Population genetics and forensic efficiency of 30 InDel markers in four Chinese ethnic groups residing in Sichuan

Fei Wang*, Guanglin He†, Zheng Wang, Mengge Wang, Jing Liu, Xing Zou, Shouyu Wang, Mengyuan Song, Ziwei Ye, Mingkun Xie and Yiping Hou

Institute of Forensic Medicine, West China School of Basic Medical Sciences & Forensic Medicine, Sichuan University, Chengdu, China

ABSTRACT

Sichuan Province is located at the transitional junction regions of the Qinghai-Tibet Plateau and the low-altitude plains. It also serves as the corridor of Sino-Tibetan-speaking population migration and expansion since neolithic expansion of Proto-Tibeto-Burman populations from Middle/Upper Yellow River during Majiayao period (3300–2000 BC). However, the population structure and the corresponding genetic diversity of forensic-related markers in this region remain unclear. Thus, we genotyped 30 insertion-deletion (InDel) markers in 444 samples from four ethnic groups (Han, Tibetan, Hui and Yi) from Sichuan Province using the Investigator® DIPplex kit to explore the characteristics of population genetics and forensic genetic focuses. All the loci were found to be in Hardy-Weinberg Equilibrium (HWE) after applying a Bonferroni correction and no pairwise loci showed prominent linkage disequilibrium. The combined matching probability (CMP) and the combined power of discrimination (CPD) are larger than 1.8089 \times 10^{-11} and 0.9999999995, respectively. Principal component analysis, multi-dimensional scaling plots and Neighbour-Joining tree among 65 worldwide populations indicated that Sichuan Hui and Han are genetically close to Hmong-Mien and Tai-Kadai-speaking populations, and Sichuan Tibetan and Yi bear a strong genetic affinity with Tibeto-Burman-speaking populations. The model-based genetic structure further supports the genetic affinity between the studied populations and linguistically close populations.

KEY POINTS

- Forensic parameters of 30 insertion-deletions (InDels) in 444 individuals from four populations are reported, which showed abundant genetic affinity and diversity among populations and high value in personal identification.
- Genetic similarities existed between the studied populations and ethnically, linguistically close populations.
- Sichuan Hui and Han are genetically close to Hmong-Mien and Tai-Kadai-speaking populations.
- Sichuan Tibetan and Yi bear a strong genetic affinity with Tibeto-Burman-speaking populations.

Introduction

Insertion-deletion (InDel) polymorphisms has been a promising and powerful supplementary tool in forensic personal identification cases and kinship testing, since they bear the advantages of lower mutation rate, smaller amplicons and absence of stutter peaks compared with traditional gold standard short tandem repeats (STRs) [1–5].

Lying in Southwest China, backland of the Eurasian continent, Sichuan Province owns a complex topography spanning Qinghai-Tibet Plateau, Hengduan Mountains, Yunnan-Guizhou Plateau, Qinba Mountains and Sichuan Basin. Evidence of both archaeology and linguistics subsequently suggested that these regions play an important role in the formation of modern Sino-Tibetan-speaking populations, especially for Tibeto-Burman-speaking populations spreading southwards and westwards from the middle and upper basins of Yellow River during Yangshao period (about 7 000–5 000 years BP) with the millet cultivation expansion [6, 7].

Nowadays, the permanent population of Sichuan has exceeded 80,000,000 according to Sichuan Population Statistics Bulletin 2018 (http://tjj.sc.gov.cn/tjxx/zxfb/201903/t20190319_277119.html), which is nearly equivalent to Germany’s population. As a multi-ethnic province, Sichuan is the biggest agglomeration for the Yi ethnic group, the second-largest community of Tibetan nationality and residence for large number of Hui group. However,
Han group is still the largest population here (93.53%). Over the past hundreds of years, the ethnic groups have been the largest immigrant population inflow and accounted for an indispensable part in the construction of genetic variability in the province. In this study, we use the Investigator® DIPplex kit (Qiagen, Hilden, Germany) to investigate population genetics and forensic efficiency in these four Sichuan ethnic groups.

Materials and methods

Sample collection

Bloodstains were collected from 444 healthy unrelated individuals from Sichuan Province (155 Chengdu Hans, 132 Liangshan Yis, 119 Tibetans of Ganzi and Aba, and 38 Chengdu Huis) after obtaining participants’ written informed consents with the approval of the Ethics Committee of the Institute of Forensic Medicine, Sichuan University (K2015008).

DNA extraction, quantification, amplification and genotyping

Genomic DNA was extracted with the DNA Blood Mini Kit (Qiagen) and quantified using the NanoDrop-2000 Spectrophotometers (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer’s instructions. InDels amplification was carried out using the Investigator® DIPplex PCR Amplification Kit (Qiagen) following the manufacturer’s protocols and analyzed with Applied Biosystems 3130 Genetic Analyzer and Gene Mapper v3.2 (Thermo Fisher).

Quality control

Control DNA 9948 (Qiagen) and ddH₂O (Qiagen) were amplified and genotyped along the population samples as the positive and negative control, respectively. All experiments were carried out in our laboratory which has been accredited by the China National Accreditation Service for Conformity Assessment (CNAS) and ISO 17025.

Reference database

To comprehensively explore the patterns of genetic affinity between our studied populations and worldwide reference populations, we merged our newly obtained population data with those of 61 worldwide populations. The final database used in the population comparison study included 6483 East Asian individuals from 38 populations [4, 5, 8–26], 494 South Asians from three populations [27–29], 300 Vietnamese of Southeast Asia [30], 194 West Asians from two populations [31], 621 African individuals from five populations [27, 30, 32], 984 European individuals from eight populations [4, 31–35] and 1071 South Americans from eight populations [36, 38]. The detailed geographical regions and population sizes of these included populations are submitted in Supplementary Table S1.

Analysis of data

Allele frequencies and forensic efficiency parameters were calculated using online software STRAF (http://cmpg.unibe.ch/shiny/STRAF/) [39]. The heat map was carried out by R Statistical Software v3.0.2 (https://www.r-project.org/). Multi-dimensional scaling analysis (MDS) was performed by SPSS software v25.0 (IBM Corporation, Armonk, NY, USA). Principal component analysis (PCA) was conducted using the MultiVariate Statistical Package (MVSP, Provalis Research, Montreal, Quebec, Canada). Neighbour-joining tree (N-J tree) was carried out by the Molecular Evolutionary Genetics Analysis (MEGA, https://www.megasoftware.net/) on basis of the Nei’s standard genetic distances (Rₛ) using the Phylogeny Inference Package (PHYLIP, http://evolution.genetics.washington.edu/phylip.html). The genetic structure composition analysis was performed by the STRUCTURE programme v2.3.4 (https://web.stanford.edu/group/ritchardlab/structure.html) [40].

Results and discussion

Forensic characteristics

All InDel markers were successfully amplified and genotyped in the 444 samples with no significant deviation from Hardy-Weinberg Equilibrium (HWE) and no significant linkage disequilibrium after performing a Bonferroni correction (P ≤ 0.0017) (see raw genotype data in Supplementary Table S2). Allele frequencies, forensic parameters, and efficiencies in the four Sichuan populations are shown in Supplementary Table S3. The highest power of discrimination (PD) loci were HLD88 (PD = 0.6519) in Sichuan Han, HLD40 (PD = 0.6294) in Sichuan Yi, HLD124 (PD = 0.6405) in Sichuan Tibetan and HLD88 (PD = 0.6607) in Sichuan Hui. The combined matching probability (CMP) and the combined power of discrimination (CPD) ranged from 1.8089 × 10⁻¹¹ to 4.6696 × 10⁻¹¹, 0.99999999995330 to 0.99999999998191, respectively. The highest and lowest combined powers of paternity exclusion (CPE) were 0.9919 in Sichuan Hui and 0.9861 in Sichuan Tibetan. These observed values indicated that the InDel system can serve as a powerful supplementary tool for existing STR system.
**Population genetic analyses**

To further explore the genetic similarities and differences between populations, 61 worldwide population data of 30 InDel markers were collected into the reference database mentioned above [4, 5, 8–38]. The heat map was built based on the insertion allele frequencies (DIP + ) of 30 InDels from four Sichuan populations and 61 previously published populations, which was shown in Supplementary Figure S1. On the left side of the heat map, 65 populations were separated into two clusters: Mexican populations and others. The second cluster was divided into two branches as African populations (Somali, Nigerian, Xhosa, Zulu) and others. The further separation happened between Asian, European and American roughly. In most Asian populations, which is contrasted with other populations, HLD111, HLD39 and HLD122 loci show prominent low insertion frequencies and HLD118, HLD99, HLD64 and HLD81 loci perform distinct high frequencies.

MDS, PCA and N-J tree were generated based on the Rst generated from PHYLIP v.3.6. (Supplementary Table S4). As is shown in MDS (Supplementary Figure S2), population stratification along the continental geographical boundaries is identifiable. Genetic similarities and differences were further dissected using PCA (Figure 1). Top three components could explain 84.254% of the total differences (PC1 = 57.019%, PC2 = 18.228%, PC3 = 9.007%). Yi and Tibetan from Sichuan are genetically close to Tibeto-Burman-speaking populations (Yunnan Yi, Tibetan from Tibet, Qinghai and Sichuan, Tujia from Hubei), while Hui and Han are localized close with Hmong-Mien and Tai-Kadai-speaking populations (Zhuang, Dong and Miao from Guangxi).

The evolutionary genetic relationship revealed by the N-J tree (Supplementary Figure S3) follows a pattern of ethnic, geographical and linguistic affinity. Asian populations are close to each other on the large scale, while the Sichuan Tibetan group remains close to other 10 Tibetan groups, and Sichuan Hui group stays next to Xinjiang Hui in the picture. Guangdong Han is considerably closer to the geographically close Tai-Kadai or Hmong-Mien-speaking populations (Dong, Miao, and Zhuang from Guangxi).

The STRUCTURE analysis was used to dissect the genetic structure of the four Sichuan populations we studied and other 32 populations on the basis of their raw genotypes, which were shown in Figure 2 [8, 10, 12–17, 19, 21–26, 30, 33]. The K represents the number of the predefined ancestry. Genetic affinity in linguistically close populations can be identified when K increases from 5 to 6. Four investigated populations share large number of ancestry components with neighbouring East Asian populations, which was in accordance with results of MDS, PCA and N-J tree. Furthermore, as we all know, the Silk Road, the earliest and most important channel of communication between eastern and western civilizations, also served as a precious opportunity for gene communication between the populations along the road. Xinjiang was the midpoint of the road. As a result, Turkish-speaking groups (Xinjiang Hui, Kazak, and Kyrgyz) bear both Asian and European ancestries and appear as the interim of genetic structure between populations from East Asia and Europe when K equals 6.

**Conclusion**

According to the obtained results, Han, Yi, Tibetan, and Hui group of Sichuan were scattered among the Asian groups in terms of genetic distance, evolutionary relationship and genetic structure. The least significant differences in the genetic relationship were found between studied populations and ethnically,

![Figure 1](image-url). Principal components analysis among 65 worldwide populations revealed by top three components (PC1 = 57.019%, PC2 = 18.228%, PC3 = 9.007%), which explains 84.254% of the total differences.
linguistically close populations. Furthermore, based on the acquired values, the Investigator DIPplex® PCR Amplification Kit has been proved to be extremely useful as a supplement of STR to apply in forensic case works in the four populations we studied.

Authors’ contributions
Fei Wang, Guanglin He, Zheng Wang, and Yiping Hou designed and conducted the experiment. Mengge Wang, Jing Liu and Xing Zou analysed data. Shouyu Wang, Mengyuan Song, Ziwei Ye and Mingkun Xie helped with background checking and picture drawing. All the authors discussed the results together. Fei Wang and Guanglin He wrote the manuscript. All authors contributed to the final text and approved it.

Compliance with ethical standards
All procedures performed in studies involving human participants were in accordance with the ethical standards of the Ethics Committee of the Institute of Forensic Medicine, Sichuan University (K2015008) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Written informed consents were obtained from all individual participants included in the study.

Disclosure statement
The authors declare that they have no conflict of interest.

Funding
This study was carried out with the grants from the National Natural Science Foundation of China [grant number 81571854], the Open Project of Key Laboratory of Forensic Genetics in Ministry of Public Security [grant number 2017FGKFKT01] and the Fundamental Research Funds for the Central Universities [grant number 20826041 A4408, YJ201651 and 2012017yjsy187].

References
[1] Weber JL, David D, Heil J, et al. Human diallelic insertion/deletion polymorphisms. Am J Hum Genet. 2002;71:854–862.
[2] Pereira R, Phillips C, Alves C, et al. A new multiplex for human identification using insertion/deletion polymorphisms. Electrophoresis. 2009;30:3682–3690.
[3] Fondevila M, Phillips C, Santos C, et al. Forensic performance of two insertion–deletion marker assays. Int J Legal Med. 2012;126:725–737.
[4] Pepinski W, Abreu-Glowacka M, Koralewska-Kordel M, et al. Population genetics of 30 INDELs in populations of Poland and Taiwan. Mol Biol Rep. 2013;40:4333–4338.
[5] Seong KM, Park JH, Hyun YS, et al. Population genetics of insertion–deletion polymorphisms in South Koreans using Investigator DIPplex kit. Forensic Sci Int-Gen. 2014;8:80–83.
[6] Zhang M, Yan S, Pan W, et al. Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic. Nature. 2019;569:112–115.
[7] Guedes JAD, Lu HL, Hein AM, et al. Early evidence for the use of wheat and barley as staple crops on the margins of the Tibetan Plateau. Proc Natl Acad Sci USA. 2015;112:5625–5630.

Figure 2. STRUCTURE analysis results of 36 worldwide populations on basis of their raw genotypes when the number of pre-defined ancestry (K) varies from 2 to 6 and the best K is 4.
F. WANG ET AL.

[23] Xie T, Guo YX, Chen L, et al. A set of autosomal polymorphisms of 30 InDel loci in Chinese ethnic population residing in Tibet. Zhonghua Yi Xue Yi Chuan Xue Za Zhi. 2015;32:410–414. Chinese.

[24] Yang CH, Yin CY, Shen CM, et al. Genetic polymorphism of 30 InDels in Fujian Han population [master’s thesis]. Fujian (China): Fujian Medical University; 2015.

[25] Zhang YD, Shen CM, Jin R, et al. Forensic evaluation and population genetic study of 30 insertion/deletion polymorphisms in a Chinese Yi group. Electrophoresis. 2015;36:1196–1201.

[26] Zhu BF, Lan Q, Guo YX, et al. Population genetic diversity and clustering analysis for Chinese Dongxdia group with 30 autosomal InDel loci simultaneously analyzed. Front Genet. 2018;9:279.

[27] Hefke G, Davison S, D’Amato ME. Forensic performance of Investigator DIPplex indels genotyping kit in native, immigrant, and admixed populations in South Africa. Electrophoresis. 2015;36:3018–3025.

[28] Shahzad M, Hussain M, Shafique M. Population genetic data of 30 insertion–deletion markers in Punjabi population of Pakistan. Int J Leg Med. 2020;134:511–512.

[29] Sharif A, Sudipta AD, Ismail H, et al. Genetic polymorphism of 30 InDel markers for forensic use in Bangladeshi population. Forensic Sci Int Genet Suppl Ser. 2013:e348–e349.

[30] Du WA, Peng ZY, Feng CL, et al. Forensic efficiency and genetic variation of 30 InDels in Vietnamese and Nigerian populations. Oncotarget. 2017;8:88934–88940.

[31] Tomas C, Poulsen L, Drobnik N, et al. Thirty autosomal insertion–deletion polymorphisms analyzed using the Investigator DIPplex Kit in populations from Iraq, Lithuania, Slovenia, and Turkey. Forensic Sci Int Genet. 2016;25:143–144.

[32] Neuvonen AM, Palo JU, Hedman M, et al. Discrimination power of Investigator DIPplex loci in Finnish and Somali populations. Forensic Sci Int Genet. 2012;6:E125–E102.

[33] Friis SL, Borsting C, Rockenbauer E, et al. Typing of 30 insertion/deletions in Danes using the first commercial indel kit—Mentype® DIPplex. Forensic Sci Int Genet. 2012;6:672–674.

[34] Kis Z, Zalan A, Volgyi A, et al. Genome deletion and insertion polymorphisms (DIPs) in the Hungarian population. Forensic Sci Int Genet. 2012;6:E125–E126.

[35] Martin P, Garcia O, Heinrichs B, et al. Population genetic data of 30 autosomal indels in Central Spain and the Basque Country populations. Forensic Sci Int Genet. 2013;7:27–E30.

[36] Martinez-Cortes G, Garcia-Aceves M, Favela-Mendoza AF, et al. Forensic parameters of the Investigator DIPplex kit (Qiagen) in six Mexican populations. Int J Leg Med. 2016;130:683–685.

[37] Palha T, Rodrigues EMR, Cavalcante GC, et al. Population genetic analysis of insertion–deletion polymorphisms in a Brazilian population using the Investigator DIPplex kit. Forensic Sci Int Genet. 2015;19:10–14.

[38] Saiz M, Andre F, Pisano N, et al. Allelic frequencies and statistical data from 30 INDEL loci in Uruguayan population. Forensic Sci Int Genet. 2014;9:E27–E29.

[39] Gouy A, Zieger M. STRAF—a convenient online tool for STR data evaluation in forensic genetics. Forensic Sci Int Genet. 2017;30:148–151.

[40] Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 2005;14:2611–2620.