



Accepted Article

Development one novel multiplex PCR system for forensic individual identification using insertion/deletion polymorphisms


Xiao-Ye Jin^{1,2,3}, Yuan-Yuan Wei^{1,2}, Wei Cui^{1,2,3}, Chong Chen^{1,2,3}, Yu-Xin Guo^{1,2,3}, Wen-Qing Zhang^{1,2}, Bo-Feng Zhu^{1,2,4}

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 and Forensics, Xi'an Jiaotong University Health Science Center, Xi'an, 710061, P. R. China

⁴ Department of Forensic Genetics, School of Forensic Medicine, Southern Medical University, Guangzhou, 510515, P. R. China

Corresponding author: Bo-Feng Zhu; e-mail: zhubofeng7372@126.com; 
<https://orcid.org/0000-0002-9038-2342>✓

Abstract:

Insertion/deletion (InDel) polymorphisms have been widely used in the fields of population genetics, genetic map constructions and forensic investigations owing to the advantages of their low mutation rates, widespread distributions in the human genome and small amplicon sizes. In order to provide more InDels with high discrimination power in Chinese populations, we selected and constructed one novel multiplex InDel panel for forensic individual identification. Genetic

Received: 09 30, 2018; Revised: 03 12, 2019; Accepted: 03 21, 2019

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1002/elps.201800412](https://doi.org/10.1002/elps.201800412).

This article is protected by copyright. All rights reserved.

distributions of these 35 InDels in five reference populations from East Asia showed low genetic differentiations among these populations. Forensic efficiency evaluations of these InDels revealed these loci could perform well for forensic individual identifications in these reference populations. In the meantime, genetic diversities and forensic parameters of these InDels were further investigated in the studied Kazak group. Mean value of polymorphism information content for 35 InDels was 0.3611. Cumulative power of discrimination of 35 InDels was 0.99999999999999603 in Kazak group. Given these results, the panel is suitable for individual identifications in the studied Kazak and these reference populations.

Keywords: Chinese populations; forensic investigation; InDels; individual identification; Kazak.

Abbreviations:

Analysis of molecular variance (AMOVA); Chinese Dai in Xishuangbanna (CDX); Han Chinese in Beijing (CHB); Southern Han Chinese (CHS); cumulative match probabilities (CMP); cumulative power of discrimination (CPD); cumulative power of exclusion (CPE); pairwise fixation index (F_{st}); expected heterozygosity (H_e); observed heterozygosity (H_o); Hardy-Weinberg equilibrium (HWE); Japanese in Tokyo (JPT); Kinh in Ho Chi Minh City, Vietnam (KHV); Linkage disequilibrium (LD), match probability (MP); principal component analysis (PCA); power of discrimination (PD); power of exclusion (PE); polymorphism information content (PIC)

Color online: See article online to view Figs. 1-6 in color.

Additional supporting information may be found in the online version of this article at the publisher's web-site.

1. Introduction

Short tandem repeats (STRs) are extensively used in forensic investigations because they possess the advantages of high informativeness and polymorphisms [1]. With the widespread applications of STR markers in forensic casework, some shortcomings of STRs gradually emerged. For instances, relatively high mutation rates of STRs [2] may provide incorrect predictions in challenging paternity tests; large amplicon sizes of STRs [3] limit their utility in dated or high degraded samples; some noise peaks like stutter may produce adverse effects on the allele determinations of STR loci. Thus, it is necessary for forensic researchers to search for novel genetic markers which can overcome these defects.

Single nucleotide polymorphism (SNP) as the promising genetic marker, brings new choices in forensic casework owing to their abundant content in human genome and low mutation rates [4]. However, unlike STRs, SNPs are sequence variations which require complex chemistry and operation approaches to detect the variations. Consequently, it is arduous for the popularization and application of SNPs in forensic practices.

Insertion/deletion polymorphisms are the insertion/deletion variations of nucleotide fragments with different lengths [5]. Similar to STRs, InDels are length polymorphisms which are compatible with extant typing equipment. Besides, InDels don't have stutter peaks [6]. The low mutation rates of InDels make them more preponderant in some cases involved in STR mutations [4]. More importantly, small amplicon sizes of InDels enhance discrimination efficiencies in some degraded samples so that it can provide more valuable information [7]. To date, some InDel panels used for forensic applications have been developed. For examples, Bastos-Rodrigues et al. [8, 9] constructed a panel of 40 InDels for individual identifications in European populations; Pereira et al. [10] provided a novel multiplex system of 38 InDels for human identifications; LaRue et al. [11] assessed genetic distributions of 114 InDels in Caucasian, African American, Hispanic and Asian populations and chose 49 InDels from the 114 InDels for human identifications; Oka et al. [12] provided 37 InDels for human identifications in Japanese populations. Besides, amplicon sizes of InDels are less than 180bp for these published panels, which are beneficial to analyze degraded samples. At present, the InDel kit commonly used in China is the Investigator DIPplex kit which is not designed for East Asian populations, especially for the Chinese populations. Although the kit performs well for forensic individual identifications in some populations from China [13-15], we developed a novel multiplex InDel system to obtain better discrimination powers in Chinese populations. Moreover, forensic efficiencies of the system were further evaluated in Chinese Kazak group.

2. Materials and methods

2.1 Selection of InDel loci for individual identification in Chinese populations

Referring to previous criteria for InDel selections [10-12], the InDel loci used for individual identifications in Chinese populations were selected from dbSNP database according to the revised criteria as below: 1. autosomal biallelic variations; 2. allelic frequencies of InDels ranged from 0.4000 to 0.6000 in Chinese populations; 3. located on no-coding regions; 4. allele length ranged from 2 to 20 bp; 5. no deviations from Hardy-Weinberg equilibrium (HWE) in Chinese populations. Finally, thirty-five InDel loci were used to develop the multiplex system.

2.2 Sample collection

Bloodstain samples of 510 unrelated healthy Kazak individuals from Xinjiang Uygur Autonomous Region were collected after receiving their written informed content. Training set including three Chinese populations (Chinese Dai in Xishuangbanna, CDX; Han Chinese in Beijing, CHB; Southern Han Chinese, CHS), Japanese in Tokyo (JPT) and Kinh in Ho Chi Minh City, Vietnam (KHV) populations was utilized to initially evaluate forensic values of selected InDels, and genetic data of these populations were downloaded from 1000 Genomes Project Phase 3 [16]. Table 1 listed these reference population information and their corresponding sample sizes. The experiment was conducted in line with the guidelines of humane and ethical research of Xi'an Jiaotong University Health Science Center, China and warranted by the ethics committee of Xi'an Jiaotong University Health Science Center, China.

2.3 Primer designs of InDel loci and the construction of multiplex PCR panel

Specific primers of 35 InDel loci were designed by Primer Premier 5.0 online tool. Specificity, primer-dimers, hairpin structure and the ability to form stable duplexes of the designed primers were further assessed by Oligo software version 7.0. To achieve multiplex amplification of 35 InDels, these InDels were classified into four groups according to their amplicon sizes, and primers of the InDels were labeled by each of four fluorochromes (FAM, HEX, TAMRA and ROX).

2.4 InDel amplification and detection

Bloodstain samples can be directly amplified on the corresponding equipment because the panel is developed into one direct amplification kit. Detailed procedures were described as below. PCR cocktail consisting of 4 μ l Nuclease-Free Water, 5 μ l 2 \times Master mix (Microread Genetics, Beijing,

China) and 1 µl primer mix was added to one well of the reaction plate which contained one 1.2 mm bloodstain disc. PCR was performed on GeneAmp PCR System 9700 Thermal Cycler (Applied Biosystems, Foster City, CA, USA) based on the following parameters: initial denaturation at 95 °C for 5 min; then 35 cycles of 94 °C for 45 s, 56 °C for 1 min, 72 °C for 1 min; the final extension at 60 °C for 60 min. Next, 1 µl PCR product was added to the cocktail of 0.5 µl Size Standard Org500 (Microread Genetics, Beijing, China) and 8.5 µl Hi-Di formamide. The mixture was denatured at 95 °C for 3 min and then immediately chilled on ice for 3 min. Finally, the sample was detected on 3500xL Genetic Analyzer (Applied Biosystems, Foster City, CA, USA). Genotyping data of InDels was determined on GeneMapper ID software version 3.2 (Applied Biosystems, Foster City, CA, USA). Deionized water and control DNA M308 (5 ng/µl) were used as negative and positive controls, respectively. Allele typing of control DNA M308 was shown in Fig. 1.

2.5 Statistical analysis

The HWE tests, observed heterozygosity (H_o) and expected heterozygosity (H_e) of 35 InDels in the Kazak group and five reference populations were estimated by Haploview software version 4.2 [17]. P -values of Linkage disequilibrium (LD) analyses of these InDels in five reference populations and Kazak group were calculated by Genepop software version 4.0 [18]. Allelic frequencies and forensic parameters including power of discrimination (PD), polymorphism information content (PIC), match probability (MP) and power of exclusion (PE) of 35 InDels in Kazak group and other training populations were calculated by PowerStats software version 1.2 (Promega, Madison, WI, USA). Pairwise fixation index (F_{st}) values of these reference populations were calculated by Genepop software version 4.0 [18]. Analysis of molecular variance (AMOVA) of these reference populations was conducted by Arlequin software version 3.5 [19] under the condition of 1000 permutations. To explore genetic relationships between Kazak group and these reference populations, principal component analysis (PCA) and phylogenetic reconstruction of these populations were performed by MVSP software version 3.1 and MEGA software version 6.0 [20], respectively.

3. Results and Discussion

3.1 General information of 35 InDels

As presented in Table 2, selected 35 InDels were distributed over 20 autosomes and their insertion or deletion fragment lengths ranged from 2-14 bp. Besides, amplicon sizes of 35 InDels ranged from 104 to 304 bp. Compared to amplicon sizes of published InDels [10-12, 21], some out of the 35 InDels possessed relatively high amplicon sizes, which might lead to the drop-out of alleles of these

InDels in degraded samples. Nevertheless, further validation evaluations of forensic efficiencies like the ability to analyze degraded samples, sensitivity, specificity, stability and so on should be performed in the latter work. Physical distances of pairwise InDels located on the same chromosomes were presented in Supporting Information Table 1. Results showed that all pairwise InDels on the same chromosomes were apart more than 10 Mb except for the two pairs (rs10629077 & rs10609615 and rs2307433 & rs3054057).

3.2 Genetic distributions of 35 InDels in five reference populations from East Asia

P-values of HWE tests of 35 InDels in five East Asian populations were given in Supporting Information Table 2-6, which revealed all InDels loci conformed to HWE in these reference populations after Bonferroni correction ($0.05/35 = 0.0014$). Even though physical distances of two pairs (rs10629077 & rs10609615 and rs2307433 & rs3054057) on the same chromosomes were less than 10 Mb, *P*-values of LD analyses for the two pairs were not statistically significant in these reference populations (data not shown).

Allelic frequencies and forensic relevant parameters of 35 InDels in five reference populations were also given in Supporting Information Table 2-6. Results indicated that insertion allelic frequencies of these loci were distributed between 0.2-0.8 except for rs3054057 locus. Furthermore, similar frequency distributions of 35 InDels could be discerned in these populations. Pairwise *Fst* values of five populations were shown in Fig. 2. *Fst* can evaluate genetic differentiations of different populations which is utilized to infer population structure and human migration [22]. A previous study [23] points out that small *Fst* values between populations meant low genetic differentiations and vice versa. The findings in Fig. 2 demonstrated low genetic differentiations ($Fst < 0.0016$) among these populations. AMOVA of these populations was performed based on genetic data of 35 InDels (Table 3). Result of AMOVA reflected most variations were from within population (0.9931), while few variations (0.0069) were from among populations. To sum up, these results revealed similar genetic distributions of 35 InDels in these reference populations. Mean values of PIC, *Ho* and *He* of 35 InDels in five reference populations ranged from 0.3568 (JPT) to 0.3603 (CHB), 0.4681 (JPT and CDX) to 0.4854 (CHB), 0.4712 (JPT) to 0.4769 (CHB), respectively. By comparing with the results of 30 InDels in Beijing Han [24] and Guangdong Han [25] populations, we found these 35 InDels showed higher diversities in Chinese Han populations. Some researchers [26] suggest that genetic markers with PIC more than 0.25 can provide reasonably genetic information. PIC values of 35 InDels in five reference populations were more than 0.25 except for rs3054057 locus, demonstrating the majority of 35 InDels possessed relatively reasonable genetic information in these reference populations.

Cumulative match probabilities (CMP) and power of discrimination (CPD) values of selected 35 InDels and other published InDel panels were presented in Table 4. Results revealed that CMP values of 35 InDels in East Asian populations was similar to those of different InDel panels in Asian populations. The InDel panel whose CMP ranges from 10^{-15} to 10^{-14} can meet requirements for forensic human identification [10]. Therefore, these InDels including developed 35 InDels can be considered valuable tools for human identification. However, since cumulative power of exclusion (CPE) values of 35 InDels was far less than the results of some common STR kits [27, 28], these InDels can provide supplementary information in paternity testing involved with STR mutations.

3.3 Forensic efficiency evaluations of 35 InDels in the studied Kazak group

HWE results of 35 InDels in Kazak group were shown in Supporting Information Table 7. After Bonferroni corrections ($P = 0.05/35 = 0.0014$), no deviations from HWE were observed. P -values for LD analysis of pairwise InDels in Kazak group were presented in Supporting Information Table 1, which demonstrated all pairwise InDels conformed to linkage equilibrium in Kazak group after Bonferroni corrections ($P = 0.05/595 = 0.000084$).

Allelic frequencies, H_o and H_e of 35 InDels were visually presented in Fig. 3. Results indicated the insertion allelic frequencies ranged from 0.3206 (rs3831219) to 0.8245 (rs3054057). Compared with the results of 30 InDels in Kazak group [29], it should be pointed out that more homogeneous frequency distributions of different alleles were seen for the majority of the selected InDels (Fig. 3). For H_o and H_e values of 35 InDels, they ranged from 0.2800 to 0.5200 and 0.2890 to 0.5000, respectively. Moreover, some loci whose frequency differences of different alleles were low tended to have high H_e values. As an example, H_e value at rs2307433 locus whose the insertion and deletion frequencies were 0.5157 and 0.4843 was 0.5000. Therefore, selections of genetic markers for forensic individual identification should give preference to those markers who show uniform allele distributions in populations, especially for biallelic markers. Forensic parameters (PIC, PD and PE) of 35 InDels in Kazak group were shown in Fig. 4. PD values of all loci were more than 0.4500; and CPD values of 35 InDels was 0.99999999999999603, showing that these 35 InDels could perform well for individual identification in Kazak group. However, PE values of 35 InDels ranged from 0.0558 to 0.2052 with the mean values of 0.1639; and CPE values of 35 InDels in Kazak group was 0.9981, revealing these loci were not good in parentage testing. More polymorphic InDel loci should be incorporated into the developed multiplex panel to enhance its application values in paternity testing. Besides, microhaplotype, the novel genetic marker, is defined by two or more genetic markers within a short distance (< 300 bp), which is of great value in forensic genetics [30]. Consequently, we

can search for informative SNPs/InDels within neighboring regions of the selected 35 InDels and those previously reported InDels [8, 10-12] to further enhance the discrimination power of these biallelic InDel loci.

3.4 Genetic relationship analyses among Kazak group and five reference populations

We assessed the genetic relationships of the studied Kazak group and five reference populations. As shown in Fig. 5, two apparent individual clusters could be observed from PCA plot: the individuals from five reference populations clustered together located in the left part; the studied Kazak individuals situated in the right part. Similar results could be discerned from the phylogenetic analysis (Fig. 6), revealing Kazak group located on one branch and the remaining populations positioned on the other. Kazak group is one old ethnic group in China, and its forefathers once lived with nomadic or semi-nomadic Uygurs, Mongols and Naimans in the mid-six century. In the 14th century, some herdsmen went eastward and mixed with Ozbeks and settled Mongols. With the growth of the population, they continually extended their pastures and eventually evolved into the Kazak group. (<http://www.china.org.cn/english/features/EthnicGroups/136924.htm>). Study on mitochondrial DNA variations revealed Altaian Kazakhs originated from a common gene pool which had a variety of West and East Eurasian maternal lineages [31]. Previous research [29, 32] based on 30 autosomal InDels and 19 X-STR loci found that Kazak group had relatively close ties with Uygur group who had mixed genetic components of European and East Asian populations. In this study, we found that the studied Kazak group showed genetic differentiations with these five East Asian populations. As more genetic data of 35 InDels in Chinese populations being published, especially for populations from northwest China, we would better understand the genetic background of Kazak group.

4. Concluding remarks

In summary, the present research developed one novel multiplex InDel panel for individual identifications in Chinese populations. Evaluations of genetic distributions for 35 InDels in five reference populations demonstrated that most loci could provide reasonable genetic information in these reference populations. Analyses of genetic diversities and forensic parameters of these loci in the studied Kazak group further supported that the panel could be used for forensic individual identification. Research on validation evaluation of the panel in forensic casework should be performed in the future. Besides, forensic value analysis of the panel in other populations in China should also be conducted.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (81525015, 81772031) and Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme (GDUPS, 2017).

The authors have declared no conflict of interest.

5. References

- [1] Edwards, A., Civitello, A., Hammond, H. A., Caskey, C. T., *Am. J. Hum. Genet.* 1991, *49*, 746-756.
- [2] Brinkmann, B., Klitschar, M., Neuhuber, F., Huhne, J., Rolf, B., *Am. J. Hum. Genet.* 1998, *62*, 1408-1415.
- [3] Golenberg, E. M., Bickel, A., Weihs, P., *Nucleic Acids Res.* 1996, *24*, 5026-5033.
- [4] Nachman, M. W., Crowell, S. L., *Genetics* 2000, *156*, 297-304.
- [5] Mills, R. E., Pittard, W. S., Mullaney, J. M., Farooq, U., Creasy, T. H., Mahurkar, A. A., Kemeza, D. M., Strassler, D. S., Ponting, C. P., Webber, C., Devine, S. E., *Genome Res.* 2011, *21*, 830-839.
- [6] Wendt, F. R., Warshauer, D. H., Zeng, X., Churchill, J. D., Novroski, N. M. M., Song, B., King, J. L., LaRue, B. L., Budowle, B., *Forensic Sci. Int. Genet.* 2016, *25*, 198-209.
- [7] Romanini, C., Catelli, M. L., Borosky, A., Pereira, R., Romero, M., Salado Puerto, M., Phillips, C., Fondevila, M., Freire, A., Santos, C., Carracedo, A., Lareu, M. V., Gusmao, L., Vullo, C. M., *Forensic Sci. Int. Genet.* 2012, *6*, 469-476.
- [8] Bastos-Rodrigues, L., Pimenta, J. R., Pena, S. D., *Ann. Hum. Genet.* 2006, *70*, 658-665.
- [9] Pimenta, J. R., Pena, S. D., *Genet. Mol. Res.* 2010, *9*, 601-607.
- [10] Pereira, R., Phillips, C., Alves, C., Amorim, A., Carracedo, A., Gusmao, L., *Electrophoresis* 2009, *30*, 3682-3690.
- [11] LaRue, B. L., Lagace, R., Chang, C. W., Holt, A., Hennessy, L., Ge, J., King, J. L., Chakraborty, R., Budowle, B., *Legal Med.* 2014, *16*, 26-32.

- [12] Oka, K., Asari, M., Omura, T., Yoshida, M., Maseda, C., Yajima, D., Matsubara, K., Shiono, H., Matsuda, M., Shimizu, K., *Mol. Cell Probes* 2014, 28, 13-18.
- [13] Meng, H. T., Zhang, Y. D., Shen, C. M., Yuan, G. L., Yang, C. H., Jin, R., Yan, J. W., Wang, H. D., Liu, W. J., Jing, H., Zhu, B. F., *Sci. Rep.* 2015, 5.
- [14] Shen, C. M., Zhu, B. F., Yao, T. H., Li, Z. D., Zhang, Y. D., Yan, J. W., Wang, B., Bie, X. H., Tai, F. D., *Sci. Rep.* 2016, 6.
- [15] Zhang, Y. D., Shen, C. M., Jin, R., Li, Y. N., Wang, B., Ma, L. X., Meng, H. T., Yan, J. W., Wang, H. D., Yang, Z. L., Zhu, B. F., *Electrophoresis* 2015, 36, 1196-1201.
- [16] Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., Abecasis, G. R., *Nature* 2015, 526, 68-74.
- [17] Barrett, J. C., Fry, B., Maller, J., Daly, M. J., *Bioinformatics* 2005, 21, 263-265.
- [18] Rousset, F., *Mol. Ecol. Resour.* 2008, 8, 103-106.
- [19] Excoffier, L., Lischer, H. E., *Mol. Ecol. Resour.* 2010, 10, 564-567.
- [20] Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., *Mol. Biol. Evol.* 2013, 30, 2725-2729.
- [21] Santos, V. R., Pena, H. B., Pena, S. D., *Genet. Mol. Res.* 2015, 14, 2947-2952.
- [22] Weir, B. S., Cockerham, C. C., *Evolution* 1984, 38, 1358-1370.
- [23] Holsinger, K. E., Weir, B. S., *Nat. Rev. Genet.* 2009, 10, 639-650.
- [24] Shi, M., Liu, Y., Bai, R., Jiang, L., Lv, X., Ma, S., *Int. J. Legal Med.* 2015, 129, 53-56.
- [25] Hong, L., Wang, X. G., Liu, S. J., Zhang, Y. M., Xue-Ling, O. U., Chen, Y., Chen, W. H., Sun, H. Y., *J. Sun Yat-sen Univ. (Med. Sci.)* 2013, 34, 299-304.
- [26] Botstein, D., White, R. L., Skolnick, M., Davis, R. W., *Am. J. Hum. Genet.* 1980, 32, 314-331.
- [27] Wang, H. D., Wu, D., Feng, Z. Q., Jing, Z. A., Li, T., Guo, Q. N., Zhang, X. P., Hou, Q. F., Guo, L. J., Kang, B., Zhang, H., Zhu, B. F., Liao, S. X., *Electrophoresis* 2014, 35, 1509-1514.
- [28] Chen, L., Lu, H. J., Qiu, P. M., Yang, X. Y., Liu, C., *Legal Med.* 2015, 17, 489-492.

[29] Kong, T. T., Chen, Y. H., Guo, Y. X., Wei, Y. Y., Jin, X. Y., Xie, T., Mu, Y. L., Dong, Q., Wen, S. Q., Zhou, B. Y., Zhang, L., Shen, C. M., Zhu, B. F., *Oncotarget* 2017, 8, 56651-56658.

[30] Oldoni, F., Kidd, K. K., Podini, D., *Forensic Sci. Int. Genet.* 2018, 38, 54-69.

[31] Gokcumen, O., Dulik, M. C., Pai, A. A., Zhadanov, S. I., Rubinstein, S., Osipova, L. P., Andreenkov, O. V., Tabikhanova, L. E., Gubina, M. A., Labuda, D., Schurr, T. G., *Am. J. Phys. Anthropol.* 2008, 136, 278-293.

[32] Liu, Y. S., Meng, H. T., Mei, T., Zhang, L. P., Chen, J. G., Zhang, Y. D., Chen, J., Guo, Y. X., Dong, Q., Yan, J. W., Zhu, B. F., *Gene* 2017, 600, 64-69.

Figure and Table Captions

Figure 1. Genetic profiles of 35 InDels in control DNA M308.

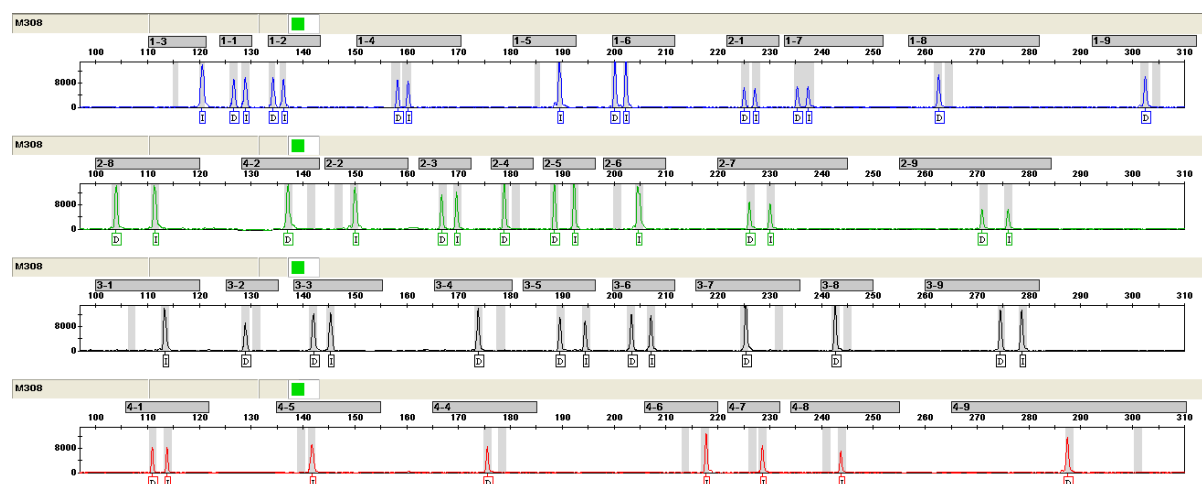


Figure 2. Pairwise F_{st} values of five reference populations. Different colors represent for the F_{st} values between certain population and the other populations: orange color for the F_{st} values between CHB population and other populations; green color for the F_{st} values between CHS population and other populations; purple color for the F_{st} values between JPT population and other populations; cyan color for the F_{st} values between KHV population and other populations.

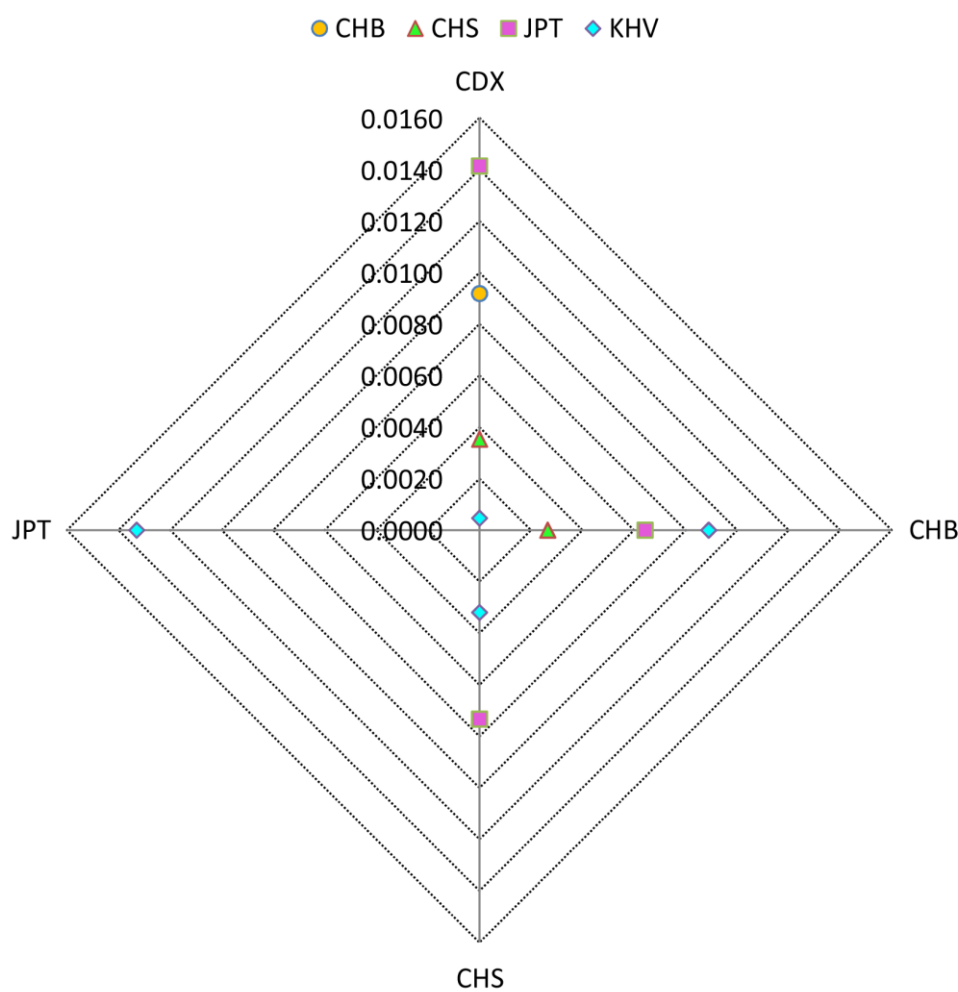


Figure 3. Allelic frequencies, observed heterozygosities and expected heterozygosities of 35 InDels in the studied Kazak group. The numerics in bar chart are frequency values of insertion and deletion alleles. The H_o and H_e in graph represent for observed heterozygosity and expected heterozygosity, respectively.

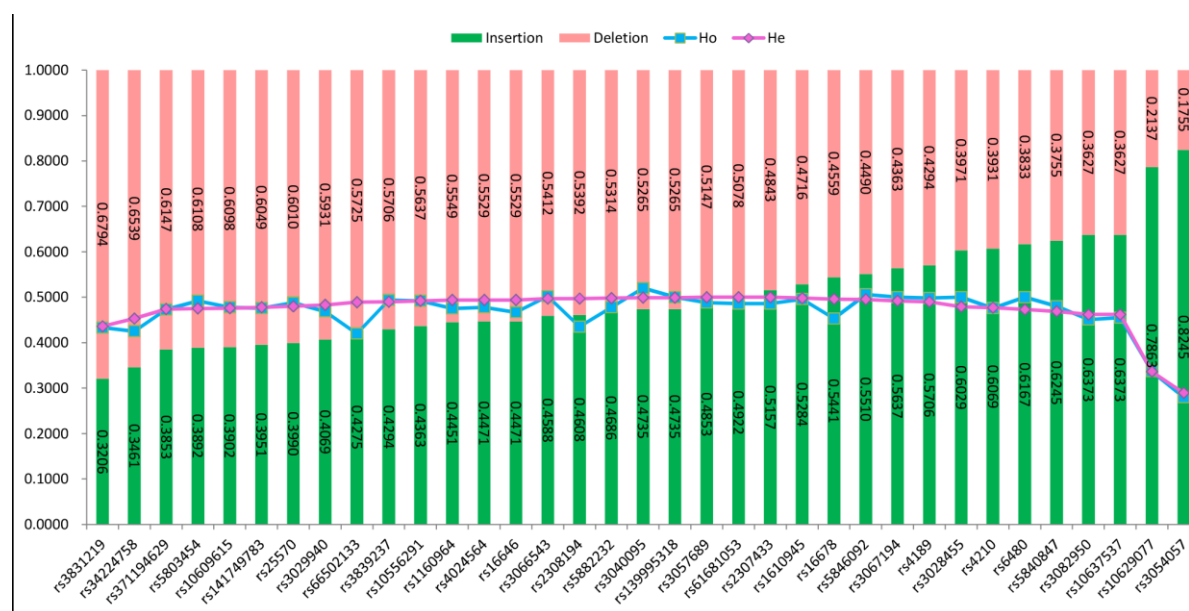


Figure 4. Forensic parameters of 35 InDels in Kazak group. Abbreviations including PD, PE and PIC denote power of discrimination, power of exclusion and polymorphism information content.

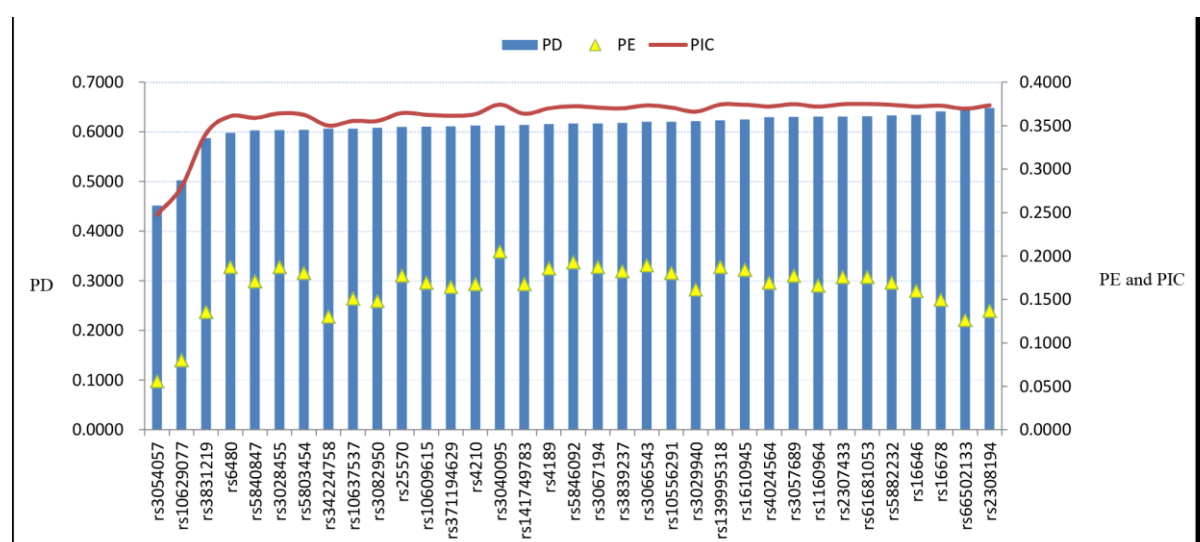


Figure 5. Principal component analysis of the studied Kazak group and other reference populations. Different colors denote different language families: red for Altaic; purple for Sino-Tibetan; cyan for Austronesian.

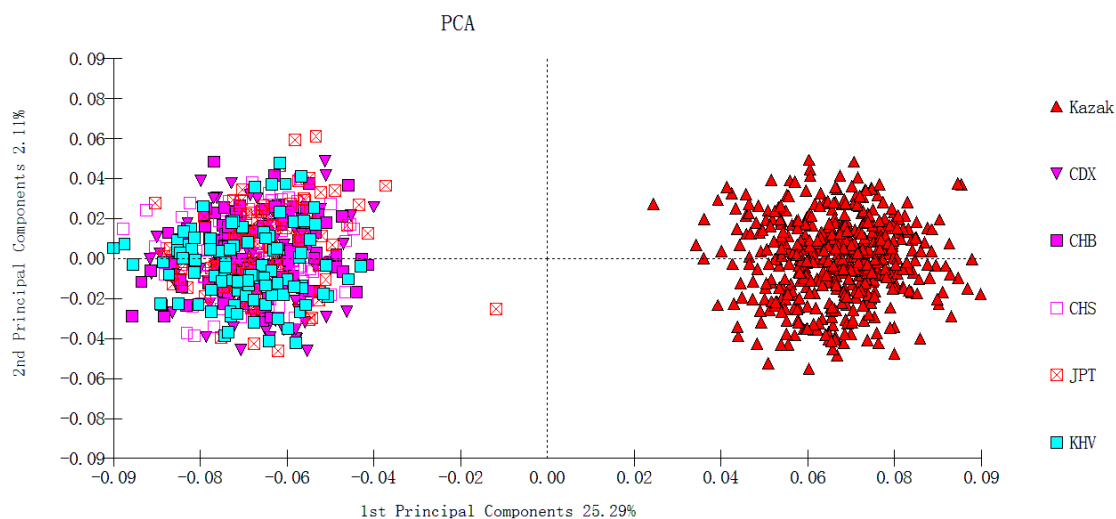


Figure 6. Neighbor-joining tree of Kazak and other compared populations. Different colors denote different language families: red for Altaic; purple for Sino-Tibetan; cyan for Austronesian.

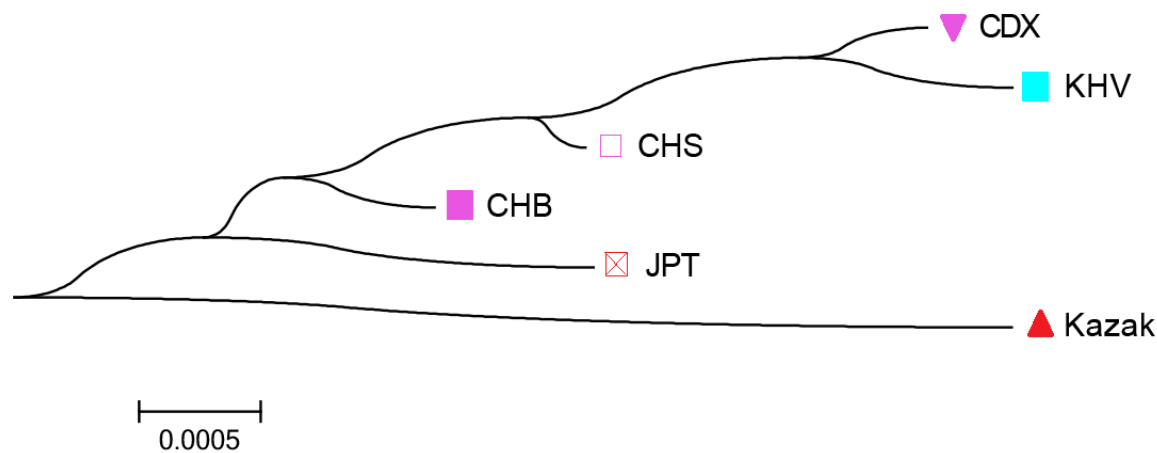


Table 1: Population information used in this study.			
Populations	Abbreviations	Sample sizes	Sources
Chinese Dai in Xishuangbanna	CDX	93	1000 Genomes Project Phase 3
Han Chinese in Beijing	CHB	103	1000 Genomes Project Phase 3
Southern Han Chinese	CHS	105	1000 Genomes Project Phase 3
Kinh in Ho Chi Minh City, Vietnam	KHV	99	1000 Genomes Project Phase 3
Japanese in Tokyo	JPT	104	1000 Genomes Project Phase 3
Kazak	-	510	Our lab

Table 2: General information of 35 InDels.

ID	Rs numbers	Fluorochromes	Alleles ^a	Chromosomes ^a	Locations ^a	Amplicon sizes (D/I)
1-3	rs3028455	FAM	-/TAAGT	1	22291466 2	115/120
1-1	rs2308194	FAM	-/TT	1	29343295	126/129
1-2	rs3067194	FAM	-/AC	1	17756224 9	134/136
1-5	rs5846092	FAM	-/CAAC	3	1143035	185/189
1-8	rs4024564	FAM	-/AC	3	14363727 0	262/264

1-9	rs3040095	FAM	-/TG	3	76783421	302/304
1-6	rs3082950	FAM	-/AA	4	21439760	200/202
1-7	rs4210	FAM	-/AC	4	88726648	235/237
2-6	rs3066543	HEX	-/CAGA	4	128139878	201/205
2-1	rs1160964	FAM	-/AA	5	16419432	225/227
2-8	rs3839237	HEX	-/CCTTTGGG	5	172149559	104/111
2-2	rs1610945	HEX	-/TGT	5	96684978	147/150
2-3	rs16678	HEX	-/TCT	6	10793065	167/170
2-4	rs25570	HEX	-/AA	6	70290638	179/181
2-7	rs3029940	HEX	-/CTTA	6	130374031	226/230
4-9	rs371194629	ROX	-/ATTTGTTTCATGCCT	6	29830804	288/301
2-5	rs16646	HEX	-/GAAA	7	103760897	188/192
2-9	rs5882232	HEX	-/TAAAG	7	9775467	271/276

3- 9	rs1063753 7	TAMRA	-/TCTT	8	84290971	275/279
3- 3	rs3831219	TAMRA	-/GAG	9	13175717	142/145
3- 4	rs3422475 8	TAMRA	-/CCAC	9	13072625 4	174/178
3- 1	rs3057689	TAMRA	-/TGGTGGA	10	30892157	107/113
3- 7	rs6168105 3	TAMRA	-/AGGCCTA	11	19852769	225/232
3- 8	rs1055629 1	TAMRA	-/TA	12	96096478	243/245
3- 5	rs5803454	TAMRA	-/ATAAC	13	49283182	190/194
3- 2	rs6650213 3	TAMRA	-/TA	14	36164592	129/131
4- 2	rs2307433	HEX	-/GTAG	15	89321085	137/141
3- 6	rs3054057	TAMRA	-/AACA	15	85467307	203/207
4- 4	rs1417497 83	ROX	-/AG	17	12716364	176/178
4- 6	rs4189	ROX	-/ACTT	18	80157550	214/217
4- 5	rs1399953 18	ROX	-/TT	19	43917495	140/142

4- 8	rs5840847	ROX	-/TCA	20	19489778	241/244
1- 4	rs1062907 7	FAM	-/AT	21	30000019	158/160
4- 1	rs1060961 5	ROX	-/CCC	21	29083308	111/114
4- 7	rs6480	ROX	-/CA	22	35314217	227/229

Note: ^a General information of InDels is presented according to dbSNP build 151. D and I denote deletion and insertion alleles of InDel loci, respectively.

Table 3: Analysis of molecular variance of five reference populations from East Asia based on 35 InDels

Source of variation	Degrees of Freedom	Sum of squares	Variance components	Percentage of variation
Among populations	4	80.0640	0.0580	0.0069
Within populations	1003	8350.3250	8.3254	0.9931
Total	1007	8430.3890	8.3834	1.0000

Table 4. Comparisons of cumulative match probabilities and power of discrimination of 35 InDels and those published InDel panels

Populations	Markers (References)	CMP	CPD
CDX ^a	35 InDels (this study)	5.866×10^{-15}	0.999 999 999 999 994134
CHB ^a	35 InDels (this study)	9.233×10^{-15}	0.999 999 999 999 990767
CHS ^a	35 InDels (this study)	6.328×10^{-15}	0.999 999 999 999 993672
JPT ^a	35 InDels (this study)	7.757×10^{-15}	0.999 999 999 999 992243
KHV ^a	35 InDels (this study)	1.410×10^{-14}	0.999 999 999 999 9859
Japanese	37 InDels (Oka et al. [12])	2.127×10^{-15}	0.999 999 999 999 997873
EAS ^b	38 InDels (Pereira et al. [10])	1.660×10^{-14}	0.999 999 999 999 9834
White Brazilians	40 InDels (Pimenta et al.[9])	3.480×10^{-17}	0.999 999 999 999 999 9652
Asian	49 InDels (LaRue et al. [11])	2.300×10^{-19}	0.999 999 999 999 999 99977

^a Genetic data of 35 InDels in these populations are downloaded from 1000 Genomes Project Phase 3. ^b

EAS includes Macanese and Taiwanese.