The Heart of Silk Road “Xinjiang,” Its Genetic Portray, and Forensic Parameters Inferred From Autosomal STRs

Atif Adnan1,2,3†, Adeel Anwar4†, Halimureti Simayijiang5†, Noor Farrukh5, Sibte Hadi2†, Chuan-Chao Wang3,6,7*, and Jin-Feng Xuan1*

1Department of Forensic Genetics, School of Forensic Medicine, China Medical University, Shenyang, China, 2Department of Forensic Sciences, College of Criminal Justice, Naif Arab University of Security Sciences, Riyadh, Kingdom of Saudi Arabia, 3Department of Anthropology and Ethnology, School of Sociology and Anthropology, Institute of Anthropology, Xiamen University, Xiamen, China, 4Department of Orthopedic Surgery, The 3rd Affiliated Hospital of Chongqing Medical University, Chongqing, China, 5Department of Forensic Medicine, School of Basic Medical Sciences, Binzhou Medical University, Yantai, China, 6State Key Laboratory of Cellular Stress Biology, School of Life Sciences, National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China, 7State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen, China

The Xinjiang Uyghur Autonomous Region of China (XUARC) harbors almost 50 ethnic groups including the Uyghur (UGR: 45.84%), Han (HAN: 40.48%), Kazakh (KZK: 6.50%), Hui (HUI: 4.51%), Kyrgyz (KGZ: 0.86%), Mongol (MGL: 0.81%), Manchu (MCH: 0.11%), and Uzbek (UZK: 0.066%), which make it one of the most colorful regions with abundant cultural and genetic diversities. In our previous study, we established allelic frequency databases for 14 autosomal short tandem repeats (STRs) for four minority populations from XUARC (MCH, KGZ, MGL, and UZK) using the AmpFlSTR® Identifier PCR Amplification Kit. In this study, we genotyped 2,121 samples using the GoldenEye™ 20A Kit (Beijing PeopleSpot Inc., Beijing, China) amplifying 19 autosomal STR loci for four major ethnic groups (UGR, HAN, KZK, and HUI). These groups make up 97.33% of the total XUARC population. The total number of alleles for all the 19 STRs in these populations ranged from 232 (HAN) to 224 (KZK). We did not observe any departures from the Hardy–Weinberg equilibrium (HWE) in these populations after sequential Bonferroni correction. We did find minimal departure from linkage equilibrium (LE) for a small number of pairwise combinations of loci. The match probabilities for the different populations ranged from 1 in \(1.66 \times 10^{-23}\) (HAN) to 6.05 \(\times 10^{-24}\) (HUI), the combined power of exclusion ranged from 0.999 999 988 (HUI) to 0.999 999 993 (UGR), and the combined power of discrimination ranged from 0.999 999 999 999 999 999 983 (HAN) to 0.999 999 999 999 999 999 999 997 (UGR). Genetic distances, principal component analysis (PCA), STRUCTURE analysis, and the phylogenetic tree showed that genetic affinity among studied populations is consistent with linguistic, ethnic, and geographical classifications.

Keywords: Uyghur, Kazakh, hui, han, autosomal STRs, phylogenetics, allelic frequency database
1 INTRODUCTION

The Altai is a mountain range in Central and East Asia and stretches around Kazakhstan and Russia in the west while Xinjiang in the northwest China and Mongolia in the East. In ancient times, Europeans were settled on the western side and Asians on the eastern side of these mountains (Li et al., 2019). Later on, the region was used as a corridor by European and Asian populations. This corridor played an important role in the formation of new ethnic groups and diversity of populations today (Ovodov et al., 2011). The Xinjiang Uyghur Autonomous Region (XUAR) is the most diversified region in China, and it is the home of almost 50 ethnic groups. XUAR has played a vital role in the early ages because it was connecting not only the Altai mountain range corridor but also western Eurasia and eastern Eurasia (Esposito, 1999). It was also the main hub for the famous Silk Road, which linked trade between the Middle East, East Asia, Central Asia, South Asia, and Europe (Esposito, 1999). XUAR is divided into two basins, Dzungarian (North Basin) and Tarim (South Basin). Many ethnic groups, including the Uyghur (UGR), Kazakh (KZK), Hui (HUI), Han (HAN), Manchu (MCH), Mongols (MGL), Kirgiz (KGZ), and Uzbek (UZK) have lived there for hundreds of years (Millward, 2007).

The Uyghurs are one of the officially recognized ethnic minorities in China, and most of them live in the southern basin (Tarim Basin). The official language is Mandarin; however, other languages are spoken in Xinjiang. The Uyghur language belongs to the Karluk branch of the Turkic language family. Since ancient times, Xinjiang has been a multiethnic region and has been mentioned in Chinese records from the 206 BC to 220 AD (Millward, 2007). The Uyghurs are the main ethnic group in Xinjiang and represent 45.84% of the total population (Millward, 2007; Wyatt and Di Cosmo, 2011).

According to some historians, the Kazakh (KZK) population emerged around the 13th century. The ancestors of modern-day Kazakh were the result of an admixture of this group with Mongol tribes from Eastern Chagatai Khanate. They speak the Kazakh language which also belongs to the Turkic language family. Till the 15th century, they were part of nomadic tribes. Kazakh is the third largest ethnic group of Xinjiang and represents 6.50% of the total Xinjiang population (Weller, 2006; Weatherford, 2012).

Ancestors of Hui came to China from Central Asia and Islamic Persia as handcraftsmen, merchants, scholars, and soldiers. This human movement started in the seventh century and continued until the 13th century. After settling in China, they started intermingling with Han Chinese, Mongols, and Uyghurs. Ultimately, their phenotype, cultural characteristics, and language became thoroughly Chinese. Hui is the fourth largest ethnic group in Xinjiang and represents 4.51% of the total Xinjiang population (Dillon, 1999).

According to historical records, all Han Chinese can trace their origins to the Huaxia tribes which were agricultural tribes living along the Yellow River and formed during the Shang and Zhou dynasties (21st–8th centuries BC) (Cioffi-Revilla and Lai, 1995). With the rise of the Han Dynasty, all these tribes formed a common tribe/ethnic group known as the Han Chinese (Du and Vincent, 1993). Han Chinese is the world’s largest ethnic group and accounts for 18% of the worldwide population. Han is the major dominating group not only in mainland China (92%) but also in Singapore (75%). Han is the second largest ethnic group of Xinjiang and represents 40.48% of the total Xinjiang population.

DNA regions with repeat units of 2–6 bps in length are called short tandem repeats (STRs), also identified as microsatellites. STR markers are present throughout the human genome and usually have stable polymorphisms, short sequence lengths, and a dense, uniform chromosomal distribution, which makes their detection and analysis smooth using PCR and sequencing (Hammond et al., 1994; Sánchez-Diz et al., 2009). In forensic investigations such as paternity cases, rape cases, kinship analysis, and missing person analysis; STRs are considered as markers of choice because of their high polymorphism (Adnan et al., 2017; 2018b; 2018c). The Goldeneye™ 20A is a five-dye kit (Beijing PeopleSpot Inc., Beijing, China), which includes 16 combined DNA index system (CODIS) core STR loci along with Penta E, Penta D, and D6S1043 (Huang et al., 2013).

There are few studies available that focus on the characterization of autosomal STRs in the main ethnic groups of Xinjiang such as Uyghur (Yuan et al., 2016) and Kazakh (Zhang et al., 2016b). The drawbacks with previous studies are that they only focus on one or a maximum of two ethnic groups. This study focused on four main ethnic groups UGR, KZK, HUI, and HAN from Xinjiang to genotype 19 autosomal STRs using the Goldeneye 20A kit. We combined the data generated in this study with our previously published work focusing on Manchu (MCH), Mongols (MGL), Kirgiz (KGZ), and Uzbek (UZK) from Xinjiang, We then compared the genotypic data with 97 other worldwide populations.

2 MATERIALS AND METHODS

2.1 Samples and DNA extraction

Blood samples were collected from 2,121 (533 Uyghur (F = 230, M = 303), 436 Kazakh (F = 204, M = 232), 593 Hui (F = 257, M = 336), and 559 Han (F = 213, M = 346) unrelated healthy individuals from the XUAR. All participants gave their informed consent either orally and with thumbprints (in case they could not write) or in writing after the study aims and procedures were carefully explained to them in their language. The study was approved by the ethical review board (dated March 20, 2019, with approval reference no. 2019-84-P) of the China Medical University, Shenyang, Liaoning Province, People’s Republic of China. All blood samples were stored at -20°C before DNA extraction. DNA was isolated from blood using the ReliaPrep™ Blood gDNA Miniprep System (Promega, Madison, WI, USA) according to the manufacturer’s instructions. The quantities of extracted DNA samples were determined using a
### TABLE 1 | Forensic parameters of the 19 autosomal STR loci in four (Uyghur, Kazakh, Hui and Han) populations residing in Xinjiang Uyghur Autonomous Region (XUAR) in China.

| STR Loci | Uyghur          | Kazakh          | Hui              | Han              |
|----------|-----------------|-----------------|------------------|------------------|
|          | GD   | PIC  | PM   | PD   | Hobs | PE   | TPI   | GD   | PIC  | PM   | PD   | Hobs | PE   | TPI   |
| CSF1PO   | 0.6505 | 0.6868 | 0.1163 | 0.8837 | 0.4760 | 1.9025 | 0.7355 | 0.6686 | 0.1191 | 0.8909 | 0.7408 | 0.4942 | 1.9292 |
| D12S317  | 0.7465 | 0.7659 | 0.8417 | 0.7959 | 0.7207 | 2.7499 | 0.8186 | 0.8488 | 0.8647 | 0.7828 | 0.8497 | 0.8180 | 2.7499 |
| D13S317  | 0.7411 | 0.8065 | 0.8616 | 0.8950 | 0.7459 | 2.7499 | 0.7396 | 0.8046 | 0.8967 | 0.7982 | 0.8497 | 0.8046 | 2.7499 |
| D16S539  | 0.7411 | 0.8065 | 0.8616 | 0.8950 | 0.7459 | 2.7499 | 0.7396 | 0.8046 | 0.8967 | 0.7982 | 0.8497 | 0.8046 | 2.7499 |
| D18S51   | 0.7411 | 0.8065 | 0.8616 | 0.8950 | 0.7459 | 2.7499 | 0.7396 | 0.8046 | 0.8967 | 0.7982 | 0.8497 | 0.8046 | 2.7499 |
| D19S433  | 0.7411 | 0.8065 | 0.8616 | 0.8950 | 0.7459 | 2.7499 | 0.7396 | 0.8046 | 0.8967 | 0.7982 | 0.8497 | 0.8046 | 2.7499 |
| D21S11   | 0.7411 | 0.8065 | 0.8616 | 0.8950 | 0.7459 | 2.7499 | 0.7396 | 0.8046 | 0.8967 | 0.7982 | 0.8497 | 0.8046 | 2.7499 |
| D2S1338  | 0.7411 | 0.8065 | 0.8616 | 0.8950 | 0.7459 | 2.7499 | 0.7396 | 0.8046 | 0.8967 | 0.7982 | 0.8497 | 0.8046 | 2.7499 |
| D3S1358  | 0.7411 | 0.8065 | 0.8616 | 0.8950 | 0.7459 | 2.7499 | 0.7396 | 0.8046 | 0.8967 | 0.7982 | 0.8497 | 0.8046 | 2.7499 |
| D5S818   | 0.7411 | 0.8065 | 0.8616 | 0.8950 | 0.7459 | 2.7499 | 0.7396 | 0.8046 | 0.8967 | 0.7982 | 0.8497 | 0.8046 | 2.7499 |
| D6S1043  | 0.7411 | 0.8065 | 0.8616 | 0.8950 | 0.7459 | 2.7499 | 0.7396 | 0.8046 | 0.8967 | 0.7982 | 0.8497 | 0.8046 | 2.7499 |
| D8S1179  | 0.7411 | 0.8065 | 0.8616 | 0.8950 | 0.7459 | 2.7499 | 0.7396 | 0.8046 | 0.8967 | 0.7982 | 0.8497 | 0.8046 | 2.7499 |
| FGA      | 0.7411 | 0.8065 | 0.8616 | 0.8950 | 0.7459 | 2.7499 | 0.7396 | 0.8046 | 0.8967 | 0.7982 | 0.8497 | 0.8046 | 2.7499 |
| Penta D  | 0.7411 | 0.8065 | 0.8616 | 0.8950 | 0.7459 | 2.7499 | 0.7396 | 0.8046 | 0.8967 | 0.7982 | 0.8497 | 0.8046 | 2.7499 |
| Penta E  | 0.7411 | 0.8065 | 0.8616 | 0.8950 | 0.7459 | 2.7499 | 0.7396 | 0.8046 | 0.8967 | 0.7982 | 0.8497 | 0.8046 | 2.7499 |
| TH01     | 0.7411 | 0.8065 | 0.8616 | 0.8950 | 0.7459 | 2.7499 | 0.7396 | 0.8046 | 0.8967 | 0.7982 | 0.8497 | 0.8046 | 2.7499 |
| TPOX     | 0.7411 | 0.8065 | 0.8616 | 0.8950 | 0.7459 | 2.7499 | 0.7396 | 0.8046 | 0.8967 | 0.7982 | 0.8497 | 0.8046 | 2.7499 |
| vWA      | 0.7411 | 0.8065 | 0.8616 | 0.8950 | 0.7459 | 2.7499 | 0.7396 | 0.8046 | 0.8967 | 0.7982 | 0.8497 | 0.8046 | 2.7499 |

GD: gene diversity; PIC: polymorphism information content; Hobs: observed heterozygosity; PD: power of discrimination; PM: matching probability; PE: power of exclusion; TPI: typical paternity index.
NanoDrop spectrophotometer (Thermo Scientific, Wilmington, DE, USA), and the final concentration of DNA was diluted to 1–2 ng/μl.

2.2 PCR amplification
Multiplex PCR amplification of 19 autosomal STR loci (D18S51, D21S11, TH01, D3S1358, FGA, TPOX, D8S1179, vWA, CSF1PO, D16S539, D7S820, D13S317, D2S1338, D19S433, D5S818, D12S391, D6S1043, Penta D, Penta E, and the gender marker Amelogenin) was performed using the Goldeneye™ 20A kit (Beijing PeopleSpot Inc.) for all extracted DNA samples. One to 2 ng of extracted DNA was amplified according to the manufacturer’s recommended protocol. Thermal cycling was conducted in a GeneAmp PCR System 9700 thermal cycler (Applied Biosystems, Foster City, CA, USA) using the following conditions: 95°C for 5 min; 30 cycles of 94°C for 30 s, 60°C for 60 s, 70°C for 60 s; and a final extension of 60°C for 30 min.

2.3 Genotyping
Amplified products were prepared using the Goldeneye™ ORG 500 internal size standard (Beijing PeopleSpot Inc.) and HiDi formamide. All samples were electrophoresed using a 3,500 genetic analyzer (Applied Biosystems, Foster City, CA, USA) according to the Goldeneye™ 20A standard protocol. POP-4™ polymer (Life Technologies, Carlsbad, CA, USA) was used for capillary electrophoresis (CE). The Goldeneye™ 20A kit allelic ladder was run for each CE injection. GeneMapper™ Software version 4.0 (Life Technologies) was used for calling all alleles,
which was based on the allelic ladder which followed the ISFG recommendations in this regard (Gusmão et al., 2006; Bodner et al., 2016).

### 2.4 Quality control

All extraction and quantitation batches included sterile water (sH2O) as negative control. Negative (sH2O) and positive (AmpFISTR Control DNA 9947A) controls were employed for PCR amplification and capillary electrophoresis. All negative controls displayed an absence of amplified product while positive controls were consistent with known genotypes.

### 2.5 Statistical analysis

The allele frequencies of STR loci, exact tests of the Hardy–Weinberg Equilibrium (HWE), and pair linkage disequilibrium tests were calculated with PowerMarker V3.25 [13]. The values for matching probability (MP), power of discrimination (PD), polymorphism information content (PIC), power of exclusion (PE) duos, typical paternity index (TPI), gene diversity (GD), and expected heterozygosity (He) were calculated by using the PowerStats software (Ver 1.2, Promega, Madison, WI, USA) [14]. Nei’s standard genetic distance between current and previously published reference populations (Law et al., 2002; Berti et al., 2005; Tie et al., 2006; Barni et al., 2007; Deng et al., 2007; Maruyama et al., 2008; Demeter et al., 2010; Stepanov et al., 2010; Zhu et al., 2010; Deng et al., 2011; Kee et al., 2011; Xing et al., 2011; Yoo et al., 2011; Chen et al., 2012; Di Cristofaro et al., 2012; Westen et al., 2012; Gaviria et al., 2013; Hedjazi et al., 2013; Nasibov et al., 2013, 2013; Park et al., 2013; Bentayebi et al., 2014; Fuji et al., 2014; Almeida et al., 2015; Gaviria et al., 2015; Polito et al., 2015; Shaqique et al., 2015; Shrivasvata et al., 2015; Zhang, 2015b, 2015a; Zhang et al., 2015; Aguilar-Velázquez et al., 2016; Hossain et al., 2016; Nakamura et al., 2016; Ng et al., 2016; Park et al., 2016; Ramos-González et al., 2016; Ristow et al., 2016; Shan et al., 2016; Tokdemir et al., 2016; Vull et al., 2016; Wang et al., 2016; Zhang et al., 2016a, 2016b; Choi et al., 2017; Guerreiro et al., 2017; Jin et al., 2017; Liu et al., 2017; Moysés et al., 2017; Ossowski et al., 2017; Perveen et al., 2017; Singh and Nandini, 2017; Taylor et al., 2017; Wu et al., 2017; Yang et al., 2017; Adnan et al., 2018c; He et al., 2018a; Wang et al., 2018; Liu et al., 2019; Adnan et al., 2020) was computed using the Phylop 3.69 package [15]. Principal component analysis (PCA) based on allele frequency correlation and was conducted using the MVSP v3.22 software (http://www.kovcomp.com). The phylogenetic tree was constructed using MEGA X, and ancestry component dissection was explored by using STRUCTURE v.2.3.4 software (Porras-Hurtado et al., 2013). The model-based analysis employed the length of the
3 RESULTS AND DISCUSSION

3.1 Forensic parameters

The allelic frequencies for four Xinjiang ethnic groups are summarized in Supplementary Table S1. Allelic frequencies ranged from 0.0009 (CSF1PO) to 0.5262 (TPOX) in the UGR population, ranged from 0.0011 (CSF1PO) to 0.5676 (TPOX) in KZK, ranged from 0.0008 (CSF1PO) to 0.5210 (TPOX) in HUI, and ranged from 0.0008 (D16S539) to 0.5152 (TH01) in HAN. A total of 229, 224, 229, and 232 unique alleles were observed for 19 STRs in UGR, KZK, HUI, and HAN populations, respectively. When we compared our newly studied populations on overlapping 15 STRs with our previously published data, 163, 160, 167, and 170 unique alleles were observed in UGR, KZK, HUI, and HAN, respectively, while 152, 165, 153, and 168 unique alleles were observed in the MCH, MOG, KYZ, and UZK populations, respectively. The combined powers of discrimination (CPDs) on 19 STRs were 0.999 999 999 999 999 999 999 997, 0.999 999 999 999 999 999 999 996, 0.999 999 999 999 999 999 999 996, and 0.999 999 999 999 999 999 999 993 in UGR, KZK, HUI, and HAN, respectively. The combined powers of exclusion (CPE) for 19 STRs were 0.999 999 999 999 999 999 999 983 in UGR, KZK, HUI, and HAN, respectively. The
matching probabilities (MP) for the different populations were $2.21 \times 10^{-24}$, $3.98 \times 10^{-24}$, $6.05 \times 10^{-24}$, and $1.66 \times 10^{-24}$ in UGR, KZK, HUI, and HAN, respectively.

On combining the data for 15 STRs extracted from our previous studies for populations (MCH, MOG, KYZ, and UZK) and the current study (UGR, KZK, HUI, and HAN), CPDs were $0.999\ 999\ 999\ 999\ 999\ 984\ 833$, $0.999\ 999\ 999\ 999\ 999\ 990\ 057$, $0.999\ 999\ 999\ 999\ 999\ 996\ 333$, $0.999\ 999\ 999\ 999\ 999\ 998\ 244$, $0.999\ 999\ 999\ 999\ 999\ 997\ 679$, $0.999\ 999\ 999\ 999\ 999\ 995\ 805$, $0.999\ 999\ 999\ 999\ 999\ 999\ 993\ 455$, and $0.999\ 999\ 999\ 999\ 999\ 999\ 986\ 23$, respectively.

CPEs were $0.999\ 999\ 416$, $0.999\ 999\ 483$, $0.999\ 997\ 932$, $0.999\ 998\ 973$, $0.999\ 998\ 617$, $0.999\ 998\ 811$, $0.999\ 998\ 151$, and $0.999\ 998\ 15$, respectively. CMPs for the different populations were $1/1.51 \times 10^{17}$, $1/1.75 \times 10^{18}$, $1/3.66 \times 10^{18}$, $1/9.94 \times 10^{18}$, $1/2.32 \times 10^{18}$, $1/4.19 \times 10^{18}$, $1/6.54 \times 10^{18}$, and $1/1.37 \times 10^{17}$, respectively. Penta E showed the greatest heterozygosities and PDs in all four newly studied populations. Penta E was also the most polymorphic locus in terms of alleles in the UGR (20), KZK (21), HUI (23), and HAN (21) while THO1 (6), TPOX (6), TH01 (7), and TH01 (7) were least polymorphic. Other forensic parameters such as GD, PIC, MP, PE, TPI, and Penta E loci
exhibited a high level of informativeness while TPOX exhibited the least level of informativeness (Table 1). The data showed that the Goldeneye™ 20A panel can be used for forensic identification and parentage testing for the four ethnic groups in the XUAR of China.

3.2 Hardy–Weinberg equilibrium
All of the loci were in the Hardy–Weinberg Equilibrium (HWE) in the HUI population (p > 0.05), while only two loci for UGR (vWA and Penta E) and one STR locus for each HAN (D13S317) and KZK (D6S1043) were out of HWE. Subsequently, when we applied sequential Bonferroni correction (Benjamini and Hochberg, 1995) to mitigate the so-called multiple comparison problem (where for a significant p-value of 0.05, 5% of tests are likely to be significant by chance), none of the loci in any of the four populations were found to be out of HWE (Supplementary Table S2). In our previous study (Zhan et al., 2018), none of the loci were out of HWE in the KYZ population, while one for MCH (D7S820), two for MOG (CSF1PO, D19S433), and four for UZK (D18S51, D2S1338, D7S820, and FGA) were out of HWE. However, after sequential Bonferroni correction (Benjamini and Hochberg, 1995), all loci in the four populations studied conformed to HWE.

3.3 Linkage equilibrium
The phenomena of linkage disequilibrium (LD) can be a result of the association between adjacent alleles co-inherited from single, ancestral chromosomes. Particularly for tightly linked genes, if a selection favors individuals with particular combinations of alleles, then it produces LD which can persist for some time. LD between two loci decays gradually in proportion to the recombination rate and time as measured in number of generations. When mutations are under positive selection, the LD surrounding the mutations is maintained because of the hitchhiking effect; thus, longer haplotypes at high frequencies can be maintained within the population. Many LD-based methods have been developed to detect positive selection. Hudson et al. (Hudson et al., 1994) proposed the first method to detect positive selection by measuring haplotype patterns. Using the extended haplotype homozygosity (EHH) test, Sabeti et al. (Sabeti et al., 2006) developed a more robust method to detect positive selection by measuring longer haplotypes at high frequencies. This method was further refined by Voight et al. (Voight et al., 2006), who standardized the EHH test using the genome-wide empirical distributions of EHH. Based on a similar rationale, Wang et al. (Wang et al., 2007) developed a new version of the LD-based method called LD decay. LD can also be caused by the rate of mutation or recombination, random genetic drift, natural selection, founder effects, nonrandom mating, recent admixture, sampling effects, and population substructure (Chakravarti, 1999). Results of exact tests for linkage equilibrium (LE) displayed that p-values of 116 pairwise combinations of STR loci (UGR 32, KZK 31, HUI 32, and HAN 21) were lower than 0.05 and thus showing LE (Supplementary Table S3). Subsequently, when we applied sequential Bonferroni correction (Benjamini and Hochberg, 1995), only 17 pairs were out of LE. These pairs were D6S1043/D2S1338, D2S1338/D5S818, D18S51/D21S11, D12S391/TPOX, TPOX/Penta D, and D6S1043/D5S818 in the UGR population; D7S820/FGA, D18S51/D8S1179, D7S820/
Penta E, CSF1PO/D3S1358, and D6S1043/D8S1179 in the KZK population; D2S1338/D5S818, TH01/D18S51, D12S391/FGA, and D13S317/TH01 in the HUI population; and CSF1PO/D3S1358 and FGA/CSF1PO in the HAN population. We tested 171 pairwise LE tests in each population, and a maximum of six was out of LE in each population. The product rule is used to estimate the random chance of detecting a given STR profile within a population. This is done by multiplying the frequencies of each of the genotypes (combination of alleles) found at all loci in the STR profile (Butler and Butler, 2010). Presentation of the “product rule” for calculating the RMP across multiple STRs can be seen in the HAN population (only two pairs were out of LE) while in the other three (UGR, KZK, and HUI) populations it was unlikely to produce significant errors. The HAN population is not an endogamous population while the other three are endogamous populations. There is no trend of external mating in Xinjiang Muslim populations. The same results were observed in our previous study for the Kyrgyz and Uzbek populations.

3.4 Ancestry content analysis with structure

STRUCTURE analysis (which uses model-based clustering algorithm) was done for four populations (Xinjiang Han, Uyghur, Kazakh, and Xinjiang Hui) and publicly available data for seven populations (Mongol, Kyrgyz, Uzbek, Manchu, Liaoning Han, Tibetan, and Jilin Korean). The data were used to explore ancestry content and genetic landscapes of Xinjiang populations. The number of inferred clusters (K) varied from 2 to 10 with 10 repetitions of each K value and a total of 10,000 burnins and 10,000 Markov chain Monte Carlo (MCMC)
FIGURE 7 | Interactivity test between the Xinjiang Han population and 104 other reference populations from worldwide.
simulations for each repetition. We observed that the best optimal number of ancestral populations was five (K = 5). Uygur and Kazakh populations shared most of their genetic components with other Turkic-speaking populations such as Uzbek and Kyrgyz (yellow) while they shared a few genetic components with Tibetan (blue) and Han, Hui, Munch, and Korean populations (pink) (Figure 1). The pink component was a common component that was present in all populations studied. Moreover, two genetic clusters were observed. One cluster mainly contained populations that are descendants of ancient Altai-speaking populations, and the second cluster contained mainly East Asian populations. Uygur and Kazakh populations appeared to be genetically closer to Altai-speaking populations while Han and Hui populations were closer to East Asian groups. Results of K2-10 are shown in Supplementary Figure S1. PCA on raw genotype data of 11 populations showed 1.39% of variations, while in the second component, 0.95% of variations were observed. We observed two main clusters on these two (1&2) components: the first cluster contains Kyrgyz, Kazakh, and Mongol populations, while the second cluster contains Uzbek, Uygur, Manchu, Hui, Tibetan, Korean, and Han populations. On PC1 and PC3 (0.87%), a total of 2.26% of variations were observed. Here we observed three clusters. The first cluster contained Kyrgyz, Kazakh, and Mongol populations. The second cluster had Uzbek, Uygur, and Tibetan populations. In the third cluster, typical East Asian populations (Han, Hui, Manchu, and Korean) were grouped together.

PC2 (0.95%) and PC3 (0.87%) resulted in three clusters. In the first cluster, only the Tibetan population was grouped while in the second and third clusters all Altai-speaking populations and typical East Asian populations were present, respectively. PC1 and PC4 (0.84%) also gave us interesting results. Here, two clusters were formed. In the first cluster, the Tibetan population along with typical East Asian populations was grouped while in the second cluster all Altai-speaking populations were present. Results of PCA (Figure 2) were in support with STRUCTURE analysis. These results were also consistent with linguistic affinity and also whole-genome sequencing and high-density genotyping data (Xu et al., 2008; Xu and Jin, 2008; Bai et al., 2018).

3.5 Xinjiang and worldwide population comparison

We have compared the data of four populations with previously published populations from Xinjiang of northwest China and worldwide regions using AMOVA, employing available data for 15 STR loci. Genetic distances between the HAN population and 12 other populations from Xinjiang (Xinjiang Uygur, Xinjiang Hui, Xinjiang Kazakh, Xinjiang Kyrgyz, Xinjiang Manchu, Xinjiang Uzbek, Xinjiang Mongols, Uygur-Xinjiang-1, Kazakh-Xinjiang-1, Kumul-Uygur-Xinjiang-3, Uygur-Xinjiang-2, Kazakh-Xinjiang-2) based on Nei’s standard formula are listed in Supplementary Table S4A. These Nei’s standard genetic distance values were used to build a neighbor-joining tree (N-J tree) between Xinjiang Han and 12 other populations (Figure 3A). The Hui population (0.0046) from Xinjiang showed the closest genetic distance with HAN followed by the Mongol population (0.0244) from Xinjiang, while the Kazakh-1 population (1.1693) showed the greatest genetic distance which is followed by the Kyrgyz population (0.4547) from Xinjiang. The first three components (extracting 91.13% genetic variations) of PCA based on allelic frequencies of 15 STRs showed that Altai-speaking populations were closely linked (Figure 3B). A heat map (Figure 3C) using the genetic distances showed two clusters. In the first cluster, Manchu, Kyrgyz, Uzbek, and Kazakh-1 were grouped together while the second cluster contained other nine populations. According to Zhang et al. (Zhang et al., 2016a), Uyghurs are genetically closer to the central Asian population and Mongolian populations from East Asia. Jin et al. (Jin et al., 2017) found that these Turkic language-speaking groups placed themselves in the middle of European and East Asian populations. Feng et al. (Feng et al., 2017) used genome-wide human SNP array and found that the Uyghur population from XUAR have four major ancestral components, which were the result of two earlier admixed groups: one of them was from the West containing European (25%–37%) and West South Asian ancestries (12–20%), while the second one was from the East, with Siberian (15%–17%) and East Asian (29%–47%) ancestries. Results of MultiWaver showed us a two-wave admixture. The earliest wave was ∼3,750 years ago (ya), and a recent wave ∼750 ya. According to Seiduvaly et al. (Seiduvaly et al., 2020), the Kazakh population has a mixed ancestry containing East Asian (32.8%), European (30.8%), North Asians (28.9%), and South Asians (6%). Wen et al. (Wen et al., 2021) found that the Northwest Chinese Kyrgyz showed a high percentage of Y haplogroup R1a1a1b2a2a-Z2125, which is related to Bronze Age Siberians, while the second dominant haplogroup was C2b1a3a1-F3796, related to Medieval Niru’un Mongols, such as the Uissun tribe from Kazakhs. Again, Wen et al. (Wen et al., 2020) found that the Kazakh population from China showed the highest frequency (80%) of haplogroup C2b1a3a1-F3796 (previous C3*-Star Cluster) which is predominantly found in Mongolian descent populations. Wang et al. (Wang et al., 2019) found that the Hui population has about 70% in total of the paternal ancestry which could be traced back to East Asia and the left 30% to various regions in West Eurasia. Zhao et al. (Zhao et al., 2020) investigated that Mongolian and Kazakh groups derived 6%–40% of their ancestry from West Eurasia while 42%–64% of their ancestry was from East Asia. He et al. (He et al., 2018b) reported that the Uyghur population has 36.30% of European-related ancestry while Hui only have 3.66% of it. Liu et al. (Liu et al., 2018) reported that the Tibetan population and Hui population have a genetic affinity with East Asian populations, while the Uyghur population showed a similar genetic makeup with South Asian populations.

Genetic distances among the HAN population and 104 worldwide populations were calculated using Nei’s standard formula and summarized in Supplementary Table S3D. Among worldwide populations, Asians living in Australia
(0.0123) showed the closest association with the HAN population followed by the Chamorro population (0.0574); on the other hand, the Haitian population (0.2322) showed a distant association followed by the amaXhosa population (0.2251) from South Africa. These Nei’s genetic distance values were used to build a neighbor-joining tree (N-J tree) between Xinjiang Han and 104 worldwide populations (Figure 4). The first 10 components of PCA (PC1 = 25.06%, PC2 = 14.70%, PC3 = 9.80%, PC4 = 6.55%, PC5 = 6.02%, PC6 = 5.40%, PC7 = 4.19%, PC8 = 3.07%, PC9 = 2.39% and PC10 = 2.08%) extracted 79.31% of genetic variations (Figure 5). A heat map of the genetic distance matrix was also generated among 105 worldwide populations (Figure 6). Populations from Xinjiang showed their affinity with Central Asian, South West Asian, West Asian, and East European populations. An interactivity test between these 105 worldwide populations showed that the results were consistent with the PCA and Nei’s formula results described above (Figure 7). Comparisons between only Chinese ethnic groups (Supplementary Figures 2A–C) and Asian ethnic groups (Supplementary Figure 3–C) are discussed in Supplementary Text S1.

4 CONCLUSION

In the current study, we genotyped 20 autosomal STR loci in the Han, Uyghur, Kazakh, and Hui ethnic groups of Xinjiang and calculated the forensic parameters. The Goldeneye® 20A panel appeared suitable for forensic investigations such as personal identification and paternity testing and had a high power of discrimination. The STR loci included in the kit showed no significant departures from HWE and minimal departure from LE for a very small number of pairwise combinations of loci. Genetic characterization showed that the Uyghur and Kazakh ethnic groups were closely related to other Turkic-speaking ethnic groups, while the Han and Hui populations showed their associations with other Sinitic language-speaking populations. Interestingly, the Kazakh population showed an affinity with the Mongols which suggested an ancient divergence between Kazakh and Mongols when Mongols originally migrated to present-day Xinjiang (Adnan et al., 2018a; Zhan et al., 2018; He et al., 2019).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

The study was approved by the ethical review board of the China Medical University (2019-21-P), Shenyang Liaoning Province, People’s Republic of China. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AA, and CW, developed the idea. AA wrote the manuscript, AA, HS and JX conducted the experiment, HS, JX, SH, NF, CW, and AA, analyzed the results. AA, SH, and CW modified the manuscript. All authors reviewed the manuscript.

FUNDING

This study was financially supported by China Medical University’s research funds for postdoctoral research grant no. 110/1210619014.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.760760/full#supplementary-material

REFERENCES

Adnan, A., Rakha, A., Kasim, K., Noor, A., Nazir, S., Hadi, S., et al. (2018a). Genetic Characterization of Y-Chromosomal STRs in Hazara Ethnic Group of Pakistan and Confirmation of DYS448 Null Allele. Int. J. Leg. Med. 133, 789–793. doi:10.1007/s00414-018-1962-x
Adnan, A., Rakha, A., Luo, O., and Kayser, M. (2018b). Mutation Analysis at 17 Y-STR Loci (Yfiler) in Father-Son Pairs of Male Pedigrees from Pakistan. Forensic Sci. Int. Genet. 36, e17–e18. doi:10.1016/j.fsigen.2018.07.001
Adnan, A., Rakha, A., Nazir, S., Rehman, Z., Lu, J., and Xuan, J. F. (2020). Genetic Characterization of 15 Autosomal STRs in the interior Sindhi Population of Pakistan and Their Phylogenetic Relationship with Other Populations. Int. J. Immunogenet. 47, 149–157. doi:10.1111/iji.12458
Adnan, A., Rakha, A., Noor, A., van Oven, M., Ralf, A., and Kayser, M. (2017). Population Data of 17 Y-STRs (Yfiler) from Punjabis and Kashmiris of Pakistan. Int. J. Leg. Med. 132, 137–138. doi:10.1007/s00414-017-1611-9

Adnan, A., Zhan, X., Kasim, K., Rakha, A., and Xin, X. J. (2018c). Population Data and Phylogenetic Structure of Han Population from Jiangou Province of China on GlobalFiler STR Loci. Int. J. Leg. Med. 132, 1301–1304. doi:10.1007/s00414-018-1815-7
Aguilar-Velázquez, J. A., Martínez-Cortés, G., Indlán-Sánchez, A., Romero-Rentería, O., Díaz-Navarro, X. X., and Rangel-Villalobos, H. (2016). Population Data of 23 STR Loci (PowerPlex Fusion System) in Mexican Mestizos from the West Region. Int. J. Leg. Med. 130, 1489–1491. doi:10.1007/s00414-016-1361-0
Almeida, C., Ribeiro, T., Oliveira, A. R., Porto, M. J., Costa Santos, J., Dias, D., et al. (2015). Population Data of the GlobalFiler Express Loci in South Portuguese Population. Forensic Sci. Int. Genet. 19, 39–41. doi:10.1016/j.fsigen.2015.06.001
Bai, H., Guo, X., Narisu, N., Lan, T., Wu, Q., Xing, Y., et al. (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. doi:10.1038/s41588-018-0250-5
Millward, J. A. (2007). Eurasian Crossroads: A History of Xinjiang. New York, N.Y.: Columbia University Press. Available at: http://rbdigital.onelickdigital.com (Accessed November 8, 2017).

Moysés, C. S., Tsutsumida, W. M., Raimann, P. E., da Motta, C. H. A. S., Nogueira, T. L. S., dos Santos, O. C. L., et al. (2017). Population Data of the 21 Autosomal STRs Included in the GlobalFiler Kits in Population Samples from Five Brazilian Regions. Forensic Sci. Int. Genet. 26, e28–e30. doi:10.1016/j.fsgen.2016.10.017

Nakamura, Y., Samejima, M., Minaguchi, K., and Nambiar, P. (2016). Population Genetics of Identifier System in Malaysia. Bull. Tokyo Dent. Coll. 57, 233–239. doi:10.2209/tcdpcpublication.2016-1400

Nasibov, E., Bulbul, O., Jabrilyali, G., Zorlu, T., Shahzad, M. S., Cengiz, S., et al. (2013). Allele Frequencies of 15 STR Loci in Azerbaijani Population. Forensic Sci. Int. Genet. 7, e99–e100. doi:10.1016/j.fsgen.2013.03.004

Ng, J., Oldt, R. F., McCulloch, K. L., Weise, J. A., Viray, J., Budowle, B., et al. (2016). Native American Population Data Based on the Globalfiler Autosomal STR Loci. Forensic Sci. Int. Genet. 24, e12–e13. doi:10.1016/j.fsgen.2016.06.014

Ossowski, A., Diepenbrock, M., Szargut, M., Zielinska, G., Jdrezeczyk, M., Berent, J., et al. (2017). Population Analysis and Forensic Evaluation of 21 Autosomal Loci Included in GlobalFiler PCR Kit in Poland. Forensic Sci. Int. Genet. 29, e38–e39. doi:10.1016/j.fsgen.2017.05.003

Ovodov, N. D., Crockford, S. J., Kuzmin, Y. V., Higham, T. F. G., Hodgins, G. W. L., and van der Plicht, J. (2011). A 33,000-Year-Old Incipient Dog from the Altai Mountains of Siberia: Evidence of the Earliest Domestication Disrupted by the Last Glacial Maximum. PLoS ONE 6, e22821. doi:10.1371/journal.pone.0022821

Park, H.-C., Kim, K., Nam, Y., Park, J., Lee, J., Lee, H., et al. (2016). Population Genetic Study for 24 STR Loci and Y Indel (GlobalFiler PCR Amplification Kit and PowerPlex Fusion System) in 1000 Korean Individuals. Leg. Med. 21, 24–28. doi:10.1016/j.legmed.2016.06.003

Park, J.-H., Hong, S.-B., Kim, J.-Y., Chung, Y., Han, S., Jeon, C.-H., et al. (2013). Genetic Variation of 23 Autosomal STR Loci in Korean Population. Forensic Sci. Int. Genet. 7, e76–e77. doi:10.1016/j.fsgen.2012.10.005

Parolin, M. L., Real, L. E., Martinazzo, L. B., and Basso, N. G. (2015). Population Genetic Analyses of the Powerplex Fusion Kit in a Cosmopolitan Sample of Chubut Province (Patagonia Argentina). Forensic Sci. Int. Genet. 19, 221–222. doi:10.1016/j.fsgen.2015.07.020

Perveen, R., Shahid, A. A., Shaﬁque, M., Shahzad, M., and Husnain, T. (2017). Genetic Variations of 15 Autosomal and 17 Y-STR Markers in Sindh Hindhi Population of Pakistan. Int. J. Leg. Med. 131, 1239–1240. doi:10.1007/s00414-017-1544-3

Porras-Hurtado, L., Ruíz, Y., Santos, C., Phillips, C., Carracedo, Á., and Lareu, M. V. (2013). An Overview of STRUCTURE: Applications, Parameter Settings, and Supportive Software. Front. Genet. 4. doi:10.3389/fgene.2013.00098

Frontiers in Genetics | www.frontiersin.org December 2021 | Volume 12 | Article 760760
Adnan et al. Uyghur, Han, Kazakh and Hui Population Typing via Autosomal STRs

Xu, S., and 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. doi:10.1016/j.ajhg.2008.08.001

Yang, M., Ren, Z., Ji, J., Zhou, H., Zhang, H., Dai, J., et al. (2017). Population Genetic Data and Mutations of 22 Autosomal STR Loci in Guizhou Han Population. Forensic Sci. Int. Genet. 29, e29–e30. doi:10.1016/j.fsigen.2017.03.008

Yoo, S. Y., Cho, N. S., Park, M. J., Seong, K. M., Hwang, J. H., Song, S. B., et al. (2011). A Large Population Genetic Study of 15 Autosomal Short Tandem Repeat Loci for Establishment of Korean DNA Profile Database. Mol. Cell 32, 15–19. doi:10.1007/s10059-011-2288-4

Yuan, L., Liu, H., Liao, Q., Xu, X., Chen, W., and Hao, S. (2016). Genetics Analysis of 38 STR Loci in Uyghur Population from Southern Xinjiang of China. Int. J. Leg. Med. 130, 687–688. doi:10.1007/s00414-015-1270-7

Zhan, X., Adnan, A., Zhou, Y., Khan, A., Kasim, K., and McNevin, D. (2018). Forensic Characterization of 15 Autosomal STRs in Four Populations from Xinjiang, China, and Genetic Relationships with Neighboring Populations. Sci. Rep. 8. doi:10.1038/s41598-018-22975-6

Zhang, H., Xia, M., Qi, L., Dong, L., Song, S., Ma, T., et al. (2016a). Forensic and Population Genetic Analysis of Xinjiang Uyghur Population on 21 Short Tandem Repeat Loci of 6-dye GlobalFiler PCR Amplification Kit. Forensic Sci. Int. Genet. 22, 22–24. doi:10.1016/j.fsigen.2016.01.005

Zhang, H., Yang, S., Guo, W., Ren, B., Pu, L., Ma, T., et al. (2016b). Population Genetic Analysis of the GlobalFiler STR Loci in 748 Individuals from the Kazakh Population of Xinjiang in Northwest China. Int. J. Leg. Med. 130, 1187–1189. doi:10.1007/s00414-016-1319-2

Zhang, L. (2015a). Population Data for 15 Autosomal STR Loci in the Bouyei Ethnic Minority from Guizhou Province, Southwest China. Forensic Sci. Int. Genet. 17, 108–109. doi:10.1016/j.fsigen.2015.04.006

Zhang, L. (2015b). Population Data for 15 Autosomal STR Loci in the Dong Ethnic Minority from Guizhou Province, Southwest China. Forensic Sci. Int. Genet. 16, 237–238. doi:10.1016/j.fsigen.2015.02.005

Zhang, L., Zhao, Y., Guo, F., Liu, Y., and Wang, B. (2015). Population Data for 15 Autosomal STR Loci in the Miao Ethnic Minority from Guizhou Province, Southwest China. Forensic Sci. Int. Genet. 16, e3–e4. doi:10.1016/j.fsigen.2014.11.002

Zhao, J., WurigemuleSun, J., Sun, J., Xia, Z., He, G., Yang, X., et al. (2020). Genetic Substructure and Admixture of Mongolians and Kazakhs Inferred from Genome-wide Array Genotyping. Ann. Hum. Biol. 47, 620–628. doi:10.1080/03014460.2020.1837952

Zhu, B.-f., Yang, G., Shen, C.-m., Qin, H.-x., Fan, S.-i., Deng, L.-b., et al. (2010). Population Genetic Analysis of 15 Autosomal STR Loci in the Russian Population of Northeastern Inner-Mongolia, China. Mol. Biol. Rep. 37, 3889–3895. doi:10.1007/s11033-010-0045-y

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor and the reviewer MS declared a past co-authorship with the authors AD, CCW, DCDM at the time of review.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Adnan, Anwar, Simayijiang, Farrukh, Hadi, Wang and Xuan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.