixture" in Uyghurs. *Am J Hum Genet*. 85:934–937.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23:1801–1806.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513:409–413.

Li C, Li H, Cui Y, Xie C, Cai D, Li W, Mair VH, Xu Z, Zhang Q, Abuduresule I, et al. 2010. Evidence that a West-East admixed population lived in the Tarim Basin as early as the early Bronze Age. *BMC Biol*. 8:15.

Li C, Ning C, Hagelberg E, Li H, Zhao Y, Li W, Abuduresule I, Zhu H, Zhou H. 2015. Analysis of ancient human mitochondrial DNA from the Xiaohe cemetery: insights into prehistoric population movements in the Tarim Basin, China. *BMC Genet*. 16:78.

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

Mair VH. 1995. Prehistoric Caucasian corpses of the Tarim Basin. *J Indo Eur Stud*. 23:281–307.

McEvoy BP, Powell JE, Goddard ME, Visscher PM. 2011. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res*. 21:821–829.

Millward JA. 2007. Eurasian crossroads: a history of Xinjiang. New York: Columbia University Press.

Millward JA, Perdue PC. 2004. Chapter 2: Political and cultural history of the xinjiang region through the late nineteenth century. In: Starr SF, editor. Xinjiang: China's muslim borderland. New York: M. E. Sharpe.

Ni X, Yang X, Guo W, Yuan K, Zhou Y, Ma Z, Xu S. 2016. Length distribution of ancestral tracks under a general admixture model and its applications in population history inference. *Sci Rep*. 6:20048.

Ni X, Yang X, Yuan K, Feng Q, Guo W, Ma Z, Xu S. unpublished data. Inference of multiple-wave admixtures by length distribution of ancestral tracks. *bioRxiv*. doi: 10.1101/096560.

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

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

Petkova D, Novembre J, Stephens M. 2014. Visualizing spatial population structure with estimated effective migration surfaces. *Nat Genet*. 48:94–100.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 38:904–909.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet*. 5:e1000519.

Pugach I, Matveev R, Spitsyn V, Makarov S, Novgorodov I, Osakovsky V, Stoneking M, Pakendorf B. 2016. The complex admixture history and recent southern origins of Siberian populations. *Mol Biol Evol*. 33:1777–1795.

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

Tursun N. 2008. The formation of modern uyghur historiography and competing perspectives toward Uyghur history. *China Eurasia Forum Quart* 6:87–100.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358–1370.

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*. 98:10244–10249.

Wood F. 2002. The Silk Road: two thousand years in the heart of Asia. Berkeley and Los Angeles: University of California Press.

Xu S, Huang W, Qian J, Jin L. 2008. Analysis of genomic admixture in Uyghur and its implication in mapping strategy. *Am J Hum Genet*. 82:883–894.

Xu S, Jin L. 2008. A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. *Am J Hum Genet*. 83:322–336.

Xu S, Jin L. 2009. Response to Li et al. *Amer J Hum Genet*. 85:937–939.

Xu S, Jin W, Jin L. 2009. Haplotype-sharing analysis showing Uyghurs are unlikely genetic donors. *Mol Biol Evol*. 26:2197–2206.

Yang X, Ni X, Zhou Y, Guo W, Yuan K, Xu S. 2016. AdmixSim: a forward-time simulator for various and complex scenarios of population admixture. *bioRxiv*. doi: 10.1101/037135.

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*. 21:2265–2280.