Hereditas

RESEARCH Open Access

Systematic evaluations of forensic effectiveness and genetic structures of two ethnic groups in Northwest China using a self-developed Multi-InDel panel

Qinglin Liang[1], Qiong Lan[1], Qinglin Liu[1], Xiaolian Wu[1], Lisiteng Luo[1], Chunmei Shen[2*] and Bofeng Zhu[1*]

Abstract

Background The use of compound markers has gained significant interest among forensic practitioners, due to their ability to enhance genetic marker polymorphisms by introducing new alleles. Two or more closely linked insertion/deletion (InDel) markers form a compound marker termed Multi-InDel, which offers the advantages of microhaplotype (MH) and can be genotyped using capillary electrophoresis (CE) platform. A multiplex amplification panel, including 41 Multi-InDel markers and the sex-determination locus Amelogenin, was developed and validated as an effective tool for forensic and population genetics applications.

Methods A total of 245 Kazakh and Kyrgyz samples from China were genotyped based on the 41 Multi-InDel markers to evaluate the forensic efficacy of the panel. In addition, Multi-InDel genotyping data from 28 reference populations were collected, and population genetic analyses were performed to elucidate the genetic backgrounds of Chinese Kazakh and Kyrgyz groups.

Conclusions The Multi-InDel markers demonstrated high genetic polymorphisms in Chinese Kazakh and Kyrgyz ethnic groups, indicating their suitability for forensic applications. For the two ethnic groups, the cumulative power of discrimination (CPD) values were 0.999999999999999999999999835993 and 0.999999999999999999999999717184, respectively, while the cumulative power of exclusion (CPE) values were 0.999998887418153 and 0.999999348634116, respectively. Using this Multi-InDel panel, an average of 98.82% of full sibling (FS) pairs could be distinguished from unrelated individual pairs (likelihood ratio > 1). Regarding population genetics, Chinese Kazakh and Kyrgyz groups were found to exhibit an East Asia-Europe admixed ancestry pattern, while maintaining closer genetic affinities with East Asian populations.

Keywords Forensic genetics, Multi-InDel, Capillary electrophoresis, Individual identification, Kinship analysis, Population genetics

*Correspondence:
Chunmei Shen
cmshen2004@126.com
Bofeng Zhu
zhubofeng7372@126.com

[1]Guangzhou Key Laboratory of Forensic Multi-Omics for Precision Identification, School of Forensic Medicine, Southern Medical University, Guangzhou, Guangdong, China
[2]Department of Clinical Laboratory, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong 510515, China

Background

The analysis of human polymorphic genetic marker is a cornerstone of forensic genetics. Currently, short tandem repeat (STR) is the predominant genetic marker used in forensic DNA identification [1, 2]. STR genotyping based on polymerase chain reaction (PCR) and capillary electrophoresis (CE) will likely remain the gold standard for forensic casework for the foreseeable future [3, 4]. However, the application of STRs has been found to have drawbacks such as a limited number of available loci, relatively high mutation rates, and long PCR amplification fragments. Single nucleotide polymorphisms (SNPs) are characterized by their high genomic density, low mutation rates, and relatively short amplicons compared to STRs, and short amplicons facilitate the genotyping of degraded biological samples [5]. In 2013, professor Kidd introduced the concept of microhaplotypes (MHs) based on SNPs at the 24th World Congress of the International Society for Forensic Genetics. MHs are defined as a single sequencing fragment with at least three haplotypes (alleles) detected [6]. When evaluated as microhaplotype, a short sequence region containing multiple SNPs within an amplicon can exhibit high level of heterozygosity. However, incorporating MHs into CE platform commonly used in forensic laboratory is challenging [7].

Insertion/deletion (InDel) genetic markers, biallelic length polymorphisms resulting from the insertion or deletion of DNA fragments, are abundant in the human genome [8, 9]. Combining the characteristics of STRs and SNPs, InDels can be genotyped on the CE platform, facilitating their implementation in routine forensic laboratories. Their lower mutation rates ensure stable inheritance, which is crucial for biogeographic ancestry inference and paternity testing [10]. Furthermore, the flexible sizes of amplification products also allow for the genotyping of degraded samples [11, 12], making InDels versatile tools in forensic genetics. Nevertheless, the information carried by the InDel as a biallelic genetic marker is limited, and the number of InDels that need to be jointly applied to achieve sufficient system efficacy is large, which also complicates the construction of multiplex amplification system. Given these considerations, researchers have endeavored to investigate the closely linked Multi-InDel genetic markers [13–17], aiming to obtain more genetic information from the same number of markers. Multi-InDel markers represent a broad type of microhaplotype and exhibit the advantages of MHs. Moreover, Multi-InDel can be genotyped on the CE platform, making them compatible with standard forensic laboratory workflow. Considering the above advantages, a multiplex amplification panel was constructed, comprising 41 Multi-InDel markers and sex-determination locus Amelogenin. This panel includes 82 InDel markers, with each Multi-InDel consisting of two closely linked InDels. In addition, the panel has been validated as an effective tool in previous studies [18–20].

According to Chinese seventh national population census (https://www.stats.gov.cn/sj/pcsj/rkpc/7rp/zk/indexch.htm), Chinese Kazakh and Kyrgyz ethnic groups number over 1.56 million and 200,000, respectively, and are recognized as significant ethnic minorities in China. These two groups primarily inhabit northwestern China, which is located at the crossroad of the Eurasian continent and is historically connected to the Silk Road. The Silk Road is a major corridor linking East Asia, Central Asia and Europe, and plays an important role in economic exchange and population migration. As long-term settled groups in the region, Chinese Kazakh and Kyrgyz groups are key to understanding the history of genetic exchange between East and West Eurasia. Recent advances have revealed that Chinese Kazakh and Kyrgyz groups exhibit considerable East-West admixture, providing deeper insights into the complex genetic relationships between Western and East Asian populations [21, 22]. However, forensic research on these two ethnic groups remains limited, particularly in the application of Multi-InDel genetic markers. Therefore, this study utilizes an self-developed panel containing 41 Multi-InDel markers to systematically evaluate its forensic applicability in Chinese Kazakh and Kyrgyz ethnic groups, as well as to explore their genetic structures and backgrounds through population genetic analyses.

Methods and materials

Sample collection

A total of 245 blood samples were collected, including 145 Kazakh and 100 Kyrgyz individuals. The participants self-reported good health, were not related within three generations, and had no history of intermarriage or migration. Prior to their participation, all volunteers were informed of the purpose of this research and provided with a written informed consent form to sign. Our sample collection and genotyping protocol have been reviewed and approved by the Ethics Committees of Southern Medical University and Xi'an Jiaotong University (No. 2019–1039). For population genetics analysis, the Multi-InDel genotyping data of 26 populations from five continents (Africa, America, Europe, East Asia, and South Asia) were acquired from the 1000 Genomes Project Phase 3 [23], as well as previously published 41 Multi-InDel markers genotyping data from Chinese Manchu and Mongolian groups [19]. All relevant information pertaining to the populations was provided in the Table S1.

DNA extraction and quantification

All blood samples were stored on FTA cards and dried before extraction. A diameter of 1 mm bloodstain sample was prepared for DNA extraction for each sample FTA card. Genomic DNA was extracted from the bloodstain sample according to the instruction following the Chelex-100 method [24]. DNA 9948 and deionized sterile water were used as positive and negative controls, respectively. The concentration and purity of template DNA were determined using a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific Waltham, MA, USA).

PCR amplification, capillary electrophoresis, and genotyping

The 41 Multi-InDel markers and an Amelogenin marker were designed into four lanes, labeled with four different dyes. Table S2 lists the location and fluorescence details of the panel. DNA amplification was performed using a GeneAmp® PCR System 9700 thermal cycler (Applied Biosystems, Foster City, California, USA). The amplification system had a total volume of 10 μL, consisting of 2 μL 2.0× master mix, 1 μL (1 ng) template DNA, 2 μL 1.0× primer mix, and 5 μL nuclease-free water. The diluted amplified product was detected via CE on a 3500xL Genetic Analyzer (Applied Biosystems, Foster City, California, USA). The DNA profiles of the 41 Multi-InDel markers were analyzed by GeneMapper® ID-X 1.3 software. The genotype of each Multi-InDel marker was determined based on the genotype of two InDel loci. Allele 0 represents simultaneous deletion fragments at both InDel loci, while an allele 3 represents the simultaneous insertion fragments at both InDel loci. Allele 1 or 2 represents one InDel locus as the insertion allele and the other locus as the deletion allele. In this study, allele 1 represents a relatively short amplicon, while allele 2 represents a relatively long amplicon. Since the two InDel loci in each of the selected Multi-InDel markers have disparate insertion or deletion fragment lengths, the lengths of the allele 1 and allele 2 amplicons differed.

Data analysis

GenALEx (version 6.5) [25] and GENEPOP (version 4.0.10) [26] were employed to analyze Hardy-Weinberg equilibrium (HWE) and linkage disequilibrium (LD) of Multi-InDel markers in Kazakh and Kyrgyz groups. And *p*-values from HWE and LD tests were adjusted using Bonferroni's correction. Gene frequencies and forensic parameters, including polymorphism information content (PIC), probability of match (PM), power of discrimination (PD), observed heterozygosity (Hobs), and power of exclusion (PE), were calculated for 41 Multi-InDel markers using the STRAF (version 2.1.5) online tool [27]. The relevant forensic parameters were visualized in a split violin plot for the two groups using the ggunchained package of the *R* software (version 4.2.1). The informativeness for assignment (I_n) was calculated using the Infocalc (version 1.1) [28], which is used to quantify the information content of Multi-InDel markers in distinguishing ancestral origions. Gene frequencies and I_n values were visualized using the online website ChiPlot (https://www.chiplot.online/).

In order to assess the efficacy of the Multi-InDel panel for kinship analysis, 10,000 full sibling (FS) pairs, 10,000 half sibling (HS) pairs, and 10,000 pairs of unrelated individuals were simulated by the Familias 3 software [29], based on the gene frequencies for the 41 Multi-InDel markers. Familias 3 was also used to calculate likelihood ratios (LR) for different relationships. The prosecution hypothesis (H0) posited that two individuals are either HS or FS, whereas the defense hypothesis (H1) posited that they are unrelated individuals. The LR distributions for these relationships were visualized using *R* software (version 4.2.1).

The paired *Nei's* genetic distances (D_A distances) [30] and fixation index (F_{ST}) values for a total of 30 populations were obtained using the Dispan program and GenAIEx (version 6.5), respectively. Subsequently, D_A distances and F_{ST} values were visualized using the ggplot2 package of *R* software (version 4.2.1). On the basis of the pairwise D_A distances, a neighbor-joining (NJ) phylogenetic tree was constructed using MEGA 11 software [31] based on the neighbor-joining method and plotted using ChiPlot. Furthermore, principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP) were used to visualize the genetic relationships of the two studied groups and 28 reference populations, which were conducted at both the population and individual levels using the Rtsne and umap packages in *R* software (version 4.2.1).

Additionally, the STRUCTURE (version 2.3.4) [32] was utilized to assess the Bayesian clustering of genotype data in the two studied groups and reference populations. The STRUCTURE program was executed with the following parameters: $K = 2–7$ (15 replicates per K), and 10,000 MCMC iterations. Subsequently, the genetic components for each K value were plotted in a stacked format using the Distruct (version 1.1) [33] software. STRUCTURE HARVESTER [34] was used to evaluate and visualize the likelihood values and to estimate ΔK. The merged Q matrices for the 15 replications of optimal K values were obtained by the CLUMPP (version 1.1.2) [35], and plotted based on CLUMPP results using the AncestryPainterV2 package in *R* (version 4.2.1).

Results

HWE and LD analyses of 41 Multi-InDel markers in the Kazakh and Kyrgyz ethnic groups

In the HWE analysis, the MI38 marker of both Kazakh group ($p = 0.000$) and Kyrgyz group ($p = 0.001$) deviated from HWE after Bonferroni's correction ($p = 0.05/41$), which was consistent with the previous research [18, 20]. Therefore, the MI38 marker was excluded from the subsequent analysis. HWE and LD analyses were performed on the remaining 40 Multi-InDel markers. After Bonferroni's correction, 40 markers in both groups were in accordance with HWE ($p > 0.05/40$), and no significant association was found between paired markers, indicating a state of linkage equilibrium ($p > 0.05/780$) in 40 markers, in these two groups.

Gene frequency distributions and forensic parameters of Multi-InDel markers in two groups

We calculated the gene frequencies and forensic parameters of 40 Multi-InDel markers in the two studied groups, respectively. The gene frequency distributions are shown in the Fig. S1 and Table S7, and most of the 40 Multi-InDel markers in two groups have three alleles. As shown in Fig. 1 and Table S3, in the Kazakh group, the forensic parameters range as follows: PIC from 0.3408 (2MI16) to 0.5898 (2MI58), PM from 0.1883 (MI32) to 0.4130 (2MI16), PD from 0.5870 (2MI16) to 0.8117 (MI32), Hobs from 0.4345 (2MI16) to 0.7310 (2MI58), and PE from 0.1363 (2MI16) to 0.4779 (2MI58). In the Kyrgyz group, these parameters show the following ranges: PIC from 0.3444 (2MI16) to 0.5923 (2MI58), PM from 0.1832 (2MI17) to 0.4024 (2MI16), PD from 0.5976 (2MI16) to 0.8168 (2MI17), Hobs from 0.4200 (2MI16) to 0.7400 (2MI07), and PE from 0.1266 (2MI16) to 0.4928 (2MI07). In addition, the cumulative matching probability (CPM), cumulative power of discrimination (CPD), and cumulative power of exclusion (CPE) of 40 markers in the Kazakh and Kyrgyz groups are 1.64007E-25 and 2.82816E-25; 0.999999999999999999999999835993 and 0.99999999999999999999999971718 4; 0.999998887418153 and 0.999999348634116, respectively, indicating that the panel could be a powerful tool for individual identification and paternity testing in the two groups.

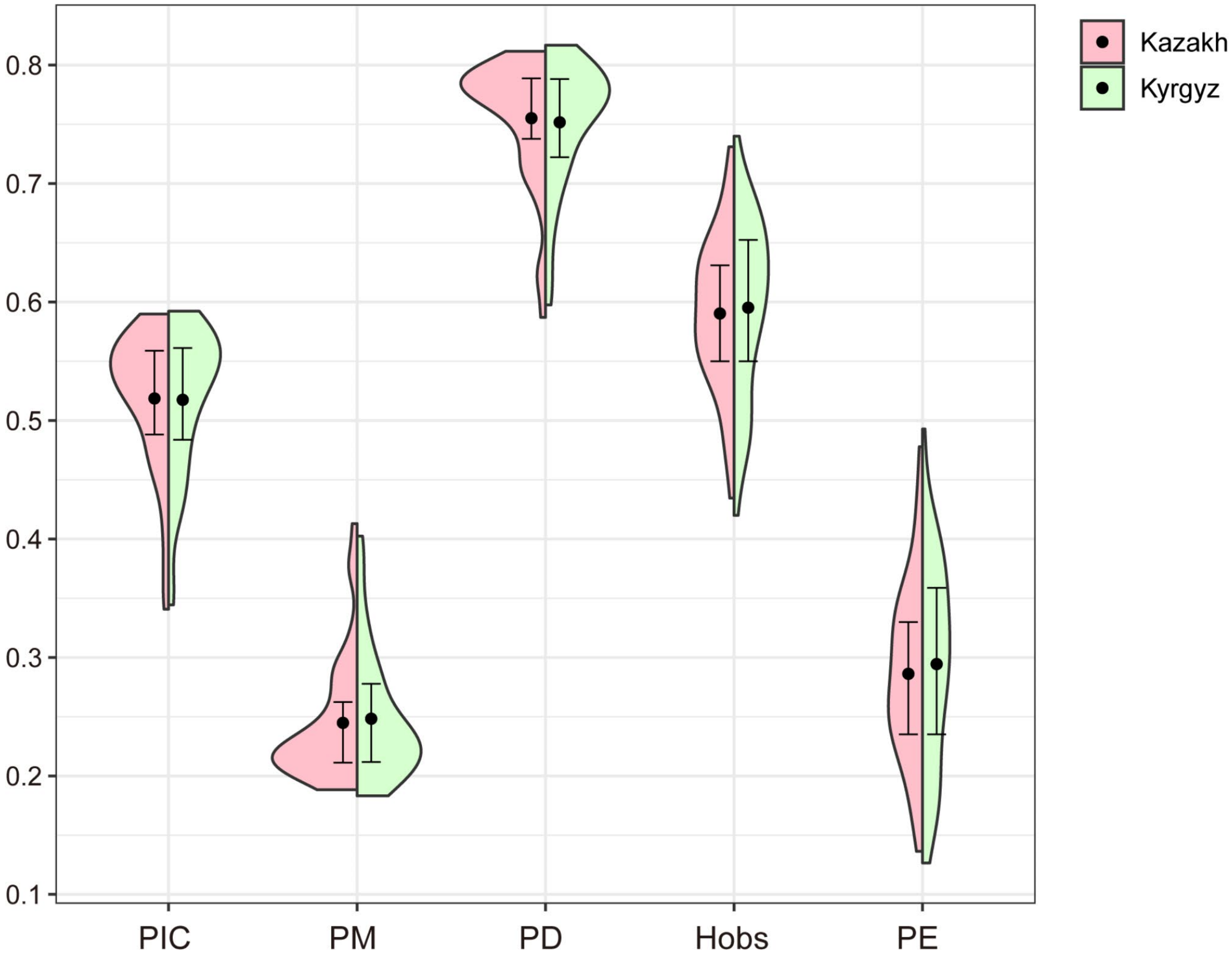

Fig. 1 Forensic parameters of the 40 Multi-InDel markers in Chinese Kazakh and Kyrgyz groups. In this split violin plot, the dots represent the mean values of forensic parameters, and the horizontal lines at both ends of the vertical lines represent the upper and lower quartiles of the data. PIC, polymorphism information content; PM, probability of match; PD, power of discrimination; Hobs, observed heterozygosity; PE, power of exclusion

The effectiveness of Multi-InDel panel in identifying full siblings and half siblings of two groups

We evaluated the forensic efficacy of the Multi-InDel panel for identifying FS and HS pairs in the Kazakh and Kyrgyz groups. The $Log_{10}LR$ distributions between FS pairs and unrelated pairs in the two groups are shown in Fig. 2. The accuracy and false positive rates for identifying FS and HS using LR thresholds are detailed in Table S4. Using the 40 Multi-InDel markers, a similar ability to distinguish FS pairs in the Kazakh and Kyrgyz groups was observed. When LR = 1, 98.89% and 98.74% of FS pairs could be differentiated from unrelated individual pairs in the two groups, with false positive rates of 1.27% and 0.95%, respectively. When the LR thresholds were set at 10, 100, 1,000, and 10,000, the average accuracies of the two groups were 96.03%, 89.12%, 76.34%, and 58.59%, respectively, while the corresponding average false positive rates were 0.27%, 0.04%, 0.00%, and 0.00%, respectively. However, the Multi-InDel panel was less effective in distinguishing HS pairs from unrelated individual pairs. When the LR thresholds were set at 1, 10, and 100, the average accuracies for identifying HS pairs in the two groups were 86.65%, 51.19%, and 1.68%, respectively.

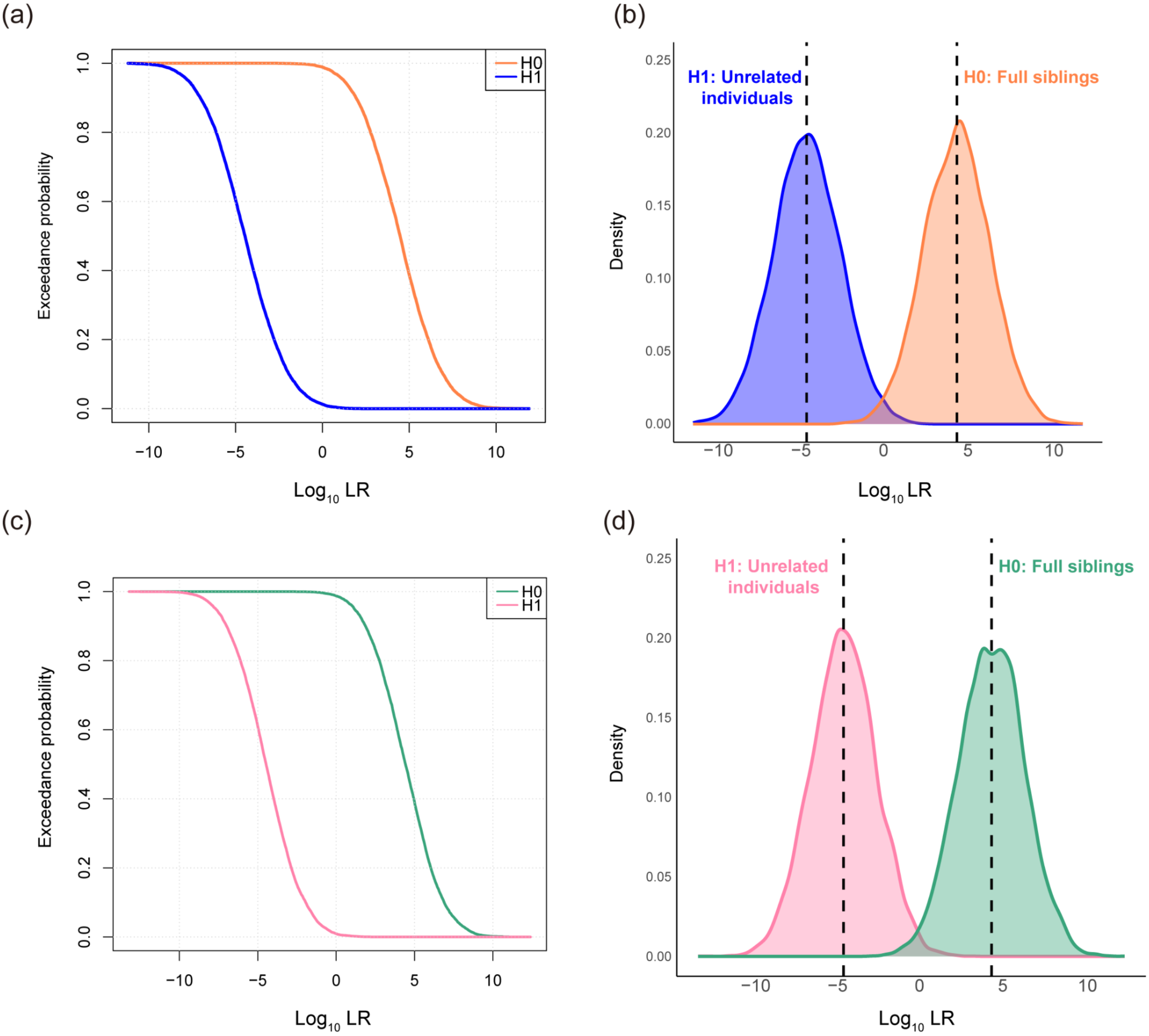

Fig. 2 Results of simulating full sibling (FS) pairs and unrelated individual pairs based on gene frequencies of 40 Multi-InDel markers in Chinese Kazakh and Kyrgyz ethnic groups. (**a**) Exceedance probability curve for $Log_{10}LR$ of FS and unrelated individuals in Chinese Kazakh group; (**b**) Kernel density profile of $Log_{10}LR$ for 10,000 FS pairs and 10,000 unrelated individual pairs simulated in Kazakh group; (**c**) Exceedance probability curve for $Log_{10}LR$ of FS and unrelated individuals in Kyrgyz group; (**d**) Kernel density profile of $Log_{10}LR$ for 10,000 FS pairs and 10,000 unrelated individual pairs simulated in Kyrgyz group

Genetic distances among the two studied groups and 28 reference populations

The F_{ST} value and D_A distance can be used to measure the degree of genetic differentiation between paired populations. Smaller value indicates a lower degree of genetic differentiation between the two populations. The F_{ST} values and D_A distances among the two groups and 28 reference populations are visualized in Fig. 3 and Table S5-S6. The African populations are the most genetically distant from the other four continents. Among these populations, the two studied groups display closer genetic relationships with East Asian populations. The average F_{ST} values between the two studied groups and seven East Asian populations are only 0.0125 (Kazakh) and 0.0121 (Kyrgyz). These values are significantly lower than the average F_{ST} values between the two studied groups and African populations, which are 0.0610 (Kazakh) and 0.0614 (Kyrgyz). It is noteworthy that the Kazakh and Kyrgyz groups exhibit the closest genetic relationship, showing the smallest F_{ST} value of 0.0020. D_A values corroborate these findings, with the average D_A values between the two groups and East Asian populations being 0.0154 (Kazakh) and 0.0152 (Kyrgyz), which are also far lower than the average D_A values between the two groups and other continental populations. The two groups still exhibit the smallest genetic distance from each other, with a D_A of 0.0017. These results highlight the close genetic relationship between the Kazakh and Kyrgyz groups, as well as their closer genetic relationships to East Asian populations compared to other continental populations.

NJ phylogenetic tree construction of the two studied groups and 28 reference populations based on pairwise D_A values

The Fig. 4 presents a NJ phylogenetic tree based on paired D_A values. The NJ phylogenetic tree shows that the 30 populations cluster into three distinct evolutionary branches. African populations form a separate evolutionary branch. East Asian populations form another distinct evolutionary branch. Within this branch, two clades are identified. One clade includes the Kazakh group and the other clade consists of the Kyrgyz, CHB, CHS, CDX, Manchu, Mongolian, JPT, and KHV populations. Whereas, European, American, and South Asian populations cluster into a single branch. And this branch is further subdivided into two clades. And one clade consists of European and American populations; the other clade consists of South Asian populations. While, the four American populations did not form a separate branch due to the presence of genetically mixed ancestral components.

I_n values for two studied groups based on 40 Multi-InDel genotype data

The capacity of the 40 Multi-InDel markers to provide information regarding an individual's ancestral information in the two studied groups and the reference

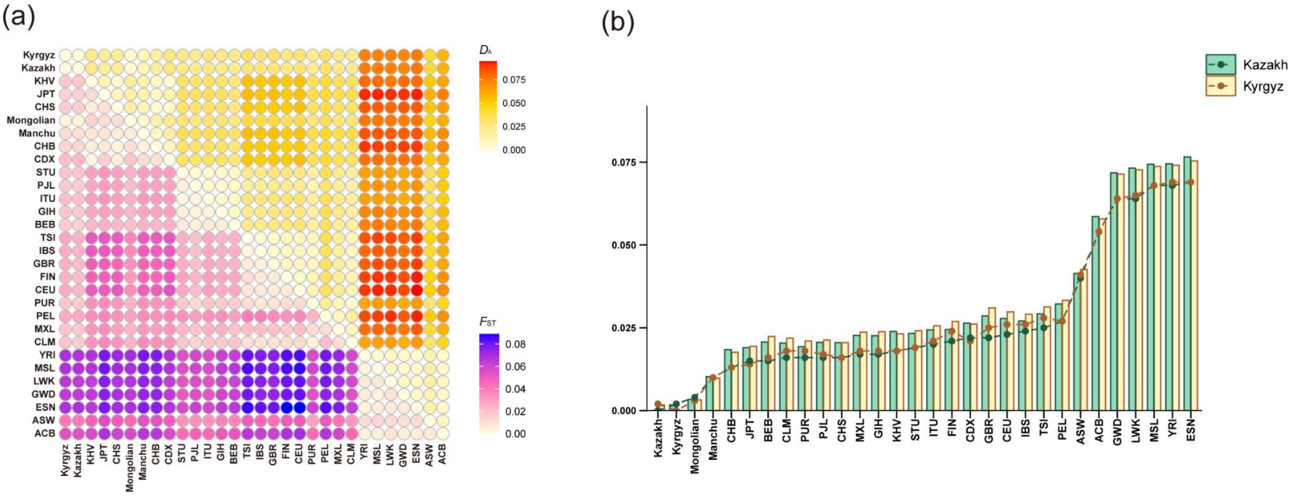

Fig. 3 F_{ST} values and D_A values for paired groups. (**a**) Heatmap of F_{ST} values (lower left corner) and D_A values (upper right corner) among paired groups. (**b**) F_{ST} values (line chart) and D_A values (histogram) among the two studied groups and other reference populations. KHV, Kinh in Ho Chi Minh City, Vietnam; JPT, Japanese in Tokyo, Japan; CHS, Southern Han Chinese, China; Mongolian; Manchu; CHB, Han Chinese in Beijing, China; CDX, Chinese Dai in Xishuangbanna, China; STU, Sri Lankan Tamil from the UK; PJL, Punjabi from Lahore, Pakistan; ITU, Indian Telugu from the UK; GIH, Gujarati Indian from Houston, Texas; BEB, Bengali from Bangladesh; TSI, Toscani in Italia; IBS, Iberian Population in Spain; GBR, British in England and Scotland; FIN, Finnish in Finland; CEU, Utah Residents (CEPH) with Northern and Western European Ancestry; PUR, Puerto Ricans from Puerto Rico; PEL, Peruvian in Lima, Peru; MXL, Mexican Ancestry from Los Angeles, USA; CLM, Colombians from Medellin, Colombia; YRI, Yoruba in Ibadan, Nigeria; MSL, Mende in Sierra Leone; LWK, Luhya in Webuye, Kenya; GWD, Gambian in Western Divisions in the Gambia; ESN, Esan in Nigeria; ASW, African Americans from the Southwest USA; ACB, African Caribbean in Barbados

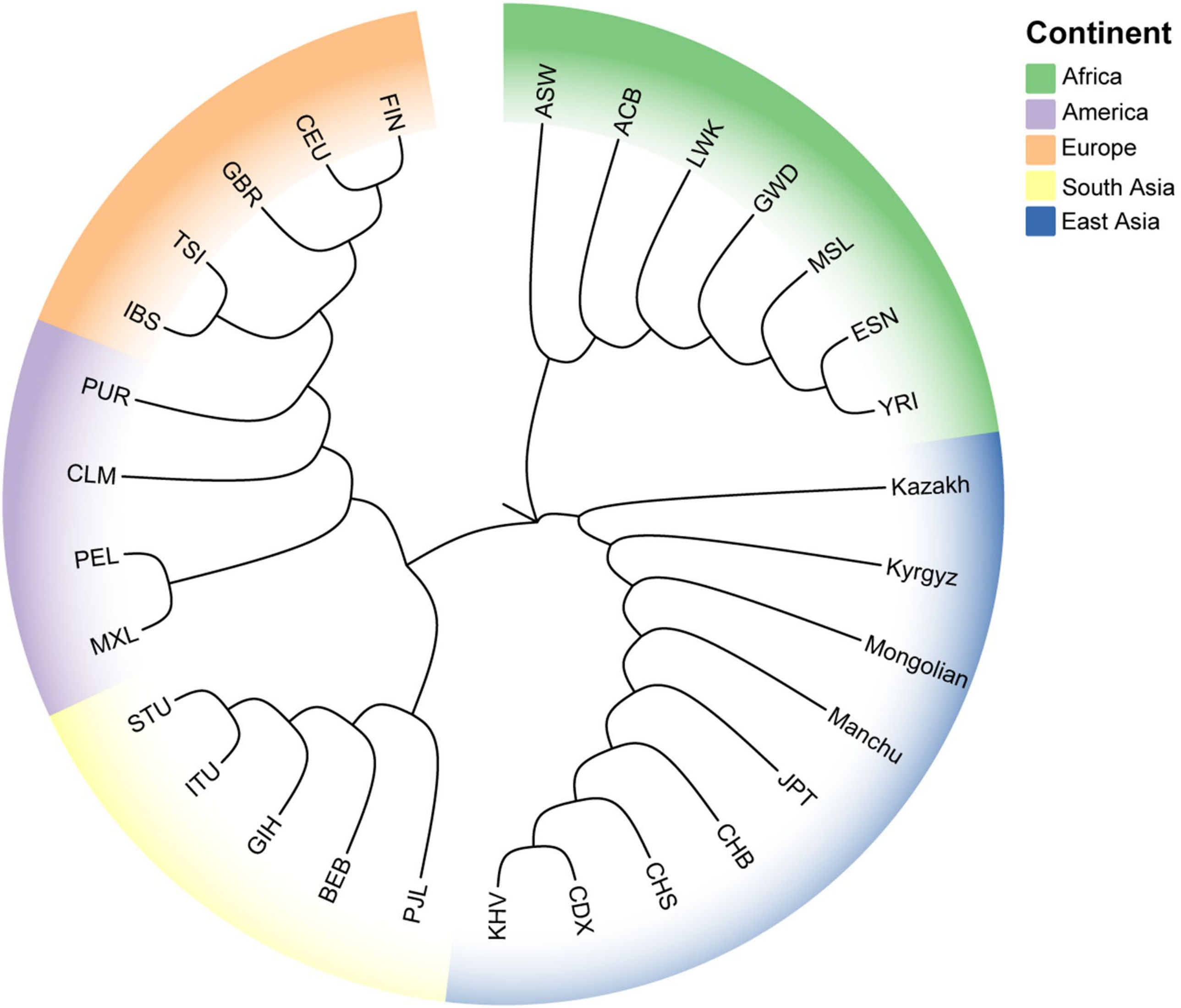

Fig. 4 NJ phylogenetic tree of the two studied groups and 28 reference populations based on paired D_A values

populations was evaluated using parameter I_n. A higher I_n value indicates a stronger ability of Multi-InDel marker to infer the ancestry of unknown individual. We calculated the I_n values for the 40 Multi-InDel markers to distinguish the two studied groups from three (Africa, Europe, and East Asia) and five (Africa, Europe, East Asia, South Asia, and America) continents, denoted as the parameters I_n3 and I_n5, respectively. As demonstrated in Fig. 5 and Table S8, the I_n values of most genetic markers are concentrated between 0.2 and 0.3. Additionally, 85% (34/40) and 80% (32/40) of the Multi-InDel markers exhibited I_n3 and I_n5 values exceeding 0.1 in the Kazakh and Kyrgyz groups. Notably, I_n5 and I_n3 were observed to be greater than 0.3 for 2MI54, 2MI49, and MI26 markers in both studied groups, indicating their efficacies in providing meaningful insights into the ancestral backgrounds of the two groups.

Dimensionality reduction analyses for 40 Multi-InDel markers of the two studied groups and 28 reference populations

To facilitate comprehension and analyses of the data structures and patterns, we projected high-dimensional data into a two-dimensional space. Population-level and individual-level PCA, t-SNE, and UMAP dimensionality reduction analyses were conducted and visualized for the two studied groups and 28 reference populations based on gene frequencies and raw genotypes of 40 Multi-InDel markers, respectively. These results are visualized in Fig. 6. At the population level, the first two principal components of PCA explained 77.3% of the total variance. The Fig. 6a depicts the 30 populations, which are roughly clustered into four clusters, i.e. the African populations cluster on the right (green), East Asian populations cluster on the top left (purple), European

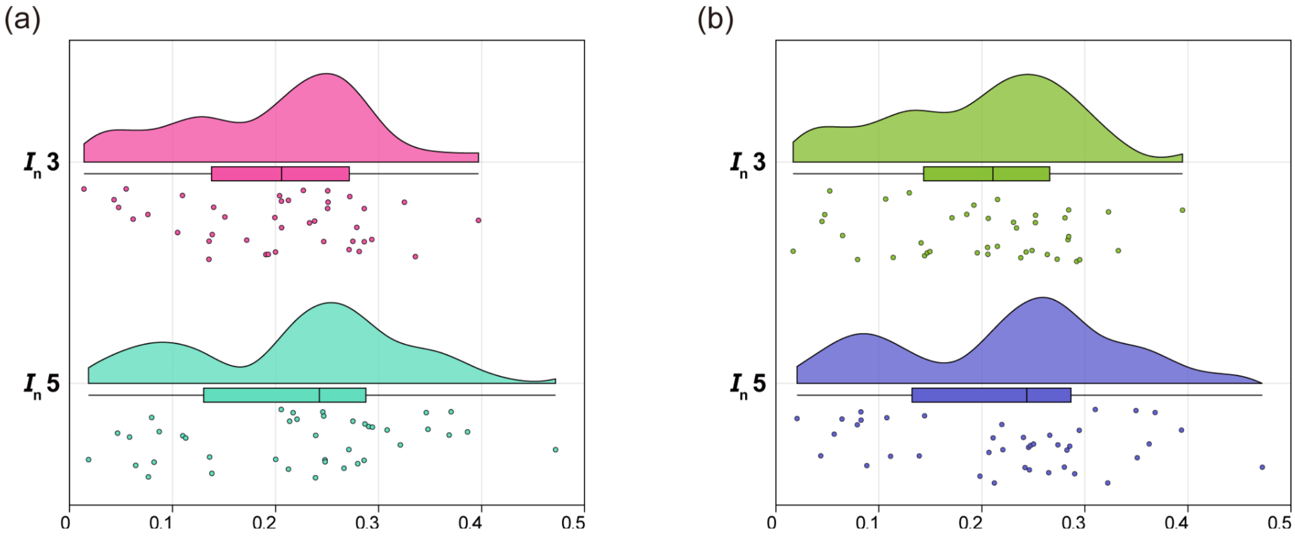

Fig. 5 *In* values based on 40 Multi-InDel markers when distinguishing between two studied groups and the populations from three or five continents. (**a**) *In* values of the 40 Multi-InDel markers for distinguishing the Kazakh group from three intercontinental populations (*In*3) and five continental populations (*In*5); (**b**) *In* values of the 40 Multi-InDel markers for distinguishing the Kyrgyz group from three intercontinental populations (*In*3) and five intercontinental populations (*In*5)

populations cluster on the bottom left (blue), and South Asian populations cluster in the center (yellow) in the PCA. In the Fig. 6b, the result of t-SNE indicates that populations from the same continent tend to cluster with each other and there is no overlap in distribution, and further distinguishes the South Asian populations compared to PCA. In the Fig. 6c, the distribution pattern of UMAP is roughly similar to that of t-SNE, but there is overlap. The genetic relationships of the two studied groups were further analyzed in relation to the East Asian, European, and African populations by individual-level dimensionality reduction with the genotypes of 40 Multi-InDel markers. The Fig. 6d-f illustrate that a total of 2,213 individuals from three continents (East Asia, Africa and Europe) are divided into three clusters. Of these, 245 Kazakh and Kyrgyz individuals are superimposed on the East Asian and European individuals, with a greater degree of overlap observed between them and the East Asian individuals. This pattern supports closer genetic affinities between the studied groups and the East Asian populations.

Population genetic structure analyses among the two studied groups and 28 reference populations

We conducted individual-level and population-level ancestral component analyses based on the genotyping data of 40 Multi-InDel markers to assess the population structures of the two studied groups and 28 reference populations. The results for $K = 2$–7 are depicted in Fig. S2, which presents the stacked plots at both the individual and population levels. When $K = 2$, the genetic structure analysis identifies African and non-African ancestral components. As the K value increases from three to five, the Europe, East Asia, South Asia and America are distinguished by exhibiting unique ancestral components. And the two studied groups (Kazakh and Kyrgyz) and East Asian populations exhibit similar genetic structures. The optimal K value of three was determined according to STRUCTURE HARVESTER. Fig. 7 illustrates the individual-level stacked plot and ternary diagram for the optimal K value ($K = 3$). Fig. 7a shows that the maximum ancestral components of the two studied groups are similar to those of the East Asian populations. The estimated East Asian ancestry components of the Kazakh and Kyrgyz groups are 63.67% and 68.02%, respectively, and a certain percentage of European ancestry components also detected, accounting for 33.08% and 28.99% of their genetic compositions, respectively.

Because the South Asian and American populations in the 1000 Genomes Project have genetic characteristics of mixed origins which are not conducive to analytical interpretation of the results, these two intercontinental populations were not included in the triangular clustering analysis. The Fig. 7b shows that the 40 Multi-InDel markers can effectively discriminate African, European and East Asian individuals. Both Kazakh and Kyrgyz individuals partially overlap with East Asian and European individuals and share similar genetic structures.

Discussion

The Multi-InDel marker was first proposed by Huang et al. in 2014 [13]. This marker exhibits the characteristics of a low mutation rate and a short amplicon, similar to that of MH. Of particular significance is its ability to be genotyped using the CE method, which is a common practice in forensic laboratory. In this study, 145 Kazakh

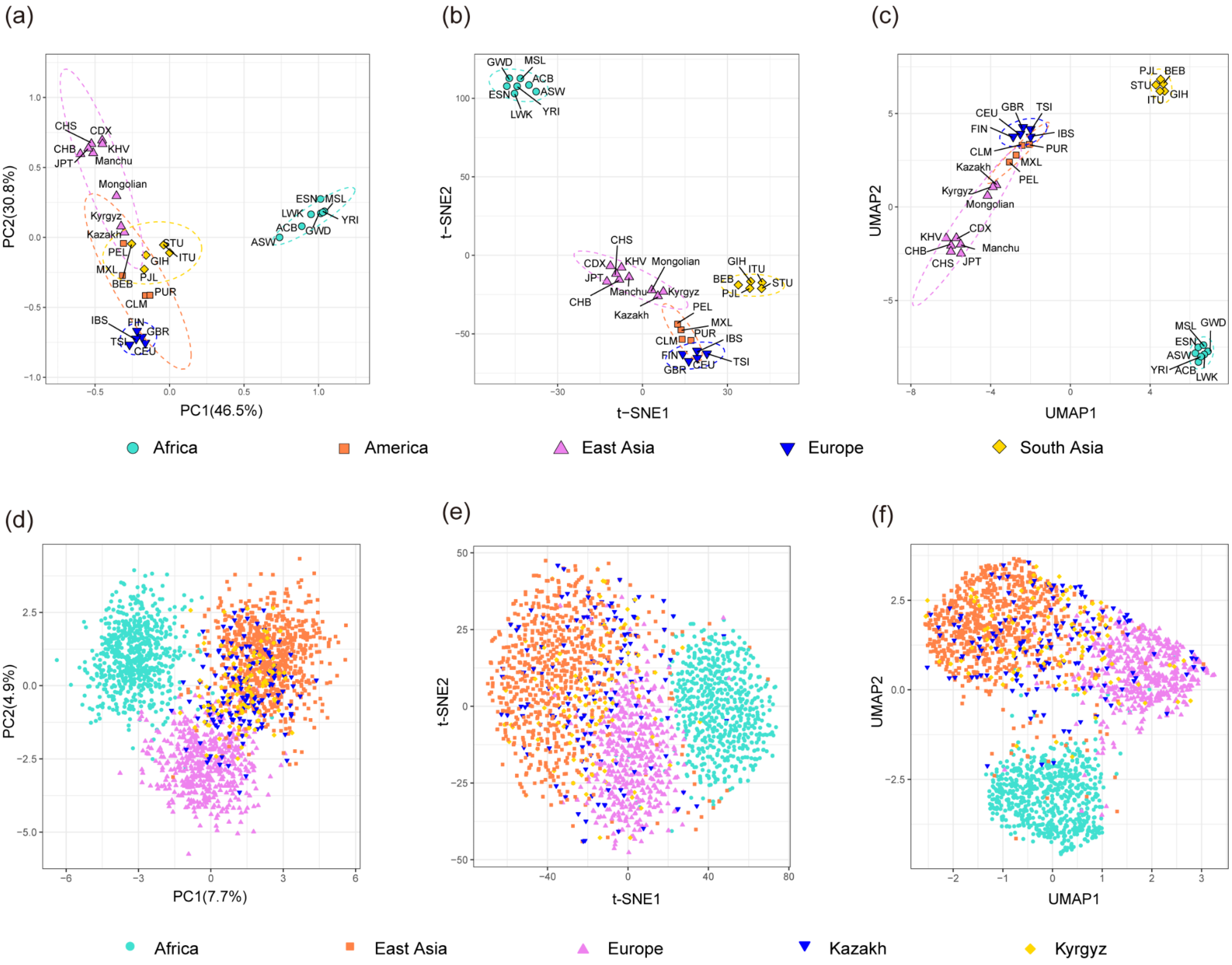

Fig. 6 Population-level and individual-level PCA, t-SNE and UMAP dimensionality reduction analyses based on 40 Multi-InDel markers. (**a**) Population-level PCA of two studied groups and the reference populations at PC1 and PC2; (**b**) Population-level t-SNE of two studied groups and the reference populations; (**c**) Population-level UMAP of two studied groups and the reference populations; (**d**) PCA of the overall individuals from three continents (Africa, Europe, and East Asia); (**e**) t-SNE of the overall individuals from three continents (Africa, Europe, and East Asia); (**f**) UMAP of the overall individuals from three continents (Africa, Europe, and East Asia)

and 100 Kyrgyz unrelated healthy individuals from China were genotyped using a panel containing 41 Multi-InDel markers and a sex-determination locus Amelogenin. Subsequently, the forensic parameters and genetic polymorphisms of the 41 Multi-InDel were subjected to comprehensive assessment. Furthermore, genotyping data on 41 Multi-InDel genotypes from 28 reference populations were collected in order to explore the genetic differentiations and relationships between the two studied groups and other reference populations. We found that the Multi-InDel panel qualified as an effective tool for individual identification and paternity testing of Chinese Kazakh and Kyrgyz groups, as well as for full sibling kinship identification. In addition, we provided evidence for the genetic relationships of Chinese Kazakh and Kyrgyz groups with East Asian and European populations.

Following Bonferroni's correction, 40 out of 41 Multi-InDel markers in two studied groups demonstrated HWE, with the exception of MI38. This deviation may be attributed to the fact that the marker was purely summed and the gene frequencies were not balanced in most of the samples in both studied groups. The 40 markers were in linkage equilibrium with each other, indicating their mutual independence. Consequently, the product rule can be applied to calculate the cumulative probabilities for this Multi-InDel panel, specifically CPD, CPE and CPM.

The Multi-InDel markers have been developed with the objective of enhancing genetic polymorphisms through the introduction of new alleles. Most of the 40 Multi-InDel markers in two studied groups have three alleles (Fig. S1), exhibiting higher polymorphisms than InDel. Genetic markers with PIC values >0.5 were considered

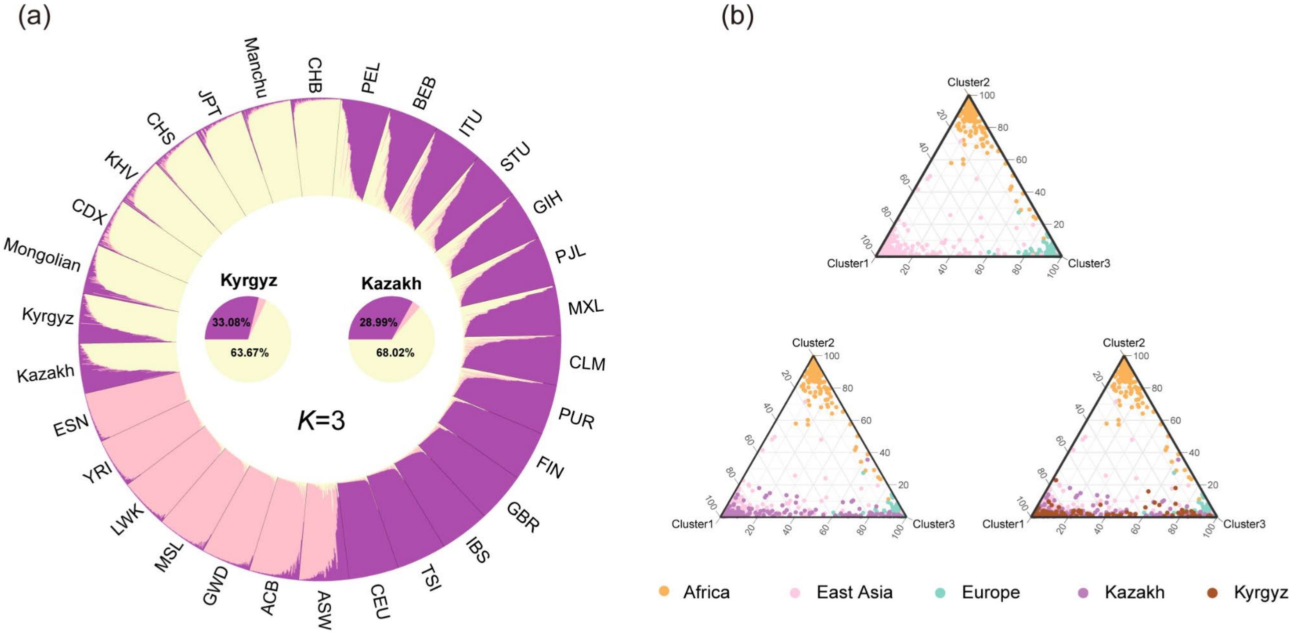

Fig. 7 Genetic structure analysis plots of the two studied groups and 28 reference populations. (**a**) Individual-level ancestral structures of the 30 populations when $K=3$, and the pie chart of ancestral compositions of Kazakh and Kyrgyz group. (**b**) Triangular clustering diagram of African, East Asian, and European individuals at $K=3$, with stepwise addition of the Kazakh and Kyrgyz individuals

to possess high information content [36]. As listed in Table S3, the mean PIC values of 40 Multi-InDel markers are 0.5186 and 0.5175 in the Kazakh and Kyrgyz groups, respectively, and both of them have 72.5% (29/40) markers with PIC values greater than 0.5. These results also indicate high polymorphisms of these markers. The CPD and CPE for the 40 Multi-InDel markers in the Kazakh and Kyrgyz ethnic groups were 0.999999999999999999999999835993 and 0.999999999999999999999999717184; 0.999998887418153 and 0.999999348634116, respectively. This suggested that analyses of the 40 Multi-InDel markers were eligible in individual identification and paternity testing. The 40 Multi-InDel markers demonstrated higher CPD and CPE values than the 20 Multi-InDel markers reported by Huang et al. [13], the 17 Multi-InDel markers by Qu et al. [16], and the 20 Multi-InDel markers by Liu et al. [17]. This may indicate that this panel is more effective for individual identification and paternity testing, though the higher efficacy may be attributed to the fact that the Multi-InDel panel contains more markers than the other panels. Additionally, the capability of this Multi-InDel panel is essentially equivalent to the STR panel for the two groups that have been studied [37–39]. Furthermore, the 40 Multi-InDel markers were employed for the identification of FS, HS and unrelated individual pairs. When the LR threshold was set at 1, the 40 Multi-InDel markers yielded meaningful conclusions in the context of FS identification cases, while its ability to differentiate HS from unrelated individuals was relatively limited.

In forensic practice, ancestry information can provide crucial insights that narrow the scope of an investigation when DNA database matches are unavailable or there is a lack of reliable eyewitness testimony. Multi-InDel markers have the potential to serve as ancestry inference markers [14]. Firstly, the populations from the same continent have small F_{ST} and D_A values, and the Kazakh and Kyrgyz groups exhibited the smallest F_{ST} and D_A values with East Asian populations, in particular the genetic distance between the Kazakh and Kyrgyz groups is the smallest. The NJ phylogenetic analysis also confirmed that there were more genetic correlations between the two studied groups and East Asian populations. Secondly, the applications of PCA, t-SNE and UMAP revealed the aggregation of populations from the same continent into a single cluster. Moreover, the latter two methods demonstrated superior performance in terms of population-level distribution effects. The t-SNE better addresses the crowding problem in high-dimensional data through its nonlinear dimensionality reduction and retention of local structure, which makes the distribution of the reduced data in the low-dimensional space more uniform and intuitive. And UMAP focuses on preserving the global structure, resulting in smaller distances within the cluster. At the individual level, the three continents (Africa, Europe, and East Asia) were well differentiated, and individuals from two groups were distributed between East Asian and European individuals, with a greater overlap observed between them and the East Asian individuals. This suggests that they were more genetically related to

East Asian populations. Furthermore, the results of the population structure analyses indicated that, within the range of *K* values between two and five, populations from five continents were progressively differentiated, and the genetic compositions of the two groups are comparable to East Asian populations. At the optimal *K* value of three, the East Asian ancestry components of Kazakh and Kyrgyz groups were 63.67% and 68.02%, with European ancestry components of 33.08% and 28.99%, respectively. The triangular clustering diagram also showed that individuals from these two groups were distributed between the East Asian and European individuals. This suggests stronger genetic affinities between these two groups and East Asian populations, as well as the possibility of gene admixture with East Asian and European populations. It should be noted, however, that only five populations in East Asia and two previously studied groups were selected as reference populations in this study, and cannot fully represent the entire East Asian populations. The same is true for the reference populations in Africa, America, Europe and South Asia. Therefore, more populations are needed in future study to further confirm the robustness of the panel.

Research based on autosomal STRs and Y-SNPs/STRs had demonstrated that the genetic component of Kyrgyz group was similar to both East Asian and European populations [40]. Furthermore, genome-wide SNP studies of present-day Chinese Kazakh [21], Mongolian [41], and Kyrgyz [42] groups also support their East Asia-Europe mixing pattern. Additionally, studies of autosomal STRs [38, 43] and DIPs [44] suggested that the Kyrgyz and Kazakh groups exhibited close genetic affinities. These findings corroborate our results. In the modern era, the Kazakh group was compelled to migrate in significant numbers to the Ili, Tacheng and Altai regions of northwest China. They subsequently continued to migrate to the northern foothill of the Tianshan Mountain, which altered and shaped the pattern of ethnic distribution and ethnic relation in the region [45]. The Kyrgyz group first inhabited the Yenisei River basin. Due to war and other factors, they experienced five westward migrations between the Western Han Dynasty and the middle of the Qing Dynasty, reaching as far as the Western Tian Shan and Central Asia [46]. Currently, they reside in mixed communities with the Kazakh, Uyghur, Mongolian, and Han Chinese groups. Both studied groups reside in the northwestern region of China, which was the route of ancient Silk Road, an important link for the exchange of good, plant, animal, and idea between the people of East Asia and Europe [47]. Due to their special geographic location and multiethnic gathering, population movements and intermarriages were inevitable [48], which may explain the fact that the two studied groups in our results consisted mainly of genetic components from East Asian and European populations and had the smallest genetic distance between them.

Conclusions

In this study, a total of 245 individuals from Chinese Kazakh and Kyrgyz ethnic groups were conducted forensic and population genetic analyses using a self-developed panel containing 41 Multi-InDel markers. The results showed that these 40 Multi-InDel markers (except MI38) can be used as effective tools for individual identification and paternity testing of two studied groups, and play a potential role in full sibling identification. In addition, population genetic analyses further elucidated the East Asia-Europe admixed ancestry components in the Kazakh and Kyrgyz groups, while demonstrating closer genetic affinities with East Asia populations. Ancestral components of five intercontinental populations can be preliminarily inferred based on the 40 Multi-InDel markers. At the same time, we expanded the genetic dataset of these two ethnic groups. Overall, this study demonstrates that the Multi-InDel panel can play an important role in forensic application and ancestry inference. Future investigations with more groups are needed to confirm the robustness of the panel.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s41065-025-00416-5.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Acknowledgements

We are very grateful to the volunteers who provided samples.

Author contributions

Bofeng Zhu and Chunmei Shen designed the study. Qinglin Liang and Qiong Lan conducted the experiments. Qinglin Liu, Lisiteng Luo, and Xiaolian Wu analyzed the raw data. Qinglin Liang wrote the manuscript and prerared Figs. 1, 2, 3, 4, 5, 6 and 7. All authors reviewed the manuscript. All authors read and finalized the manuscript.

Funding

This study was supported by the National Natural Science Foundation of China (No. 82471917) and the Guangdong Provincial Science and Technology Programme (2024A1111120006).

Data availability

The raw genotype data used and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The idea, technical route and implementation of this study which involved human samples were approved by the Ethics Committees of Southern Medical University and Xi'an Jiaotong University (No. 2019–1039). All the volunteers were informed of the purpose and significance of the study, and signed written informed consents. To protect the individual privacies of these volunteers, all the samples were anonymized by numbering during the experiments.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 21 October 2024 / Accepted: 12 March 2025
Published online: 16 May 2025

References

1. Butler JM. Forensic DNA typing: biology, technology, and genetics of STR markers. Elsevier; 2005.
2. Barrio PA, Martín P, Alonso A, Müller P, et al. Massively parallel sequence data of 31 autosomal STR loci from 496 Spanish individuals revealed concordance with CE-STR technology and enhanced discrimination power. Forensic Sci International: Genet. 2019;42. https://doi.org/10.1016/j.fsigen.2019.06.009.
3. Ensenberger MG, Lenz KA, Matthies LK, Hadinoto GM, et al. Developmental validation of the PowerPlex(®) Fusion 6 C System. Forensic Sci Int Genet. 2016;21. https://doi.org/10.1016/j.fsigen.2015.12.011.
4. Haidar M, Abbas FA, Alsaleh H, Haddrill andPR. Population genetics and forensic utility of 23 autosomal PowerPlex Fusion 6 C STR loci in the Kuwaiti population. Sci Rep. 2021;11(1). https://doi.org/10.1038/s41598-021-81425-y.
5. Butler JM, Coble MD, Vallone andPM. STRs vs. SNPs: thoughts on the future of forensic DNA testing. Forensic Sci Med Pathol. 2007;3(3). https://doi.org/10.1007/s12024-007-0018-1.
6. Kidd KK, Pakstis AJ, Speed WC, Lagace R, et al. Microhaplotype loci are a powerful new type of forensic marker. Forensic Sci International: Genet Supplement Ser. 2013;4(1). https://doi.org/10.1016/j.fsigss.2013.10.063.
7. Kidd KK, Pakstis AJ, Gandotra N, Scharfe C, et al. A multipurpose panel of microhaplotypes for use with STR markers in casework. Forensic Sci International: Genet. 2022;60. https://doi.org/10.1016/j.fsigen.2022.102729.
8. Weber JL, David D, Heil J, Fan Y, et al. Human diallelic insertion/deletion polymorphisms. Am J Hum Genet. 2002;71(4). https://doi.org/10.1086/342727.
9. Mills RE, Luttig CT, Larkins CE, Beauchamp A, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. 2006;16(9). https://doi.org/10.1101/gr.4565806.
10. Gao TZ, Yun LB, He W, Gu Y, et al. The application of multi-InDel as supplementary in paternity cases with STR mutation. Forensic Sci International: Genet Supplement Ser. 2015;5. https://doi.org/10.1016/j.fsigss.2015.09.087.
11. Jin R, Cui W, Fang Y, Jin X, et al. A novel panel of 43 insertion/deletion loci for human identifications of forensic degraded DNA samples: development and validation. Front Genet. 2021;12. https://doi.org/10.3389/fgene.2021.610540.
12. Avellaneda LL, Johnson DT, Gutierrez R, Thompson L, et al. Development of a novel five-dye panel for human identification insertion/deletion (INDEL) polymorphisms. J Forensic Sci. 2024;69(3). https://doi.org/10.1111/1556-4029.15475.
13. Huang J, Luo H, Wei W. andY. Hou. A novel method for the analysis of 20 multi-Indel polymorphisms and its forensic application. Electrophoresis. 2014;35(4). https://doi.org/10.1002/elps.201300346.
14. Sun K, Ye Y, Luo T. andY. Hou. Multi-InDel analysis for ancestry inference of Sub-Populations in China. Sci Rep. 2016;6. https://doi.org/10.1038/srep39797.
15. Sun K, Yun L, Zhang C, Shao C, et al. Evaluation of 12 Multi-InDel markers for forensic ancestry prediction in Asian populations. Forensic Sci Int Genet. 2019;43. https://doi.org/10.1016/j.fsigen.2019.102155.
16. Qu S, Lv M, Xue J, Zhu J et al. Multi-Indel: A microhaplotype marker can be typed using capillary electrophoresis platforms [Original Research]. Front Genet. 2020;11.
17. Liu J, Zhang X, Zhang X, Li W, et al. A new set of 20 Multi-InDel markers for forensic application. Electrophoresis. 2022;43(11). https://doi.org/10.1002/elps.202100361.
18. Mei S, Yi S, Cai M, Zhang Y, et al. Exploring the forensic effectiveness and population genetic differentiation by self-constructed 41 multi-InDel panel in Yunnan Zhuang group. Gene. 2023;860. https://doi.org/10.1016/j.gene.2023.147180.
19. Lan Q, Cai M, Lei F, Shen C, et al. Systematically exploring the performance of a self-developed Multi-InDel system in forensic identification, ancestry inference and genetic structure analysis of Chinese Manchu and Mongolian groups. Forensic Sci Int. 2023;346. https://doi.org/10.1016/j.forsciint.2023.111637.
20. Yuan X, Wang X, Lan Q, Li S, et al. Using two self-developed indel panels to explore forensic traits and ancestral components in the Hui group. Genomics. 2024;116(1). https://doi.org/10.1016/j.ygeno.2023.110756.
21. Lei C, Liu J, Zhang R, Pan Y, et al. Ancestral origins and admixture history of Kazakhs. Mol Biol Evol. 2024;41(7). https://doi.org/10.1093/molbev/msae144.
22. Chen H, andS. Xu. Population genomics advances in frontier ethnic minorities in China. Sci China Life Sci. 2024. https://doi.org/10.1007/s11427-024-2659-2.
23. Auton A, Brooks LD, Durbin RM, Garrison EP, et al. A global reference for human genetic variation. Nature. 2015;526(7571). https://doi.org/10.1038/nature15393.
24. Walsh PS, Metzger DA. andR. Higuchi. Chelex® 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material [Article]. BioTechniques. 1991;10(4).
25. Peakall andP R, Smouse E. GenAlEx 6.5: genetic analysis in excel. Population genetic software for teaching and research–an update. Bioinformatics. 2012;28(19). https://doi.org/10.1093/bioinformatics/bts460.
26. Rousset F. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Mol Ecol Resour. 2008;8(1). https://doi.org/10.1111/j.1471-8286.2007.01931.x.
27. Gouy A, andM. Zieger. STRAF-A convenient online tool for STR data evaluation in forensic genetics. Forensic Sci Int Genet. 2017;30. https://doi.org/10.1016/j.fsigen.2017.07.007.
28. Rosenberg NA, Li LM, Ward R, Pritchard andJK. Informativeness of genetic markers for inference of ancestry**. Am J Hum Genet. 2003;73(6). https://doi.org/10.1086/380416.
29. Kling D, Tillmar AO. andT. Egeland. Familias 3 - Extensions and new functionality. Forensic Sci Int Genet. 2014;13. https://doi.org/10.1016/j.fsigen.2014.07.004.
30. Nei M, Tajima F. andY. Tateno. [Article]ccuracy of [Article]stimated phylogenetic [Article]rees from molecular data - II. Gene frequency data [Article]. J Mol Evol. 1983;19(2). https://doi.org/10.1007/BF02300753.
31. Tamura K, Stecher G. andS. Kumar. MEGA11: molecular evolutionary genetics analysis version 11. Mol Biol Evol. 2021;38(7). https://doi.org/10.1093/molbev/msab120.
32. Pritchard JK, Stephens M. andP. Donnelly. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2). https://doi.org/10.1093/genetics/155.2.945.
33. Rosenberg NA. Distruct: a program for the graphical display of population structure. Mol Ecol Notes. 2004; 4(1).
34. Earl andB DA, vonHoldt. M. Structure harvester: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour. 2012;4(2). https://doi.org/10.1007/s12686-011-9548-7.
35. Jakobsson andN M, Rosenberg A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics. 2007;23(14). https://doi.org/10.1093/bioinformatics/btm233.
36. Botstein D, White RL, Skolnick M, Davis andRW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. 1980;32(3).
37. Feng C, Wang X, Wang X, Yu H, et al. Genetic polymorphisms, forensic efficiency and phylogenetic analysis of 15 autosomal STR loci in the Kazak population of Ili Kazak autonomous Prefecture, Northwestern China. Ann Hum Biol. 2018;45(2). https://doi.org/10.1080/03014460.2018.1445289.
38. Guo Y, Chen C, Xie T, Cui W, et al. Forensic efficiency estimate and phylogenetic analysis for Chinese Kyrgyz ethnic group revealed by a panel of 21 short tandem repeats. R Soc Open Sci. 2018;5(6). https://doi.org/10.1098/rsos.172089.
39. Zhang H, Yang S, Guo W, Ren B, et al. Population genetic analysis of the globalfiler STR loci in 748 individuals from the Kazakh population of Xinjiang in Northwest China. Int J Legal Med. 2016;130(5). https://doi.org/10.1007/s00414-016-1319-2.
40. Fang Y, Mei S, Zhang Y, Teng R, et al. Forensic and genetic landscape explorations of Chinese Kyrgyz group based on autosomal SNPs, Y-chromosomal SNPs and STRs. Gene. 2022;832. https://doi.org/10.1016/j.gene.2022.146552.
41. Zhao J, Wurigemule J, Sun Z, Xia, et al. Genetic substructure and admixture of Mongolians and Kazakhs inferred from genome-wide array genotyping. Ann Hum Biol. 2020;47:7–8. https://doi.org/10.1080/03014460.2020.1837952.

42. Halili B, Yang X, Wang R, Zhu K et al. Inferring the population history of Kyrgyz in Xinjiang, Northwest China from genome-wide array genotyping. Am J Biol Anthropol. 2023;181(4).
43. Chen P, Zou X, Wang B, Wang M, et al. Genetic admixture history and forensic characteristics of Turkic-speaking Kyrgyz population via 23 autosomal STRs. Ann Hum Biol. 2019;46(6). https://doi.org/10.1080/03014460.2019.1676918.
44. Xie M, Li Y, Wu J, Song F, et al. Genetic structure and forensic characteristics of the Kyrgyz population from Kizilsu Kirghiz autonomous Prefecture based on autosomal dips. Int J Legal Med. 2022;136(2). https://doi.org/10.1007/s00414-020-02277-1.
45. Wang Z. The migration of Kazakhs and the Interactions and interchanges of ethnic groups in modern Xinjiang. Ethno-National Stud. 2023;(03).
46. Yang Y. Westward migration and Lts influence on the formation of Kirgiz. J North Minzu University(Philosophy Social Science). 2015;(03).
47. X, An. The role of cross-border Kazakh ethic group in the construction of the silk road economic belt. West Leather. 2017;39(23).
48. Liang H. Study on the Inter-ethnic marriage of floating population and its changing trend. Popul Soc. 2023;39(01). https://doi.org/10.14132/j.2095-7963.2023.01.005.

Publisher's note