Phylogenetic relationship and genetic history of Central Asian Kazakhs inferred from Y-chromosome and autosomal variations

Atif Adnan[1,2+*], Guanglin He[3,6+], Allah Rakha[4], Kaidirina Kasimu[1], Jianxin Guo[3], Sibt-E-Hassan[1], Sibte Hadi[5], Chuan-Chao Wang[3*] and Jin-feng Xuan[2*]

1Department of Human Anatomy, School of Basic Medicine, China Medical University, Shenyang, 110122, China

2Department of Forensic Genetics and Biology, School of Forensic Medicine, China Medical University, Shenyang, 110122, China

3Department of Anthropology and Ethnology, Institute of Anthropology, National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen 361005 China

4Department of Forensic Sciences, University of Health Sciences Lahore 54000 Pakistan

5*University of Central Lancashire, School of Forensic and Investigative Sciences, Preston, UK*

6 *Institute of Forensic medicine, West China School of Basic Science and Forensic Medicine, Sichuan University, Chengdu 610041, China*

[*]Corresponding authors:

Atif Adnan (mirzaatifadnan@gmail.com)

Chuan-Chao Wang (wang@xmu.edu.cn)

Jinfeng Xuan (jfxuan@cmu.edu.cn)

+ Contributed equally

ABSTRACT

The Xinjiang Uyghur Autonomous Region of China (XUARC) with 47 ethnic groups is a most colourful ethnic region of China, harboring abundant genetic and cultural diversity~~, which harbors 47 ethnic groups~~. The Kazakhs are the third largest ethnic group (7.02%) after Uyghur (46.42%) and Han (38.99%) in Xinjiang, but their genetic diversity and forensic characterization are poorly understood. In the current study, we genotyped 15 autosomal short tandem repeat (STR) loci and 10 Y-STRs on 889 individuals (659 male and 230 female) collected from Kazak population of the Ili Kazak Autonomous Prefecture using AGCU Expressmarker 16 and 10Y-STR Kit (EX16+10Y). For autosomal STRs, we observed a total of 174 different alleles ranging from 6 ~~and~~ to 34.2 repeat units and FGA showed the greatest power of discrimination (20 alleles) in Ili Kazakh population. We have not observed departures from Hardy–Weinberg equilibrium (HWE) after sequential Bonferroni correction and only found a minimal departure from linkage equilibrium (LE) for a very small number of pairwise combinations of loci. The combined power of exclusion (CPE) was 0.99999998395 and combined power of discrimination (CPD) was 99.99999999999999798%. For Y-STRs, we observed a total of 496 different haplotypes on these 10 Y-STR loci. The gene diversities ranged from 0.5023 (DYS391) to 0.8357 (DYS385a/b). The overall haplotype diversity (GD) was 0.9985 with random matching probability (RMP) 0.0015. The results of population genetic analysi~~e~~s based on both autosomal and Y-chromosome

STRs demonstrated that the genetic affinity among populations are generally consistent with ~~exists along the~~ ethnic, linguistic, and continental geographical classifications~~boundary~~.

Keywords: Kazakh, Xinjiang Uyghur Autonomous Region, Expressmarker 16+10Y STR Kit, Forensic genetics,

INTRODUCTION

The Kazakhs (also spelt Kazak or Khazak) have a long history and been considered as descendants of Turkic people and Wusun people from ancient times ~~considered as descendants of Turkic people and Wusun people~~. The Ashina clan of Turk established Turkic Khanate around the Altay Mountains in the late sixth century and mixed with the former~~old~~ inhabitants (Wusun people). Qarliq, Kereit, Qidans (Khitans), Naimans, Mongols and Uyghurs of the Kipchak and Jagatai ~~k~~Khanates later mixed with the ancestors of Kazakhs (Esposito, 1999; Weller, 2006). Some historians think that the Kazakhs are one of the Mongolian tribes ~~which~~ formed around the 13th century. They moved to the south of Balkhash lake. The admixture of this initial group and Mongol tribes from Eastern Chagatai Khanate gave birth to the ancestor of modern Kazakh populations. Their language belongs to Turkic language family and they were ~~was~~ part of nomadic tribes which separated from the Uzbek Empire in fifteen century. Most of the Kazakhs live in the Altai Mountains, Ili (also called Ghulja in Uyghur language) Valley, TianShan, and Lake Issyk Kul in the northwestern areas of China and Central Asia. The idea of "Silk Road" was initially developed by Kazakh.

The ~~L~~literal meaning~~s~~ of Kazakh is "separators" or "brave and free people" (Wyatt and Di Cosmo, 2011). The area of the Ili Kazak Autonomous Prefecture is 26.91 million square kilometers and located in the XUARC. It is bordered by Russia in the north, Kazakhstan in west and Mongolia in the east. The Ili Kazak Autonomous Prefecture harbors 38 ethnic groups, among them Han representing 35.7% while other minorities accounting for~~represent~~ 64.3%. Among minorities of the Ili Kazak Autonomous Prefecture, Kazakhs comprise 26.88% of the population, Uyghur account for 21.53% and Hui contribute 11.7%.

Short tandem repeats (STRs), also known as microsatellites, ~~are found in noncoding regions of the human genome. STRs~~ are the repetitive sequences of 2-7 bp in noncoding regions of the human genome. The higher mutation rates of STRs make them~~of DNA which have repeat units of 2–7 bp and make them~~ less stable ~~when~~ compared to SNPs. The~~se~~ STRs can be easily amplified and detected using polymerase chain reaction (PCR) followed by sequencing. ~~structures which have short sequence lengths, stable in polymorphism, compact chromosomal spreading makes them detectable using PCR followed by sequencing.~~ These characteristics make STRs important for forensic DNA investigations (Hammond et al., 1994; Sánchez-Diz et al., 2009). Autosomal STRs are the most commonly used genetic markers in~~for~~ forensic investigations such as paternity testing and personal identification cases (Zhan et al., 2018). Though, Y-chromosomal STRs (Y-STRs) have an extra advantage in cases like sexual assault which involves mixture DNA ~~which has~~containing predominant female DNA. ~~Y-STRs are important in determination of paternal relationships among male individuals especially in deficiency paternity cases where mother is no longer available~~ Since ~~because~~ STRs on the non-recombining (NR) region of Y-chromosome are not~~doesn't~~ involved in meiotic recombination and remain~~s~~ intact when inherited from father to son (Adnan et al., 2018a, 2018b, 2016), Y-STRs are important in the determination of

paternal relationships among male individuals especially in deficiency paternity case where the genotype of the mother is no longer available. ~~In some cases where biological evidence is limited and individual identification and paternity test must be carried out at the same time to resolve the case.~~

Presently, autosomal STRs and~~or~~ Y-STRs are used independently for forensic investigations. To cut off the cost and human resources, different companies like Promega, USA (PowerPlex Fusion) and Life Technologies, USA (GlobalFiler) designed commercial kits to include only a few ~~Y~~ Y-STRs along with extended autosomal STRs. However, Expressmarker 16 + 10Y kit contains 15 autosomal and 10 Y-STR~~s~~ loci, which will be suitable not only for paternity testing but also for personal identification and the reconstruction of paternal lineage (Kayser et al., 2004).

To have a better understanding of the forensic characterization and genetic relationship of Ili Kazakh people, in the current study, we have investigated the polymorphisms of 25 STRs (15 autosomal STRs and 10 Y-STRs) in 889 individuals (659 male and 230 female) from Kazak population of the Ili ~~which is~~ Kazak Autonomous Prefecture ~~in the~~ in ~~Xinjiang Uyghur Autonomous Region of China (~~XUARC~~)~~. We have calculated forensic parameters and compared our population with previously published populations from nationwide and worldwide to infer~~find~~ the genetic relationships among them.

MATERIALS AND METHODS

Sample Collection and DNA extraction

Blood samples from 889 unrelated individuals (659 male and 230 female) from Kazak population of the Ili Kazak Autonomous Prefecture in XUARC were collected. All participants gave their informed consent either orally and with thumb prints (in case they could not be able to write) or in writing after the study aims and procedures were carefully explained to them in their own language. The study was approved by the ethical review board of the China Medical University, Shenyang Liaoning Province, People's Republic of China and in accordance with the standards of the Declaration of Helsinki. All blood samples were stored at -20 ºC before DNA extraction. Phenol chloroform procedure was used to extract DNA. Quantification of DNA was carried out using Quantifiler™ Human DNA Quantification Kit (Applied Biosystems, Foster City, CA, USA) according to manufacturer's instructions.

PCR amplification

PCR co-amplification of fifteen autosomal ~~STR loci~~ and ten Y-chromosomal STR loci (autosomal loci: CSF1PO, D13S317, D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D5S818, D7S820, D8S1179, FGA, TH01, TPOX, vWA; whereas the Y-STRs include~~s~~ DYS385a/b, DYS390, DYS391, DYS392, DYS393, DYS438, DYS456, DYS458 and DYS635) were performed in a fluorescence-based multiplex reaction using the AGCU Expressmarker 16+10Y STR Kit (EX16+10Y, AGCU ScienTech Incorporation, Wuxi, Jiangsu, China). 1 to 2 ng of the template DNA was used to amplify according to the manufacturer's instructions. Thermal cycling was carried out under the following conditions: 95 ºC for 120 sec; 10 cycles of 94 ºC for 30 sec, 62 ºC for 60 sec, 72 ºC for 60 sec;, 20 cycles of 90 ºC for 30 sec, 60 ºC for 60 sec, 72 ºC for 60sec; a final extension at 60 ºC 60 min; final hold at 4 ºC. Amplification

of 25 STRs wasere performed using AGCU EX16+10YExpressmarker 16 and 10Y STR Kit (EX16+10Y) in a GeneAmp PCR System 9700 thermal cycler (Applied Biosystems, Foster City, CA).

Genotyping

Amplified products were analyzed with reference of theto internal size standard (ISS) AGCU SIZ-500 and Allelic Ladder provided by EX16+10Y using an ABI (Applied Biosystems, Foster City, CA) 3500 genetic analyzer according to the EX16+10Y Expressmarker 16 + 10Y (AGCU ScienTech Incorporation, Wuxi, Jiangsu, China) standard protocol (Zhou et al., 2016). Data obtained from genetic analyzer was analyzed using GeneMapper software v3.5.

Quality control

Negative (autoclaved deionized H2O) and positive (AmpFlSTR Control DNA 9947A) controls were employed for DNA extraction, DNA quantification, PCR amplification and capillary electrophoresis. All negative controls did not show any amplified product while positive controls were consistent with known genotypes.

Statistical analysis

STRAF (STR Analysis for Forensics), a newly developed online tool (Gouy and Zieger, 2017), was used to calculate the allelic frequencies and forensic statistical parameters like power of discrimination (PD), polymorphism information content (PIC), power of exclusion (PE) and probability of matching (PM). Expected heterozygosity (He), observed heterozygosity (Ho), exact tests for Hardy–Weinberg equilibrium (HWE), pairwise *F*st and linkage equilibrium (LE) between pairwise combinations of loci were performed using software Arlequin v3.5 (Excoffier and Lischer, 2010) based on a likelihood ratio test for unknown gametic phase. Empirical distributions were obtained from 10,000 permutations. MVSP 3.1 software was used to calculate Principal components analysis (PCA) based on allelic frequencies of the 15 autosomal STR loci. Nei's genetic distances between Ili Kazakhs and previously published populations were calculated using the Phylip-3.69 Software (http://evolution.gs.washington.edu/phylip.html).

For 10 Y-STRs, haplotype diversity (HD) was calculated according to:;

$$\mathrm{HD} = \frac{n}{n-1}\left(1 - \sum_i p_i^2\right)$$

where n is the male population size and p_i is the frequency of ith haplotype. Match probabilities (MP) were calculated as the sum of squared haplotype frequencies. The discrimination capacities (DC) were calculated by dividing the number of different haplotypes by the total number of samples. Genetic distances based on Y-STRs (*Rst* and Reynolds) between Ili Kazakhs and other reference populations were generated using Arlequin v3.5 (Excoffier and Lischer, 2010). Reduced dimensionality spatial representation of the populations was performed based on *Rst* values using multi-dimensional scaling (MDS) with IBM SPSS Statistics for Windows, Version 23.0 (IBM Corp., Armonk, NY, USA). Neighbor-Joining (NJ) Phylogenetic trees were generated and visualized with MEGA7 software (Kumar et al., 2016). Finally, STRUCTURE (version 2.3.4) (Pritchard et al., 2000) was used to determine population structure within and between the Ili Kazakh and other 7 populations (Manchu, Mongol, Kyrgyz, Uzbek, Liaoning Han, Korean, and Tibetan) from Xinjiang in China, Northern China and Tibet in China.

RESULTS AND DISCUSSIONS

Forensic parameters of 15 autosomal STRs

We have successfully obtained the genotypes of 889 individuals (Supplementary Table 1). Distribution of allelic frequencies and forensic parameters at 15 STR loci for the studied populations are shown in **Table 1**. A total of 174 different alleles were observed in Ili Kazakh population on 15 autosomal STRs. The combined power of exclusion (CPE) was 0.99999998395 and combined power of discrimination (CPD) was 99.999999999999798%. FGA was the most polymorphic STR with 20 alleles with the frequencies ranging from 0.00056 to 0.1918, while THO1 was the least polymorphic STR among the 15 autosomal STRs ~~,~~ with only 7 alleles. GD values ranged from 0.6429 (TPOX) to 0.8797 (D2S1338). The observed heterozygosity and excepted heterozygosity ranged from 0.6378, 0.64302 (TPOX) to 88639, 0.87944 (D2S1338), respectively. All studied 15 autosomal STRs were fairly informative with PIC values ranged from 0.5903 (TPOX) to 0.8671 (D2S1338~~),~~) (PIC > 0.5), ~~hence all studied 15 autosomal STRs were fairly informative (PIC > 0.5)~~ which make these STRs suitable for differentiation of individuals and for paternity testing in Ili Kazakh ethnic group.

Hardy-Weinberg equilibrium (HWE)

Out of 15 loci, we observed thirteen loci were in Hardy-Weinberg Equilibrium (HWE) in the Ili Kazakh population ($p > 0.05$). However, D5S818 and D21S11 showed HWE P-values of 0.00674 and 0.00794, respectively **(Supplementary Table 2)**. Conversely, when we applied a sequential Bonferroni correction (Benjamini and Hochberg, 1995) to reduce the so-called "multiple comparison problems" (where for a significant p-value of 0.5, 5% of tests are likely to be significant by chance), no loci were found to be out of HWE.

Linkage equilibrium (LE)

In the exact tests for linkage equilibrium (LE), initially 11 pairwise combinations of STR loci showed the *p*-values below 0.05 and thus showing LD **(Supplementary Table 3)**. After applying sequential Bonferroni correction (Benjamini and Hochberg, 1995), ~~out of 11 pairwise combinations~~ only 5 pairs out of 11 pairwise combinations were still in LE. These five pairs were (D2S1338, D5S818), (D13S317, D5S818), (FGA, TPOX), (FGA, D19S433), and (D21S11, D8S1179). These pairs ~~which were~~ in LD may be the results of mutation, recombination, founder effects, genetic admixture or natural selection. To check the population hierarchy existence in Kazakh ethnic group and 7 other East Asian populations (Uzbek, Kyrgyz, Manchu, Mongol, Tibetan, Han from Liaoning and Yanb~~a~~ian Korean) with raw genotypic data, we check the genetic homozygosity or heterozygosity via ~~principal component analysis (~~PCA~~)~~. A total of 3.01% genetic variations can be extracted by the first three PCs **(Figure 1A~D)**. Above mentioned PCA results are later confirmed with pairwise *Fst* genetic distances **(Supplementary Table 4)** and phylogenetic relationship reconstruction **(Figure S1)**.

Comparison with other populations

We have compared our currently studied population with previously published populations from China and worldwide using an AMOVA for comparison by employing the same 15 STR loci. Genetic distances (F*st*) and associated *p*-values for each locus are given in **Supplementary Table 5**. We have observed close genetic relation between Ili Kazakh and other Xinjiang minority groups (Uzbek, Kyrgyz and Mongols) at 11 STRs while Uyghurs have similarities only at 3 STRs (D7S820, THO1 and vWA). Kazakhs have more genetic similarities with Mongols, which is also confirmed with previous studies (Adnan et al., 2018a; Bai et al., 2018; Zerjal et al., 2003). Initially, we have applied PCA to normalized allele frequencies at the 15 STR loci between Ili Kazakhs and 71 other Chinese populations (Uyghur, Uzbek, Manchu, Kyrgyz, Mongols, Hui, Yi, Tibetan, Kazakh, Bai, Vietnamese, Miao, Zhuang, Hani, Xibe, Korean and Han). Ili Kazakh population clustered in the upper--middle position of the plot with Uyghur, Kazakh and Mongol populations of Xinjiang **(Figure 2)**. Genetic distances between Ili Kazakhs and other 70 reference populations based on Nei's formula are listed in **Supplementary Table 6**. These values were used to construct a neighbor-joining tree between Ili Kazakhs and 70 other populations **(Figure S2)**. The closest genetic distances were observed between previously studied Ili Kazakhs (0.0019) and Uyghurs from Hotan (0.0056). A heat map of genetic matrix showed that Kazakh, Uyghur and Tibetan ethnic groups showed significant genetic variations when compared with other East Asian ethnic groups **(Figure S3)**. We also have performed interactivity test between these populations and found the result which was in consistence with above-mentioned PCA and Nei's formula results **(Supplementary Table 7 and Figure S4)**. To explore the genetic similarities and differences between Ili Kazakhs and other 32 worldwide populations, we again performed PCA to normalize allele frequencies at the 15 STR loci and found Ili Kazakh lined up with Australian Asians **(Figure S5)**. A genetic distance between Ili Kazakhs and other 32 worldwide reference populations based on Nei's formula (Supplementary Table 8) was used to construct the N-J tree **(Figure S6)**. We found that Australian Asians (0.0158) and Koreans have closest genetic distance while South African amaZulu (0.1441) and South African amaXhosa (0.1919) has the the longest largest distance among the studied populations. Interactivity test was also performed between Ili Kazakhs and 32 worldwide populations, which was in consistence with above-mentioned PCA and Nei's formula results **(Supplementary Table 9 and Figure S7)**.

Cluster analysis with STRUCTURE

Genetic landscapes of Ili Kazakh were further dissected by employing the model-based clustering algorithm in STRUCTURE software in the context of the genetic variations from Manchu, Mongolian, Kyrgyz, Uzbek, Liao Ning Han Chinese, Jilin Korean, and Tibetan. As shown in **Figure 3**, we identify the best optimal predefined populations in fiveat (K=5). Ili Kazakh shared most of the genetic components with Uzbek, Kyrgyz and Mongolian (magenta component), but shared and shares few genetic alleles or genetic drift with Tibetan of the in green component and Liaoning Han and Manchu of thein yellow component (**Figure 3B**). We can also identify a common blue component existing in all of these populations with different proportions. When the K values increasing from 2 to 10, a new ancestry component appears in each population with different followed proportions (**Figure 3C**). In total, two genetical clusters were observed: one comprises Manchu, Han Chinese, Jilin Korean and Tibetans, which are typical East Asian populations with the dominant East Asian ancestry components; while the other one consists of Kazakh, Kyrgyz, Uzbek and Mongolian populations, which are decedents of ancient Altai-speaking populations residing in

central and north Asia. The genetic differences between the above two clusters observed in this study may be explained by the complicated genetic admixture between genetic sources from South Asian, East Asian and West Eurasians (Damgaard et al., 2018). Our STRUCTURE results demonstrate that Ili Kazakhs areis genetically closer with to Altai-speaking populations than to other East Asian groups, which is consistent with linguistic affinity and also the genomic results on the basis of whole-genome sequencing and high-density genotyping data (Bai et al., 2018; Xu et al., 2008; Xu and Jin, 2008).

Forensic parameters of the 10 Y-STR loci

Allelic frequencies, haplotype diversity (GD) and the number of observed alleles for 10 Y-STRs are summarized in **(Supplementary Tables 10)**. Allelic frequencies ranged from 0.0015 to 0.6464. Among the single copy Y-STRs, DYS458 was the most polymorphic allele (GD= 0.8091) with 17 alleles, while the least polymorphic STRs was DYS391 (GD= 0.5023) with 6 alleles. Overall DYS385a/b has the highest gene diversity value 0.8357 with 70 different alleles. A total of 496 haplotypes were observed with haplotype diversity (GD) of 0.9985 and discriminations capacity (DC) of 0.7526. Random matching probability (RMP) was 0.0015 and 61.76% haplotypes were unique as shown in **Table 2**.

MDS based on the 10 Y- STRs

Multidimensional scaling analysisMDS was performed between Ili Kazakh and other 47 populations based on *Rst* values at 10 Y-STRs. *Rst* values are summarized in **(Supplementary Table 11)** and the results are shown in **(Figure S8)**. Ili Kazakhs were genetically closer to a Russian ethnic group Tambovskaja (-0.00436), , a Russian ethnic group followed by another Russian ethnic group Lipezkaja (-0.00388), while Lebanese (0.21632) and Iraqi (0.24607) has the greatlargest genetic differences among the studied groups based on *Rst*. According to the MDS plot, Kazakh populations are on the lower left side on the lower half along with Filipino and European populations. We performed interactivity test and found Kazakh population aligned with Ladakh Himalaya followed by Kazakhs from Kazakhstan **(Figure S9)**.

Phylogenetic analysis based on the 10 Y-STRs

We constructed an N-J tree **(Figure 4)** based on Reynolds genetic distance **(Supplementary Table 12)** between Ili Kazakh and other 47 populations and found they are lined up with Ladakh Himalayas and, a Russian ethnic group Lipezkaja a Russian ethnic group. We also generated a heat map using the genetic matrix **(Figure 5)** and found there wereare minor genetic variations among studied populations with the exceptions of Middle Eastern populations and a few Ssouthwest Asian populations.

Conclusion

In conclusion, we have presented the first comprehensive study which focusinged on the characterization of 25 STR loci including (15 autosomal and 10 Y-STRs) in the Ili Kazakh was performed. We found the AGCU EX16+10Y Expressmarker 16+10Y STR Kit (EX16+10Y) is appropriate for population studies, paternity testing and forensic

identity testing in the Kazakh population. Additionally we have studied the genetic affinity~~origin~~ of Kazakh population by comparing with 71 local and 32 worldwide populations on autosomal STRs and 48 worldwide populations on 10 Y-STRs. Population genetic analyses indicated that the Kazakhs had a close genetic relationship with Mongols and other Xinjiang minorities such as ~~(~~Kyrgyz and Uzbek~~)~~, which is in accordance with historical documents previously described ~~stories by historians~~ (Esposito, 1999; Weller, 2006). The data generated in this study will serve as a reference database in ~~which is~~ addressing both autosomal and Y-STR related cases at the same time for Kazakhs.

ACKNOWLEDGEMENTS

This study was financially supported by the National Natural Science Foundation of China (31801040), Nanqiang Outstanding Young Talents Program of Xiamen University (X2123302), and Fundamental Research Funds for the Central Universities (ZK1144).

COMPLIANCE WITH ETHICAL STANDARDS

Conflict of interest

The authors declare that they have no competing interests.

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the China Medical University, Shenyang, Liaoning Province, People's Republic of China and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Informed consent

Informed consent was obtained from all individual participants included in the study.

Data availability

All our data are submitted as supplementary materials.

AUTHOR CONTRIBUTIONS

J.F., C.W., and A.A. designed this study, A.A wrote the manuscript, A.A., K.K., J.G., and G.H, conducted the experiment, A.A., G.H., K.K, A.R., Se. H., S.H. J.F., and C.W.~~,~~ analyzed the results, A.A., and C.W. modified the manuscript. All authors reviewed the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

FIGURES:

Figure 1: Genetic homology between Ili Kazakhs and 7 other population from China revealed by principle component analysis.

Figure 2: Principal component analysis based on the 15 autosomal STR loci of Ili Kazakh population of Xinjiang and already published 70 populations from China.

Figure 3: Estimated population genetic structure of 8 different populations at K = 2 to K=10.

Figure 4: Neighbor-joining phylogenetic tree based on pairwise Reynolds genetic distance values of overlapping 10 Y STRs between Ili Kazakhs population of Xinjiang and already published 47 worldwide populations.

Figure 5: A heat map tree based on pairwise Reynolds genetic distance values of overlapping 10 Y STRs between Ili Kazakhs population of Xinjiang and already published 47 worldwide populations.

Figure S1: Genetic homology between Ili Kazakhs and 7 other population from China revealed by NJ phylogenetic tree based on Fst values.

Figure S2: Neighbor-joining phylogenetic tree based on Nei's genetic distance between Ili Kazakhs and other 70 reference populations across China.

Figure S3: A heat map of Nei's genetic distance values between Ili Kazakhs and other 70 reference populations across China.

Figure S4: Interactivity test based on allelic frequencies of overlapping 15 autosomal STRs between Ili Kazakhs and other 70 reference populations across China.

Figure S5: Principal component analysis based on the frequency of overlapping 15 autosomal STRs between Ili Kazakh population of Xinjiang and already published 32 worldwide populations.

Figure S6: Neighbor-joining phylogenetic tree based on Nei's genetic distance between Ili Kazakh population of Xinjiang and already published 32 worldwide populations.

Figure S7: Interactivity test based on allelic frequencies of overlapping 15 autosomal STRs between Ili Kazakhs population of Xinjiang and already published 32 worldwide populations.

Figure S8: A two-dimensional multidimensional scaling plot of Ili Kazakh population of Xinjiang and other regional populations based on pairwise Rst values.

Figure S9: Interactivity test based on allelic frequencies of overlapping 10 Y STRs between Ili Kazakhs population of Xinjiang and already published 47 worldwide populations.

TABLES:

Table 1: Allelic frequencies and forensic statistical parameters of 15 autosomal STRs in the Kazakh population from Ili Kazak Autonomous Prefecture from Xinjiang (n=889)

Table 2: Forensic statistical parameters of 10 Y STRs in the Kazakh population from Ili Kazak Autonomous Prefecture from Xinjiang (n=659)

ELECTRONIC SUPPLEMENTARY MATERIALS

Table S1: The haplotypes of 25 STR loci in Ili Kazakhs from Xinjiang using AGCU Expressmarker 16 and 10Y-STR Kit (EX16+10Y) (n=889)

Table S2: Sequential Bonferroni corrections for p values from exact tests for HWE of 15 autosomal STRs in Kazakh population from Ili Kazak Autonomous Prefecture from Xinjiang (n=889)

Table S3: p-values from exact tests for linkage equilibrium (LE) employing pairwise combinations of 15 autosomal STRs in Kazakh population from Ili Kazak Autonomous Prefecture from Xinjiang (n=889)

Table S4: Pairwise Fst genetic distances between Kazakh population from Ili Kazak Autonomous Prefecture from Xinjiang and 7 Chinese reference populations on the basis of raw genotype data of 15 STRs

Table S5: Genetic distance (Fst) and associated p-values for population differentiation between Kazakh population from Ili Kazak Autonomous Prefecture from Xinjiang and other published populations

Table S6: Nei's pairwise genetic distance between Kazakh population from Ili Kazak Autonomous Prefecture from Xinjiang and other published populations

Table S7: Interactivity test between Kazakh population from Ili Kazak Autonomous Prefecture from Xinjiang and other published populations

Table S8: Nei's pairwise genetic distance between Kazakh population from Ili Kazak Autonomous Prefecture from Xinjiang and other Worldwide published 32 populations

Table S9: Interactivity test between Kazakh population from Ili Kazak Autonomous Prefecture from Xinjiang and other 32 Worldwide published populations

Table S10: Allelic distribution of 10 Y STRs in Kazakh population from Ili Kazak Autonomous Prefecture from Xinjiang

Table S11: Pairwise Rst genetic distance between Kazakh population from Ili Kazak Autonomous Prefecture from Xinjiang and other published populations

Table S12: Renolyd distance between Kazakh population from Ili Kazak Autonomous Prefecture from Xinjiang and other published populations

REFERENCES:

Adnan, A., Rakha, A., Kasim, K., Noor, A., Nazir, S., Hadi, S., Pang, H., 2018a. Genetic characterization of Y-chromosomal STRs in Hazara ethnic group of Pakistan and confirmation of DYS448 null allele. International Journal of Legal Medicine. https://doi.org/10.1007/s00414-018-1962-x

Adnan, A., Rakha, A., Lao, O., Kayser, M., 2018b. Mutation analysis at 17 Y-STR loci (Yfiler) in father-son pairs of male pedigrees from Pakistan. Forensic Science International: Genetics. https://doi.org/10.1016/j.fsigen.2018.07.001

Adnan, A., Ralf, A., Rakha, A., Kousouri, N., Kayser, M., 2016. Improving empirical evidence on differentiating closely related men with RM Y-STRs: A comprehensive pedigree study from Pakistan. Forensic Sci Int Genet 25, 45–51. https://doi.org/10.1016/j.fsigen.2016.07.005

Bai, H., Guo, X., Narisu, N., Lan, T., Wu, Q., Xing, Y., Zhang, Yong, Bond, S.R., Pei, Z., Zhang, Yanru, Zhang, Dandan, Jirimutu, J., Zhang, Dong, Yang, X., Morigenbatu, M., Zhang, L., Ding, B., Guan, B., Cao, J., Lu, H., Liu, Yiyi, Li, W., Dang, N., Jiang, M., Wang, S., Xu, H., Wang, D., Liu, C., Luo, X., Gao, Y., Li, X., Wu, Z., Yang, L., Meng, F., Ning, X., Hashenqimuge, H., Wu, K., Wang, B., Suyalatu, S., Liu, Yingchun, Ye, C., Wu, H., Leppälä, K., Li, L., Fang, L., Chen, Y., Xu, W., Li, T., Liu, X., Xu, X., Gignoux, C.R., Yang, H., Brody, L.C., Wang, J., Kristiansen, K., Burenbatu, B., Zhou, H., Yin, Y., 2018. Whole-genome sequencing of 175 Mongolians

uncovers population-specific genetic architecture and gene flow throughout North and East Asia. Nat. Genet. 50, 1696–1704. https://doi.org/10.1038/s41588-018-0250-5

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society, Series B Vol. 57, 289–300.

Damgaard, P. de B., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., Korneliussen, T., Moreno-Mayar, J.V., Pedersen, M.W., Goldberg, A., Usmanova, E., Baimukhanov, N., Loman, V., Hedeager, L., Pedersen, A.G., Nielsen, K., Afanasiev, G., Akmatov, K., Aldashev, A., Alpaslan, A., Baimbetov, G., Bazaliiskii, V.I., Beisenov, A., Boldbaatar, B., Boldgiv, B., Dorzhu, C., Ellingvag, S., Erdenebaatar, D., Dajani, R., Dmitriev, E., Evdokimov, V., Frei, K.M., Gromov, A., Goryachev, A., Hakonarson, H., Hegay, T., Khachatryan, Z., Khaskhanov, R., Kitov, E., Kolbina, A., Kubatbek, T., Kukushkin, A., Kukushkin, I., Lau, N., Margaryan, A., Merkyte, I., Mertz, I.V., Mertz, V.K., Mijiddorj, E., Moiyesev, V., Mukhtarova, G., Nurmukhanbetov, B., Orozbekova, Z., Panyushkina, I., Pieta, K., Smrčka, V., Shevnina, I., Logvin, A., Sjögren, K.-G., Štolcová, T., Taravella, A.M., Tashbaeva, K., Tkachev, A., Tulegenov, T., Voyakin, D., Yepiskoposyan, L., Undrakhbold, S., Varfolomeev, V., Weber, A., Wilson Sayres, M.A., Kradin, N., Allentoft, M.E., Orlando, L., Nielsen, R., Sikora, M., Heyer, E., Kristiansen, K., Willerslev, E., 2018. 137 ancient human genomes from across the Eurasian steppes. Nature 557, 369–374. https://doi.org/10.1038/s41586-018-0094-2

Esposito, J.L. (Ed.), 1999. Central Asia and China: The Oxford History of Islam, in: The Oxford History of Islam. Oxford University Press, Oxford, p. 433.

Excoffier, L., Lischer, H.E.L., 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources 10, 564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x

Gouy, A., Zieger, M., 2017. STRAF-A convenient online tool for STR data evaluation in forensic genetics. Forensic Sci Int Genet 30, 148–151. https://doi.org/10.1016/j.fsigen.2017.07.007

Hammond, H.A., Jin, L., Zhong, Y., Caskey, C.T., Chakraborty, R., 1994. Evaluation of 13 short tandem repeat loci for use in personal identification applications. Am. J. Hum. Genet. 55, 175–189.

Kayser, M., Kittler, R., Erler, A., Hedman, M., Lee, A.C., Mohyuddin, A., Mehdi, S.Q., Rosser, Z., Stoneking, M., Jobling, M.A., Sajantila, A., Tyler-Smith, C., 2004. A Comprehensive Survey of Human Y-Chromosomal Microsatellites. The American Journal of Human Genetics 74, 1183–1197. https://doi.org/10.1086/421531

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Molecular Biology and Evolution 33, 1870–1874. https://doi.org/10.1093/molbev/msw054

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Sánchez-Diz, P., Acosta, M.A., Fonseca, D., Fernández, M., Gómez, Y., Jay, M., Alape, J., Lareu, M.V., Carracedo, A., Restrepo, C.M., 2009. Population data on 15 autosomal STRs in a sample from Colombia. Forensic Sci Int Genet 3, e81-82. https://doi.org/10.1016/j.fsigen.2008.08.002

Weller, R.C., 2006. Rethinking Kazakh and Central Asian nationhood: a challenge to prevailing western views. Asia Research Associates, Los Angeles.

Wyatt, D.J., Di Cosmo, N., 2011. Political frontiers, ethnic boundaries and human geographies in Chinese history. Routledge, London.

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. https://doi.org/10.1016/j.ajhg.2008.01.017

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. https://doi.org/10.1016/j.ajhg.2008.08.001

Zerjal, T., Xue, Y., Bertorelle, G., Wells, R.S., Bao, W., Zhu, S., Qamar, R., Ayub, Q., Mohyuddin, A., Fu, S., Li, P., Yuldasheva, N., Ruzibakiev, R., Xu, J., Shu, Q., Du, R., Yang, H., Hurles, M.E., Robinson, E., Gerelsaikhan, T., Dashnyam, B., Mehdi, S.Q., Tyler-Smith, C., 2003. The Genetic Legacy of the Mongols. The American Journal of Human Genetics 72, 717–721. https://doi.org/10.1086/367774

Zhan, X., Adnan, A., Zhou, Y., Khan, A., Kasim, K., McNevin, D., 2018. Forensic characterization of 15 autosomal STRs in four populations from Xinjiang, China, and genetic relationships with neighboring populations. Scientific Reports 8. https://doi.org/10.1038/s41598-018-22975-6

Zhou, H., Bi, G., Zhang, C., Liu, Y., Chen, R., Li, F., Mei, X., Guo, Y., Zheng, W., 2016. Developmental validation of forensic DNA-STR kits: Expressmarker 16 + 10Y and expressmarker 16 + 18Y. Forensic Science International: Genetics 24, 1–17. https://doi.org/10.1016/j.fsigen.2016.05.011