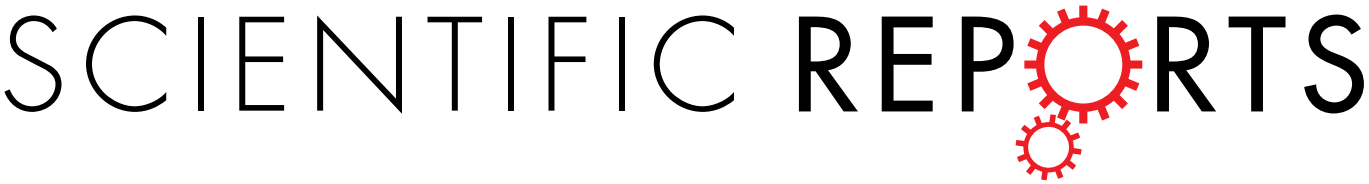

OPEN

Autosomal DIPs for population genetic structure and differentiation analyses of Chinese Xinjiang Kyrgyz ethnic group

Received: 27 October 2017
Accepted: 4 July 2018
Published online: 23 July 2018

Yuxin Guo[1,2,3], Chong Chen[1,2,3], Xiaoye Jin[1,2,3], Wei Cui[1,2,3], Yuanyuan Wei[1,2], Hongdan Wang[4], Tingting Kong[1,2], Yuling Mu[3] & Bofeng Zhu[1,2,3]

In recent years, deletion and insertion polymorphisms (DIPs) were treated as a novel complementary tool with huge potential for forensic applications. In this study, we utilized 30 DIP loci to make a comprehensive research of allele frequency distribution and compute forensic parameters to evaluate the efficiency of forensic applications in the 295 unrelated healthy individuals of Kyrgyz group, and in addition, infer the genetic relationships between Kyrgyz group and 24 other previously studied groups. No significant departures from Hardy-Weinberg equilibrium and linkage disequilibrium were observed at these 30 DIP loci. The combined power of discrimination and the combined probability of exclusion for all 30 DIP loci in Kyrgyz group were 0.9999999999989 and 0.9939, respectively. Furthermore, the results of the interpopulation differentiations, phylogenetic reconstruction, population genetic structure and principal component analyses suggested that Kyrgyz group had relatively close genetic relationships with Kazakh and Uygur groups. However, it was also important to stress that 15 loci were selected out from these 30 DIP loci using the method of selecting ancestry markers, which could be utilized for further ancestry inference study relatively.

The majority of human genome sequence variation could be attributable to nucleotide substitution polymorphisms, with the rest attributable to deletion and insertion polymorphisms (DIPs)[1]. DIPs could in turn be split into those with multiple alleles (multiallelic) and with only two alleles (diallelic)[1]. Nearly all of the multiallelic DIPs were based on tandem repeats, mostly short tandem repeats (STRs)[1], however, the 30 DIP loci chosen in this study were diallelic.

Further, it was particularly noteworthy that DIPs possessed the desirable properties of both STRs and single nucleotide polymorphisms (SNPs), which could be summarized as followings: (i) length polymorphisms allowing them amenable to analyze through simple capillary electrophoresis in common forensic DNA laboratories[2]; (ii) an abundance of distributions in human genome with the density ranking only second to that of SNPs[3]; (iii) small amplicon size improving the probabilities of successfully analyses for highly degraded DNA[4]; (iv) lower rates of mutations, which made them more stable than STRs[2]; (v) PCR amplification without the generation of stutter peaks, making the allelic genotyping results more concise and precise[5]; (vi) the marked differences of allele frequencies in some loci between diverse populations from geographically separated regions, therefore they had potential to be applied in biogeographic ancestry analyses[6,7].

The Kyrgyz ethnic minority with the population totaling over 0.18 million belongs to the 56 ethnic groups officially published by the People's Republic of China[8], which are mainly found in the southwestern part of the Xinjiang Uygur Autonomous Region, China[8]. Now we use Qiagen Investigator DIPplex reagent (Qiagen, Hilden, Germany), a commercial kit, to analyze 30 DIP loci distributed on 19 pairs of chromosomes and in addition,

[1]Key Laboratory of Shaanxi Province for Craniofacial Precision Medicine Research, College of Stomatology, Xi'an Jiaotong University, Xi'an, 710004, P. R. China. [2]Clinical Research Center of Shaanxi Province for Dental and Maxillofacial Diseases, College of Stomatology, Xi'an Jiaotong University, Xi'an, 710004, P. R. China. [3]College of Medicine & Forensics, Xi'an Jiaotong University Health Science Center, Xi'an, 710061, P. R. China. [4]Medical Genetics Institute of Henan Province, Henan Provincial People's Hospital, Zhengzhou University People's Hospital, Zhengzhou, 450003, P. R. China. Correspondence and requests for materials should be addressed to B.Z. (email: zhubofeng7372@126.com)

Loci	DIP+	DIP-	PD	PIC	PE	Ho	He
HLD40	0.5610	0.4390	0.6451	0.3712	0.1336	0.4305	0.4926
HLD131	0.4542	0.5458	0.6406	0.3729	0.1505	0.4542	0.4958
HLD56	0.5831	0.4170	0.6356	0.3680	0.1407	0.4407	0.4862
HLD136	0.4627	0.5373	0.6354	0.3736	0.1635	0.4712	0.4972
HLD77	0.5017	0.4983	0.6338	0.3750	0.1717	0.4814	0.5000
HLD6	0.5492	0.4509	0.6259	0.3726	0.1773	0.4881	0.4952
HLD58	0.3932	0.6068	0.6243	0.3633	0.1455	0.4475	0.4772
HLD67	0.6153	0.3848	0.6217	0.3614	0.1431	0.4441	0.4734
HLD93	0.5542	0.4458	0.6200	0.3720	0.1860	0.4983	0.4941
HLD101	0.5068	0.4932	0.6188	0.3750	0.1981	0.5119	0.4999
HLD133	0.4441	0.5559	0.6179	0.3719	0.1890	0.5017	0.4937
HLD88	0.5542	0.4458	0.6165	0.3720	0.1920	0.5051	0.4941
HLD92	0.5559	0.4441	0.6144	0.3719	0.1950	0.5085	0.4937
HLD124	0.6525	0.3475	0.6128	0.3506	0.1160	0.4034	0.4535
HLD70	0.6186	0.3814	0.6100	0.3605	0.1635	0.4712	0.4718
HLD97	0.3729	0.6271	0.6099	0.3583	0.1556	0.4610	0.4677
HLD128	0.3797	0.6203	0.6078	0.3601	0.1662	0.4746	0.4710
HLD125	0.4983	0.5017	0.6053	0.3750	0.2206	0.5356	0.5000
HLD48	0.5051	0.4949	0.6053	0.3750	0.2206	0.5356	0.4999
HLD122	0.3373	0.6627	0.5975	0.3471	0.1383	0.4373	0.4470
HLD114	0.4102	0.5898	0.5971	0.3668	0.2075	0.5220	0.4839
HLD45	0.6610	0.3390	0.5953	0.3477	0.1455	0.4475	0.4481
HLD83	0.3492	0.6509	0.5715	0.3512	0.2012	0.5153	0.4545
HLD84	0.7170	0.2831	0.5677	0.3235	0.1079	0.3898	0.4059
HLD64	0.7339	0.2661	0.5535	0.3143	0.1039	0.3831	0.3906
HLD118	0.7356	0.2644	0.5524	0.3133	0.1020	0.3797	0.3890
HLD99	0.7356	0.2644	0.5490	0.3133	0.1139	0.4000	0.3890
HLD81	0.7424	0.2576	0.5425	0.3094	0.1139	0.4000	0.3825
HLD111	0.2593	0.7407	0.5346	0.3104	0.1383	0.4373	0.3841
HLD39	0.2085	0.7915	0.4967	0.2756	0.0761	0.3288	0.3300

Figure 1. Plots of allele frequencies and forensic parameters were mapped on account of 30 DIP loci in the Chinese Xinjiang Kyrgyz group.

this kit was put into use in the previous population studies which were already published[9–11]. We gathered the bloodstain samples of Kyrgyz group in Xinjiang Uygur Autonomous Region and used the kit mentioned above to obtain population data to acquire more information about the Kyrgyz ethnic minority's genetic background.

Results and Discussion

The analyses of allelic frequency distributions and forensic parameters. The systematically experimental operations and analyses of the samples had been conducted under the laboratory stringent criteria before the data of Kyrgyz group obtained. There were no significant departures from Hardy-Weinberg equilibrium (HWE) in the 30 DIPs after applying a Bonferroni correction ($p = 0.05/30 = 0.0017$). Allele frequencies and forensic efficiency parameters of 30 DIPs in Kyrgyz group were depicted in Fig. 1. The expected heterozygosity (He) values ranged from 0.3300 (HLD39) to 0.5000 (HLD77 and HLD125) and the observed heterozygosity (Ho) values varied from 0.3288 (HLD39) to 0.5356 (HLD48 and HLD125). The values of polymorphic information content (PIC) were in the range of 0.2756 to 0.3750 with a mean value of 0.3524. Additionally, the maximum value of power of exclusion (PE) was 0.2206 at HLD48 and HLD125 loci, whereas the minimum was 0.0761 at HLD39 locus. The combined probability of exclusion (CPE) for all 30 DIP loci in Kyrgyz group was 0.9939. However, the CPE value was relatively low (compared with which of STRs[12]) implying that the panel of 30 DIP loci could be a complementary tool for STR typing system in forensic paternity cases. We also detected the power of discrimination (PD) ranging from 0.4967 (HLD39) to 0.6451 (HLD40), and combined power of discrimination (CPD) reached 0.9999999999989, which was able to meet the satisfactory levels for the individual identification of forensic demands[13]. Among these forensic parameters, it was significantly pronounced that the lowest values of He, Ho, PD, PE, PIC were obtained at HLD39 locus, indicating that HLD39 locus showed relative low forensic efficiency in the studied Kyrgyz ethnic group.

In addition, the values of minor allele frequency (MAF) of 0.0700 to 0.2000 were found at 12 loci (Supplementary Table 1), including HLD39, HLD48, HLD58, HLD81, HLD83, HLD84, HLD99, HLD111, HLD114, HLD122, HLD125 and HLD128 loci. In this study, the MAF values of some loci were generally low, indicating that the panel of 30 DIP loci might have great potential to detect population structure and analyze population genetic relationships[14]. Therefore, to confirm our hypothesis, we did the following analyses of 25 populations based on these 30 DIP loci to explore the population origin and genetic structure of Xinjiang Kyrgyz.

Linkage disequilibrium analyses. Linkage disequilibrium (LD) tests among these 30 DIP loci in Kyrgyz group were performed using the SNPAnalyzer program. As shown in the Supplementary Figure 1, the pairwise LD analyses indicated that no significant LD with the coverage of thick black curve existed in the plot, showing that these 30 DIP loci were independent with each other in the studied Kyrgyz ethnic group.

Interpopulation differentiations. To explore the hereditary similarities and differences, the studied Kyrgyz group was compared with previously published groups at these 30 DIP loci utilizing the analysis of molecular variance (AMOVA) method on the basis of Arlequin software version 3.1. The locus-by-locus p values were shown in Supplementary Table 2 and the number of loci with significant differences ($p < 0.05$) between Kyrgyz and the 24 reference populations were presented intuitionally in bar diagram format combined with the result of structure analysis ($K = 4$) in Fig. 2. Statistically significant differences were detected between the studied Kyrgyz

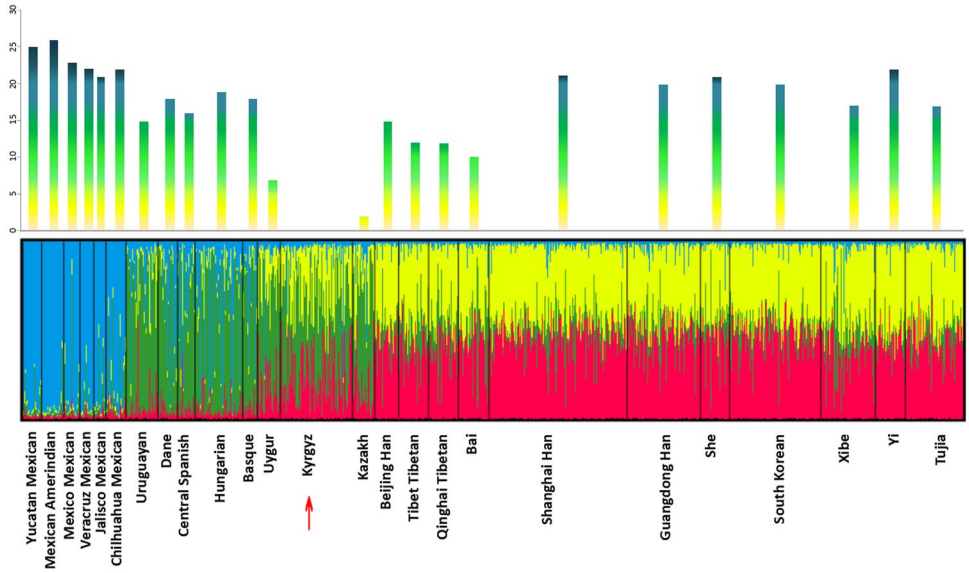

Figure 2. Clustering structure for the full-loci dataset assuming $K = 4$ of the 25 groups combined with the bar diagram representing various numbers of pairwise p-value with significant differences ($p < 0.05$).

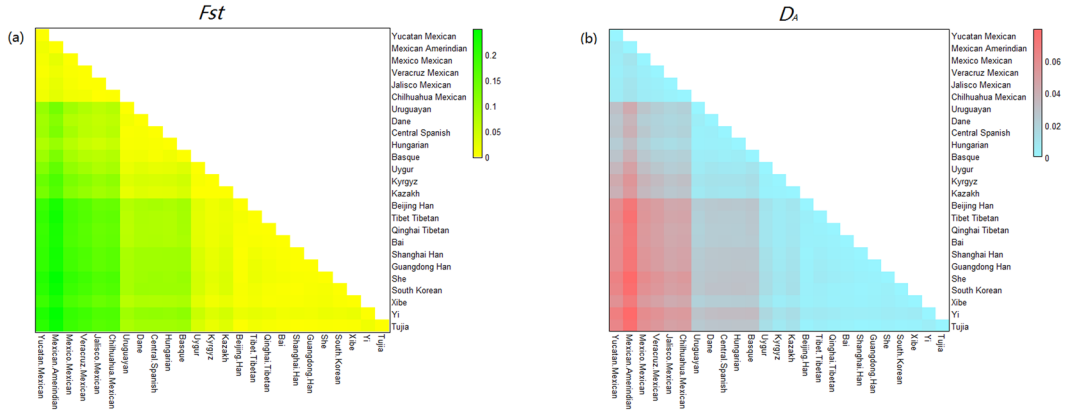

Figure 3. (**a**) Heat map of pairwise *Fst* values of 30 DIP loci in Xinjiang Kyrgyz and 24 previously studied populations based on *R* software. (**b**) Heat map of pairwise D_A values of 30 DIP loci among Xinjiang Kyrgyz and 24 previously published populations conducted with *R* software.

group and Kazakh group[11] at two loci; Uygur group[11] at seven loci; Bai group[15] at ten loci; Tibet Tibetan and Qinghai Tibetan group[16] at 12 loci; Beijing Han[11] and Uruguayan group[17] at 15 loci; Central Spanish group[18] at 16 loci; Tujia[19] and Xibe group[20] at 17 loci; Basque[18] and Dane group[21] at 18 loci; Hungarian group[22] at 19 loci; Guangdong Han[23] and South Korean group[24] at 20 loci; Shanghai Han[25] and She group[25] at 21 loci; Yi group[26] at 22 loci; and six Mexican groups[27] at 21–26 loci, respectively. According to the diagram, the Kyrgyz group had the lowest genetic divergence with Kazakh group (significant differences found at two loci) in contrast with Mexican Amerindian group (significant differences found at 26 loci). Furthermore, some loci with high ethnic diversities could be observed: two loci (HLD81 and HLD111) all showed significant differences at 21 compared groups, on the contrary, HLD101 and HLD88 at four and six respectively. It was suggested that the abilities of some DIP loci to distinguish ethnic groups were at different levels[3]. Note that, studies of more DIP loci in more ethnic populations should be required for the different application purposes in forensic science.

In addition, we conducted two heat maps using *R* statistical software. Based on 30 DIP loci, one heat map of pairwise fixation index (*Fst*) values (Supplementary Table 3) calculated by GENEPOP program was labeled on Fig. 3a, revealing the genetic differentiations among the studied Kyrgyz group and 24 reference populations. *Fst* is directly related to the variance in allele frequency among populations. The larger *Fst* value is, the higher genetic divergence between pairwise populations is, and vice versa[28]. As presented in Fig. 3a, the deeper green color stood for the larger *Fst* value, which meant the more differentiation existed between pairwise populations; conversely, the deeper yellow color meant the smaller *Fst* value as well as the less differentiation[28]. We could also detect intuitively that the 25 studied populations could be separated into four clusters based on the depth of color:

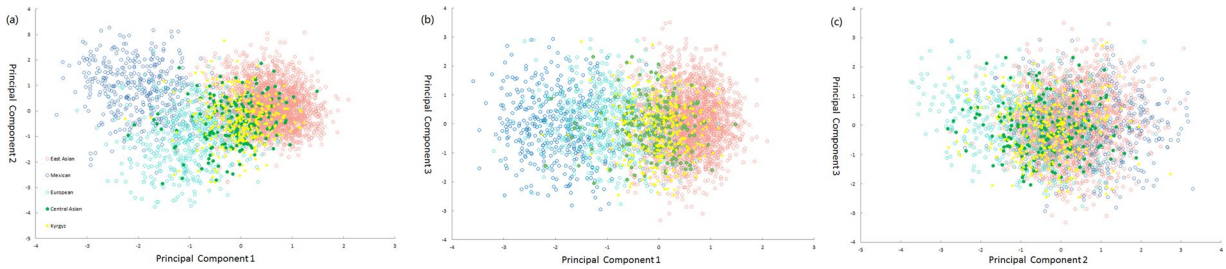

Figure 4. The PCA plots were analyzed at individual level using 30 DIP loci. (**a**) based on PC1and PC2 (**b**) based on PC1 and PC3 (**c**) based on PC2 and PC3. The red dots represented for East Asians, deep blue dots for Mexicans, light blue dots for Europeans, green dots for Central Asians, and yellow dots for Kyrgyzs.

the Mexican groups, European groups and Uruguayan, Central Asian groups, and East Asian groups. Focusing on the Central Asian groups, we could come to a conclusion that the studied Kyrgyz had lower *Fst* values with Kazakh and Uygur groups. For further detail, the studied Kyrgyz had deeper yellow color with Kazakh group rather than Uygur group, which indicated Kyrgyz group might have a closer genetic relationship or more similar origin with Kazakh group.

For further elucidation of the Chinese population affiliations, another heat map exhibited in Fig. 3b plotted based on pairwise D_A values (Supplementary Table 4) which were carried out with DISPAN program, showing the various genetic distances among the studied Kyrgyz and 24 reference populations. The deeper red color represented the greater D_A value meant the bigger genetic distance, while the deeper blue color meant the smaller D_A value along with the closer genetic distance. In addition, each pairwise population with a closer genetic distance also had a smaller genetic divergence. In this study, small D_A value with deep blue color were also found in four clusters: the Mexican groups, European groups and Uruguayan, Central Asian groups, and East Asian groups, which was consistent with the result of *Fst* heat map. On the purpose of more direct analyses, a bar chart of both *Fst* and D_A values between the studied Kyrgyz group and 24 reference populations were shown in the Supplementary Figure 2, respectively, displaying the high consistency of trend between both kinds of values. It was evident that the studied Kyrgyz group had the shortest genetic distance with the Central Asian groups (Kazakh and Uygur groups), which meant that these three groups with small genetic divergence mentioned above might have similar consanguineous relationships to some extent.

Principal component analysis. The genetic relationships between Kyrgyz group and other 24 populations were presented by three plots of principal component analysis (PCA) utilizing the SPSS 18.0 software (SPSS, Chicago, IL, USA). As shown in Fig. 4a, 25 populations were divided into four colored clusters based on PC1 (9.844%) and PC2 (4.373%), including Central Asian groups (green) without Kyrgyz group, East Asian groups (pink), six Mexican (deep blue) and European groups (light blue). Then, Kyrgyz group was represented by yellow points, with one yellow dot standing for an individual. Nevertheless, Fig. 4c based on PC2 (4.373%) and PC3 (3.698%) was in a blended cluster with small capacity to discriminate each continent apart contrasting with Fig. 4a and in addition, Fig. 4b on the basis of PC1 (9.844%) and PC3 (3.698%) had a relatively limited discrimination between Fig. 4a,c.

As presented in Fig. 4a, all individuals from 25 populations were partitioned into four main regions keeping in line with their intercontinental distributions roughly, and individuals from Kyrgyz, Kazakh and Uygur groups were scattered between East Asians and Europeans as expected, conforming to the previous studies and ethnic migration records[29,30]. Since Western Han Dynasty to the middle of the Qing Dynasty, mainly from the Yenisai River to the Tianshan Mountains and Central Asia, Kyrgyz group experienced five westward migrations which were basically facilitated by warfare[29]. The studied Kyrgyz group which inhabits the southwestern region of Xinjiang broadly assimilated Western Regions culture after the long term of mixed dwelling with the Uygurs, Kazakhs, Hans and Mongolians etc[29]. In contrast, the ancestry of Xinjiang Xibe is different from Kyrgyz group. Xibe traditionally resided in northeast China and immigrated to Xinjiang during the middle of the eighteenth century. Thus, Xinjiang Xibe group had the same pattern with other East Asian groups as shown in structure analysis of Fig. 2, displaying the same cluster pattern as previously reported[31]. That was the reason why the Xinjiang Xibe group came from the same region as Kyrgyz group was treated as a member of East Asian groups in Fig 4. The above result also showed the close genetic relationships between the Kyrgyz, Kazakh and Uygur groups, and implied that Kyrgyz group, in this study, might play an important role in culture exchange and gene flow between East Asians and Europeans[8].

Multidimentional scaling analysis. For further investigation of genetic correlations among 25 populations, multidimentional scaling (MDS) analysis was performed using SPSS 18.0 software (SPSS, Chicago, IL, USA) and provided a two-dimensional representation of genetic relationships based on pairwise *Fst* values calculated by GENEPOP program. As shown in Fig. 5, each dot in the two dimensional space indicated one population and it was given a color according to language family that the population belonged to. The various distances between different dots showed different genetic relationships among the populations. In detail, the closer the two dots were, the closer the genetic relationships they had. In the light of various distances between dots, 25 populations mentioned above were divided into four clusters roughly: the East Asian groups, the Mexican groups,

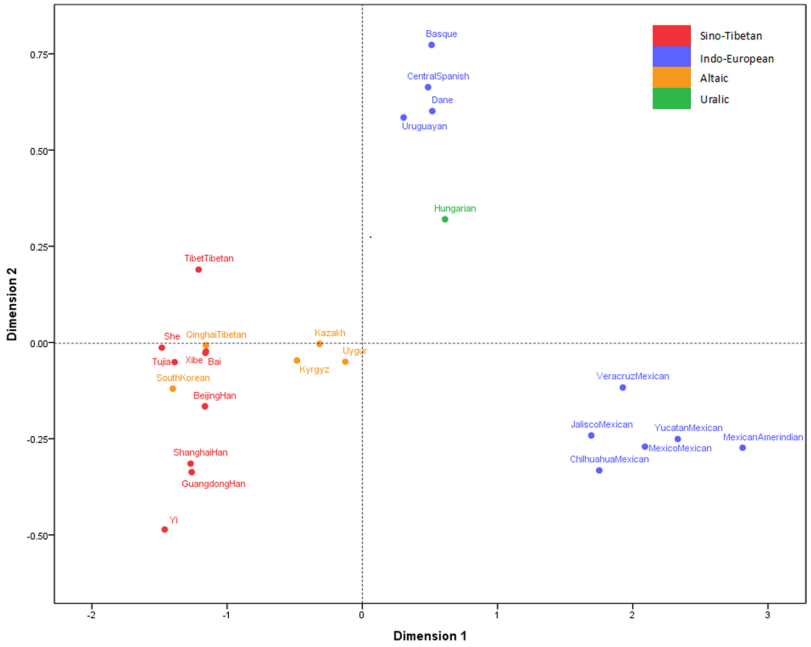

Figure 5. MDS analysis of Xinjiang Kyrgyz group and other 24 reference populations were conducted by SPSS18.0 based on pairwise *Fst* values.

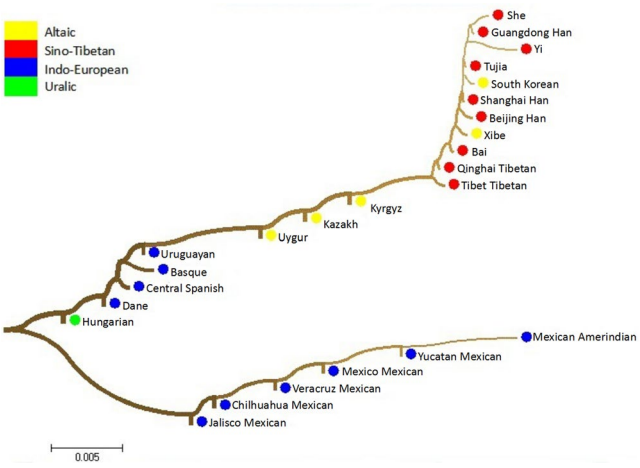

Figure 6. Phylogenetic tree constructed by the neighbour joining method based on the D_A distances among the 25 populations.

European groups and Uruguayan, and Central Asian groups, which was in concordance with their geographic distributions.

Phylogenetic analysis. In order to estimate the studied population affiliations, based on the D_A values, phylogenetic reconstruction tree encompassing two main branches was performed to verify the genetic relationships between the Xinjiang Kyrgyz and 24 reference populations, which was reconstructed utilizing neighbor joining method by MAGA software v5.0. As shown in Fig. 6, two main branches could be clearly identified in the phylogenetic tree. The upper branch was composed of three clusters: eleven East Asian groups on the upper-right side including South Korean and most Chinese populations such as Tibetan populations from two different regions, Han populations from three different regions, She, Yi, Tujia, Xibe and Bai groups; Uruguayan group and four European groups on the upper-left side; and Central Asian groups (Kazakh, Kyrgyz and Uygur groups) in the middle of the upper branch. The lower branch merely consisted of six Mexican populations. The clustering results of dendrogram focusing on the Xinjiang Kyrgyz, Kazakh and Uygur populations roughly congruent with the former analyses such as the results of PCA and MDS analysis.

In long-term development of Chinese history, the Kazakh group was formed as a combination of the Turkic, Wusun, Khitan and Mongolian people, and Uygur group was stemmed from a branch of Turkic people[11], while

Kyrgyz group partially mixed the genetic component of Mongol, Khitan, Turkic, Uygur and Han groups since the ancient time[29,32]. Furthermore, Kyrgyz, Kazakh and Uygur groups almost possessed the same religious belief and the close geographic distance, more likely leading to the cultural exchange among these three groups[29].

Clustering analyses. To assess the population stratification and calculate the proportion of different ancestry components in various populations, the structure analyses based on the genotyping data of the studied Kyrgyz group and 24 reference populations were conducted with Structure 2.3.4 software which could infer individual genetic ancestry coefficients by controlling the values of *K* that represented the number of hypothetical ancestral populations[33]. As shown in Supplementary Figure 3, 25 populations were separated by black lines and each single vertical line represented one individual was partitioned into several colored segments on behalf of the individual's estimated membership fractions[34]. At $K=2$, the East Asian groups were distinguished from both European groups (including Uruguayan, Dane, Central Spanish, Basque and Hungarian groups) and Mexican groups (including Yucatan Mexican, Mexican Amerindian, Mexico Mexican, Veracruz Mexican, Jalisco Mexican and Chilhuahua Mexican groups), with the constitution of entirely red components. At the same time, the Central Asian groups, including Kyrgyz, Kazakh and Uygur groups, shared obviously mixed memberships in red and green color. Five European groups and six Mexican groups were almost filled with green components; therefore, European and Mexican groups could not separate from each other at $K=2$. Whereas, the Central Asian groups were separated from other populations evidently with the combination of red, green and yellow components in different proportions at $K=4$ (Fig. 2), which was verified as the most suitable *K* value relying on the output posterior probability results[35]. As a result, genetic clusters were roughly in accordance with collections of geographically similar populations[34].

Based on the analyses mentioned above, we have a sufficient reason to insist that some loci with great genetic divergence were existed in these 30 DIP loci, which were valid for detecting population structure and distinguishing Kyrgyz ancestry information from other populations distributed in different administrative divisions. However, abilities of these 30 DIP loci appeared to be diverse at the power of describing population clusters. In order to pinpoint the loci that more contributed to population discrimination, 15 loci among these 30 DIP loci, including HLD39, HLD45, HLD48, HLD56, HLD58, HLD70, HLD64, HLD81, HLD83, HLD111, HLD114, HLD118, HLD122, HLD125 and HLD128, were selected out based on population-specific allele frequencies (δ values >0.29)[36], using the method of selecting ancestry information markers (AIMs) that discrepant values of average insertion allele frequencies among the clusters (four clusters as detected above: the Mexican groups, European groups and Uruguayan, Central Asian groups, and East Asian groups) could be over 0.29[37]. After that, we again performed structure analyses of the studied Kyrgyz group and 24 reference populations with the result ($K=4$) showing in the Supplementary Figure 4b, significantly contrasting with Fig. 4a which was performed by the rest eliminated 15 loci. We could draw the same conclusion in the structure Fig. 4b that 25 populations were partitioned into four clusters roughly, three Central Asian groups and compared with 11 East Asian groups, five European populations and six Mexican populations, the studied Kyrgyz group had more intimate membership with Kazakh and Uygur groups[38]. However, in the Fig. 4a, the population stratification could hardly be detected, which indicated the ability of ancestry inference of the rest eliminated 15 loci was relatively insufficient.

In brief, to explore the genetic background and genetic structure of the studied Kyrgyz and other populations further, we could choose efficient AIMs with reference to the method mentioned above to acquire more comprehensive and accurate population genetic information and to lay a solid foundation for the ancestry inference study in the future.

Conclusion

In this study, the allele frequencies and statistically forensic parameters of the autosomal 30 DIP loci were obtained for the researches of population genetics and forensic applications. The panel, as a useful forensic tool, was suit for individual identification, but could barely be treated as supplementary markers for STR loci in paternity testing. The results of interpopulation differentiations, genetic distances, principal component, multidimentional scaling, phylogenetic and structure analyses indicated close genetic relationships between Kyrgyz and the two Central Asian groups (Kazakh and Uygur groups). Furthermore, we selected out 15 loci with sufficient capacity of ancestry inference from these 30 DIP loci, which could be implemented in ancestry inference study. For the sake of better understand the origin and genetic evolution of Kyrgyz group, further study should be performed in later research.

Material and Methods

Sample collections and DNA extraction. The bloodstain samples of 295 unrelated healthy individuals were collected from Kyrgyz group residing in Xinjiang Uygur Autonomous Region, China. During the course of collecting samples, we excluded the samples gathered from two individuals who had blood relationships within three generations. All participants concerned to this study provided the written informed consents. The research was in accordance with the human and ethical research principles and approved by the ethics committee of Xi'an Jiaotong University Health Science Center.

PCR amplification and DIP genotyping. On GeneAmp PCR System 9700 thermal cycler (Applied Biosystems, Foster City, CA, USA), a multiple PCR amplification with five-color fluorescence of autosomal 30 DIPs was performed with Investigator DIPplex reagent on the basis of manufacturer's instructions. Genotyping of DIPs was analyzed by capillary electrophoresis on ABI 3500 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA) and processed by GeneMapperv3.2 software (Applied Biosystems, Foster City, CA, USA).

Quality control. The study was conducted following ISFG recommendations on the analysis of the DNA polymorphisms as described by Schneider[39].

Statistical analyses. Forensic statistical parameters of 30 DIP loci in Kyrgyz group, such as the values of HWE, Ho, PE, PD, PIC and allele frequency distributions were calculated by modified Powerstate (version1.2) spreadsheet (Promega, Madison, WI, USA), and He values were calculated based on allele frequencies. Arlequin software (version3.0)[40] was chosen to evaluate locus-by-locus *p* values using AMOVA method. The heat maps were performed by *R* statistical software v3.0.2[41] based on *Fst* and D_A values which were calculated by allele frequencies and raw population data, respectively. The SNPAnalyzer (version2.0 Istech, South Korea)[42] was selected to test LD for all pairwise DIP loci. Three plots of PCA and a plot of MDS analysis were carried out by SPSS 18.0 software (SPSS, Chicago, IL, USA). Population genetic structure analyses were conducted by STRUCTURE v2.2 program[33]. The phylogenetic tree was described using MAGA v6.06 software based on D_A values calculated by DISPAN program[43].

Data availability. The datasets analyzed during the current study are available from the corresponding author upon reasonable request.

References

1. Weber, J. L. *et al.* Human diallelic insertion/deletion polymorphisms. *American Journal of Human Genetics* **71**, 854–862 (2002).
2. Pereira, R. *et al.* A new multiplex for human identification using insertion/deletion polymorphisms. *Electrophoresis* **30**, 3682–3690 (2009).
3. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
4. Fondevila, M. *et al.* Forensic performance of two insertion-deletion marker assays. *International Journal of Legal Medicine* **126**, 725 (2012).
5. Wendt, F. R. *et al.* Massively parallel sequencing of 68 insertion/deletion markers identifies novel microhaplotypes for utility in human identity testing. *Forensic Science International Genetics* **25**, 198 (2016).
6. Yang, N. *et al.* Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. *Human Genetics* **118**, 382 (2005).
7. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
8. Zulfiya, Y. *et al.* Genetic polymorphisms of pharmacogenomic VIP variants in the Kyrgyz population from northwest China. *Gene* **529**, 88–93 (2013).
9. Larue, B. L., Ge, J., King, J. L. & Budowle, B. A validation study of the Qiagen Investigator DIPplex® kit; an INDEL-based assay for human identification. *International Journal of Legal Medicine* **126**, 533–540 (2012).
10. Neuvonen, A. M., Palo, J. U., Hedman, M. & Sajantila, A. Discrimination power of Investigator DIPplex loci in Finnish and Somali populations. *Forensic Science International Genetics* **6**, e99 (2012).
11. Wei, Y. L., Qin, C. J., Dong, H., Jia, J. & Li, C. X. A validation study of a multiplex INDEL assay for forensic use in four Chinese populations. *Forensic Science International Genetics* **9**, e22 (2014).
12. Li, Y. *et al.* Genetic polymorphism of 21 non-CODIS STR loci in Chengdu Han population and its interpopulation analysis between 25 populations in China. *Legal Medicine* **31**, 14–16 (2018).
13. Lai, J. H., Zhang, H. B., Zhu, B. F., Chen, T. & Zheng, H. B. STR polymorphism in Bai minority in Yunnan province. *Journal of Xian Medical University* **23**, 242–245 (2002).
14. De, lC. O. & Raska, P. Population structure at different minor allele frequency levels. *Bmc Proceedings* **8**, 1–5 (2014).
15. Yang, C. H. *et al.* Genetic variation and forensic efficiency of autosomal insertion/deletion polymorphisms in Chinese Bai ethnic group: phylogenetic analysis to other populations. *Oncotarget* **8** (2017).
16. Guo, Y. *et al.* Population Differentiations and Phylogenetic Analysis of Tibet and Qinghai Tibetan Groups Based on 30 InDel Loci. *Dna & Cell Biology* **35**, 787–794 (2016).
17. Saiz, M. *et al.* Allelic frequencies and statistical data from 30 INDEL loci in Uruguayan population. *Forensic Science International Genetics* **9**, e27 (2014).
18. Martín, P. *et al.* Population genetic data of 30 autosomal indels in Central Spain and the Basque Country populations. *Forensic Science International Genetics* **7**, e27 (2013).
19. Shen, C. *et al.* A 30-InDel Assay for Genetic Variation and Population Structure Analysis of Chinese Tujia Group. *Sci Rep* **6**, 36842 (2016).
20. Meng, H. T. *et al.* Genetic polymorphism analyses of 30 InDels in Chinese Xibe ethnic group and its population genetic differentiations with other groups. *Sci Rep* **5**, 8260 (2015).
21. Friis, S. L. *et al.* Typing of 30 insertion/deletions in Danes using the first commercial indel kit—Mentype®; DIPplex. *Forensic Science International Genetics* **6**, e72–e74 (2012).
22. Z, K. *et al.* Genome deletion and insertion polymorphisms (DIPs) in the Hungarian population. *Forensic Science International Genetics* **6**, e125 (2012).
23. Hong, L. *et al.* Genetic Polymorphisms of 30 Indel Loci in Guangdong Han Population. *Journal of Sun Yat-sen University(Medical Sciences)* **34**, 299–304 (2013).
24. Seong, K. M. *et al.* Population genetics of insertion–deletion polymorphisms in South Koreans using Investigator DIPplex kit. *Forensic Science International Genetics* **8**, 80–83 (2014).
25. Wang, Z. *et al.* Population genetics of 30 insertion-deletion polymorphisms in two Chinese populations using Qiagen Investigator® DIPplex kit. *Forensic Science International Genetics* **11**, e12–e14 (2014).
26. Zhang, Y. D. *et al.* Forensic evaluation and population genetic study of 30 insertion/deletion polymorphisms in a Chinese Yi group. *Electrophoresis* **36**, 1196 (2015).
27. Martínez-Cortés, G. *et al.* Forensic parameters of the Investigator DIPplex kit (Qiagen) in six Mexican populations. *International Journal of Legal Medicine* **130**, 683–685 (2016).
28. Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nature Reviews Genetics* **10**, 639–650 (2009).
29. Yang, Y. X. Westward Migration and Its Influence on the Formation of Kirgiz. *Journal of Beifang University of Nationalities* **123**, 30–33 (2015).
30. Lou, H. *et al.* Copy number variations and genetic admixtures in three Xinjiang ethnic minority groups. *European Journal of Human Genetics* **23**, 536–542 (2015).
31. Meng, H. T. *et al.* Chinese Xibe population genetic composition according to linkage groups of X-chromosomal STRs: population genetic variability and interpopulation comparisons. *Annals of Human Biology*, **1** (2017).
32. Peng, M. S. *et al.* Mitochondrial genomes uncover the maternal history of the Pamir populations. *European Journal of Human Genetics Ejhg* **26** (2017).
33. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
34. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science (New York, N.Y.)* **298**, 2381–2385 (2002).

35. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611–2620 (2005).
36. Shriver, M. D. *et al*. Ethnic-affiliation estimation by use of population-specific DNA markers. *American Journal of Human Genetics* **60**, 957 (1997).
37. Halder, I., Shriver, M., Thomas, M., Fernandez, J. R. & Frudakis, T. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. *Human Mutation* **29**, 648 (2008).
38. Porrashurtado, L. *et al*. An overview of STRUCTURE: applications, parameter settings, and supporting software. *Frontiers in Genetics* **4**, 98 (2013).
39. Schneider, P. M. Scientific standards for studies in forensic genetics. *Forensic Science International* **165**, 238 (2007).
40. Excoffier, L., Laval, G. & Schneider, S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* **1**, 47–50 (2005).
41. Coreteam, R. R. A language and environment for statistical computing. *Computing* **1**, 12–21 (2015).
42. Jinho, Y., Youngbok, L., Yujung, K., Young, R. S. & Yangseok, K. SNPAnalyzer 2.0: A web-based integrated workbench for linkage disequilibrium analysis and association analysis. *Bmc Bioinformatics* **9**, 290 (2008).
43. Ota, T. *DISPAN: Genetic Distance and Phylogenetic Analysis*. (Pennsylvania State Univ. 1993).

Acknowledgements

This study was supported by the National Natural Science Foundation of China (NSFC, No. 81525015, 81772031).

Author Contributions

Y.G. and B.Z. performed the data acquisition and wrote the main manuscript text, B.Z., C.C. and X.J. designed the research, W.C., Y.W. and H.W. did the data processing and the manuscript modification, T.K. and Y.M. prepared the figures. All authors reviewed the manuscript. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-29010-8.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.