Frontiers in Bioscience Landmark

Front. Biosci. (Landmark Ed) **2022**; 27(2): 067
http://doi.org/10.31083/j.fbl2702067

Original Research

Genetic insights and evaluation of forensic features in Mongolian and Ewenki groups using the InDel variations

Qiong Lan[1], Congying Zhao[1], Cong Wei[2], Hui Xu[1], Chunmei Shen[3,4,*], Bofeng Zhu[1,5,6,*]

[1]Guangzhou Key Laboratory of Forensic Multi-Omics for Precision Identification, School of Forensic Medicine, Southern Medical University, 510515 Guangzhou, Guangdong, China
[2]Department of Ophthalmology, The Six Affiliated Hospital of Guangzhou Medical University, Qingyuan People's Hospital, 511500 Guangzhou, Guangdong, China
[3]Zhujiang Hospital, Southern Medical University, 510282 Guangzhou, Guangdong, China
[4]Institute of Brain and Behavioral Sciences, College of Life Sciences, Shaanxi Normal University, 710062 Xi'an, Shaanxi, China
[5]Key Laboratory of Shaanxi Province for Craniofacial Precision Medicine Research, College of Stomatology, Xi'an Jiaotong University, 710000 Xi'an, Shaanxi, China
[6]Clinical Research Center of Shaanxi Province for Dental and Maxillofacial Diseases, College of Stomatology, Xi'an Jiaotong University, 710000 Xi'an, Shaanxi, China
*Correspondence: zhubofeng@i.smu.edu.cn (Bofeng Zhu); cmshen2004@126.com (Chunmei Shen)
Academic Editor: Graham Pawelec
Submitted: 15 October 2021 Revised: 22 December 2021 Accepted: 12 January 2022 Published: 14 February 2022

Abstract

Background: InDel polymorphisms show great potential for use with challenging DNA samples in forensic practice due to having similar advantages to STRs and SNPs. Large-scale InDel genotype data are becoming available world wide populations, thus providing an alternative for investigating genetic architectures in rarely studied populations from a genome perspective. **Methods**: Here, we genotyped 47 highly polymorphic InDel variations in 157 Mongolian and 100 Ewenki individuals from the Inner Mongolia Autonomous Region of China in order to evaluate their utility for forensic purposes. **Results**: The CDPs of the 47 InDels for these groups were calculated to be 0.99999999999999999874 and 0.99999999999999999677, respectively, while the CPEs were 0.99981 and 0.99975, respectively. The 47 InDel variations were therefore an efficient tool for forensic personal identification in the Mongolian and Ewenki ethnic groups. Comparison of results from the present study with datasets from previously published literatures and from the 1000 Genomes Project revealed a prominent East Asian ancestry component in the gene pool of both ethnic groups, hinting at the close genetic relationships between Mongolian, Ewenki and most East Asian populations. Furthermore, Han populations from northern China showed even closer genetic affinities with the Mongolian and Ewenki groups. **Conclusions**: The data presented here would facilitate the forensic application of InDels for Mongolian and Ewenki ethnic minorities and expand our knowledge regarding the genetic diversities of these populations.

Keywords: InDel; Mongolian; Ewenki; forensic efficiency; genetic structure

1. Introduction

The efficiency of novel molecular genetic markers for forensic genetic applications has been widely investigated in recent years. A new era started with the first application of DNA fingerprinting in a paternity test case [1]. Subsequently, short tandem repeats (STRs) became the most common genetic markers used in human identification and parentage testing because of their high polymorphisms and genotyping convenience [2]. STR-based genetic profiling has become the gold standard in forensic DNA analysis due to the generalized establishment of STR databases in worldwide populations [3,4]. However, the poor resolution of degraded or low-template DNA samples caused by adverse environmental conditions remains very challenging for forensic DNA laboratories, despite the substantial progress made in STR genotyping technology. Apart from the stutter peaks, forensic case samples sometimes also show genotyping failures of large amplicons in STR profiling. This has led to a search for more suitable molecular genetic markers. Whole genome sequencing technology can identify many novel variants and offer tremendous potential for uncovering better genetic markers for more challenging forensic samples [5,6]. The superior feasibility of single nucleotide polymorphisms (SNPs) has already been documented for the individual identification, biogeographic ancestry inference, external visible trait prediction, and genealogical inference [7–10]. However, the complex experimental procedure of the SNaPshot assay make it time-consuming and susceptible to contamination, thus hampering the widespread use of SNPs in routine forensic DNA laboratories.

The di-allelic insertion/deletion (InDel) genetic marker is a length-based polymorphism that can overcome some of the drawbacks of STRs and SNPs [11,12]. The characteristics of genome abundance, relatively low mutation rate, small amplicon size and compatibility

with current capillary electrophoresis (CE) platform favor the use of InDels as a supplementary tool for STRs. Besides, it is recently reported that InDels are useful for individual identification, mixed DNA identification and deconvolution, as well as for population genetic analyses such as biogeographic ancestry inference in forensic DNA analysis [13–16]. The first commercially available InDel-based panel is developed by QIAGEN for the multiplex amplification of 30 InDels plus a sex-determining marker (Amelogenin) [17]. By taking advantage of smaller amplicon size and CE platform compatibility, the Investigator DIPplex kit provides better resolution for degraded DNA, while having similar sensitivity to STR- and SNP-based approaches for human identification [13,18,19]. To further establish the application of InDels as an acceptable supplementary tool for STRs, additional InDel-based genotyping systems must be developed and validated in different populations. A newly released amplification system that incorporates 47 autosomal InDels, 2 Y-chromosomal InDels and Amelogenin have recently been used to obtain genotype data for different populations within China [20]. However, this amplification system is still in its infancy and more population data are needed before it can be widely used.

So far, genetic variations in the ethnic minorities of northern China have not been systematically studied. The Inner Mongolia Autonomous Region (IMAR), located in northern China, is a centuries-old settlement area for the Han population and many other ethnic minorities. Geographically speaking, IMAR is adjacent to more than 10 provinces from the northeast and southwest parts of China, while bordering Russia and Mongolia to the north. This unique biogeographic location provides an important scenario for the study of gene interactions among different ethnic minorities in the region, with great potential to conduct population genetics-related research. The Mongolian group is the most centralized ethnic minority in IMAR, with a population of 4.8 million according to the 2010 census. The Mongolians are well known in the world for the Mongol Empire, founded by Genghis Khan in the 13th century. As documented in historical records, the Mongolians played a significant role in guiding the genetic drift of populations they came into contact with, especially after construction of the Mongol Empire and conquering of the Eurasian continent under Genghis Khan. The territorial expansion of the Mongol Empire facilitated gene interactions between Mongols and Europeans, leaving notable effects on the genetic structure of Eurasians [21,22]. Genome-wide analysis has revealed that the largest ancestral component in the genome of Mongolians matches that of East Asian genetic components [23]. Mongolians display a closer genetic relationship with East Asians than South Asians. The genetic exploration of Mongolians may shed light on the background of the Mongolian group and its neighboring populations.

The Ewenki group is one of the ethnic minorities in northern China with rich cultural forms. The size of the Ewenki population is 30,875 according to the 2010 census. Ewenkis live mainly in the Hulun Buir area of IMAR and in the Great Khingan Mountains in Heilongjiang Province, China. The mixed residence of Ewenkis with neighboring ethnic minorities means that gene interactions are inevitable during their populations' history. It is therefore essential to reveal these relations from a genetic perspective. Moreover, the Ewenki group faces a crisis of cultural ablation and marginalization due to their small population and the influence of cultural globalization. Scholars from diverse research fields have made considerable efforts to study the customs of Ewenki and to strengthen the protection and passage of traditional Ewenki culture and genetic resources. There is also considerable interest in determining the genetic affinities between the Ewenki group and its neighboring populations.

Currently, the genetic characteristics of ethnic minorities in IMAR are not well described in existing historical documentation and in population genetic data. InDel-based methods may provide an alternative approach for analyzing degraded biological samples when conventional STR analysis fails to produce reliable DNA profiling in forensic practice. Genetic exploration of InDel variations in different populations is therefore highly desirable for both anthropological and forensic research. In the present study, we explored the feasibility of 47 InDel loci for forensic purposes in two ethnic minorities, the Mongolian and Ewenki groups. By incorporating previously published population data and the population genetic dataset from the 1000 Genomes Project phase 3, we made genetic comparisons among these two ethnic minorities and the reference populations in order to obtain a more comprehensive genetic structure of the Mongolian and Ewenki groups.

2. Materials and methods

2.1 Sample collection, DNA extraction and quantification

Volunteers were recruited to the study if they complied with the following conditions: (1) self-reported healthy condition; (2) no biological kinship related to anteriorly recruited participants within at least three generations; (3) no immigration or inter-marriage events in their family histories. In total, 257 blood samples were collected from two ethnic minorities (Mongolian and Ewenki) of IMAR, China. The sample sizes were 157 Mongolian and 100 Ewenki volunteers, respectively. Blood samples were stored at –20 °C before DNA extraction using the paramagnetic particle method. The concentration of DNA samples was estimated by measuring their absorbance at 260 nm using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) according to the manufacturer's recommendation. DNA samples were then diluted to 1 ng/μL. Written informed consents were obtained from all participants before the study began. The

study received approvals from the Ethics Committees of Southern Medical University, Guangzhou, China and Xi'an Jiao Tong University, Xi'an, China (No. XJTULAC201).

2.2 PCR amplification and subsequent InDel genotyping

PCR amplification of the overall loci was carried out in a single PCR system using the reagents and reaction conditions as described in the manufacturer's protocol for the AGCU InDel 50 kit (AGCU, Wuxi, China) and performed in a GeneAmp PCR 9700 Thermal Cycler (Applied Biosystem, Foster City, CA, USA). The 25 μL reaction volume for the multiplex amplification system included 10 μL of reaction mix, 5 μL of primer mix, 1 μL of heat-activated C-Taq DNA polymerase, 1 ng of genomic DNA, and filled to the total volume with sdH_2O. Thermal cycling conditions were programmed as follows: the initial pre-degeneration step was 95 °C for 2 min, followed by 29 cycles of denaturation at 94 °C for 30 s, annealing at 60 °C for 1 min, extension at 72 °C for 1 min, and a final extension at 60 °C for 30 min. The ABI 3500 xL Genetic Analyzer (Applied Biosystem, Foster City, CA, USA) was used to detect PCR product in a 1 μL volume added to 0.5 μL of AGCU Marker SIZ-500 and 12 μL of deionized formamide. Subsequently, InDel genotyping was performed by GeneMapper ID-X version 1.5 Software (Applied Biosystem, Foster City, CA, USA) with the analytical threshold of peak height at 50 relative fluorescence units (RFU) for allele calling. Positive and negative controls were included to ensure the accuracy of results for InDel genotyping.

2.3 Reference population dataset

Genotype data for the 47 autosomal InDel variations in 26 reference populations from the 1000 Genomes Project Phase 3 were downloaded from the online website Ensemble Genome Browser (http://grch37.ensembl.org/index.html) in order to conduct inter-population genetic analyses. Previously published datasets describing genetic diversities in the 47 polymorphic InDel loci from different geographic regions in China were also integrated. These included data for the 47 InDels from the Hainan Han (HAH), Hainan Li (HNL), Zunyi Gelao (ZGL), Beijing Han (BJH), Henan Han (HNH), Heilongjiang Han (HLJH), Shandong Han (SDH), Shanxi Han (SXH), Tibetan, Chengdu Han (CDH), Yi, and Guangdong Han (GDH) populations [20,24–26]. We then integrated the genotype data for CHB from the 1000 Genomes Project Phase 3 and BJH from previous research to create a combined dataset of 256 Han individuals from Beijing city, referred to as BJH in this study. Finally, a combined dataset was generated for the subsequent analyses by combining Mongolian, Ewenki and other individuals from 37 reference populations. Detailed information for the 26 continental reference populations derived from the 1000 Genomes Project Phase 3 is shown in **Supplementary Table 1**.

2.4 Forensic statistical parameters

Forensic statistical parameters for the 47 InDel variations in the Mongolian and Ewenki ethnic minorities were calculated using the browser-based STRAF application (http://cmpg.unibe.ch/shiny/STRAF/) [27]. These included match probability (MP) [28], discrimination power (DP), probability of exclusion (PE), polymorphism information content (PIC) [29] and observed heterozygosity (Hob). Linkage disequilibrium (LD) analyses of pairwise InDel loci in each group were carried out using SNPAnalyzer version 2.0 (Istech, Goyang, South Korea) Software [30]. An r^2 threshold of 0.8 was used to determine the strong linkage state for pairwise InDel loci. Genepop software (version 4.04, Rousset, France) was used to perform Hardy-Weinberg equilibrium (HWE) tests for all loci in the two ethnic minorities, and to calculate the pairwise F-statistic (F_{ST}) index in the 39 populations [31]. The cumulative discrimination power (CDP) and the combined probability of exclusion (CPE) were calculated with Excel by reference to the relevant formulae.

2.5 Inter-population comparison analyses

The Dispan program was used to calculate *Nei's* D_A genetic distances (D_A distances) among the overall populations. Principal component analysis (PCA) was performed at the individual level to reveal the spatial distribution pattern for all samples from the two ethnic minorities and from the reference populations. The efficiency of the 47 InDel loci in assigning these individuals to the corresponding biogeographic regions was not immediately intuitive. We therefore generated different clustering patterns at the population level for five continents (Africa, Europe, America, East Asia and South Asia) and three continents (Africa, Europe and East Asia) using the *R* program. Genetic differences between the two ethnic groups and the reference populations were more intuitively reflected by simultaneously analyzing the clustering patterns of the overall populations using different PCA plots. The Jensen–Shannon divergence values [32] for the 47 InDel variations were calculated using the built-in algorithm "Thorough analysis of population data of a custom Excel file" in the online Snipper site (http://mathgene.usc.es/snipper/analysispopfile2_new.html) [33]. The Jensen–Shannon divergence values were then converted to the "informativeness-for-assignment" metric (I_n) by referring to the previously reported formula [34]. Arlequin software (version 3.5.1.2, Laurent Excoffier & Heidi Lischer, Switzerland) [35] was used to compute locus-by-locus *p* values of the pairwise populations within China in order to reveal intra-population differentiations within the same geographic region by the method of molecular variation analysis of variance (AMOVA) [36].

To investigate in more detail the substructure patterns for the two ethnic minorities from IMAR and to infer the proportions of different ancestral genetic components, STRUCTURE software (version 2.3.4, Pritchard &

Stephens & Donnelly, United Kingdom) was used to perform unsupervised ancestral component prediction of the populations [37]. The length of burn-in period was 100,000 times followed by 100,000 MCMC repetitions. Initial runs were performed without any prior information about the sample origin and were based on the admixture model and correlated allele frequencies pattern. The number of hypothetical ancestry clusters (K) was set from 2 to 5, with 15 independent runs performed for each of the tested K values. The optimal K was identified using the online web Harvester program (http://taylor0.biology.ucla.edu/structureHarvester/) [38]. The average permutated individual and population Q-matrices for 15 replicates of each K were assessed by CLUMPP software (version 1.1.2, Jakobsson & Rosenberg, USA) [39]. Subsequent plotting (bar plot) was conducted using DISTRUCT software (version 1.1, Jakobsson & Rosenberg, USA) [40] with the input of CLUMPP results. Phylogenetic construction of the two ethnic groups and of the reference populations was generated based on D_A distances with the neighbor-joining method applied by MEGA software (version 6.06, Tamura, Japan) [41]. Subsequent tree visualization and management were performed with the online tool ITOL [42].

To better visualize the parameter distributions and gain a comprehensive view of the differences, Rstudio software (https://www.rstudio.com/products/rstudio/download/) was used to draw the boxplot of the forensic statistical parameters for the two ethnic minorities, the heat map of the insertion allelic frequencies in five continental populations, and the heat map of F_{ST} and D_A values for pairwise populations.

3. Results

3.1 HWE exact tests for all InDel loci and LD analyses of pairwise loci

HWE exact tests for the 47 InDel loci in the Mongolian and Ewenki groups were performed to confirm the validity of sample collection. After Bonferroni's correction, p value of <0.0011 (0.05/47 = 0.0011) was considered to represent deviation from the equilibrium state. All InDel loci were found to comply with HWE, with a minimum p value of 0.008 for the rs142221201 locus in Mongolians, and 0.014 for the rs151335218 locus in the Ewenki group.

Before using the product law in the CDP and CPE calculations, LD tests of pairwise InDel loci were performed to assess the independence of each InDel locus. The results of LD analyses were mirrored by data matrix of inverted triangles. As shown in **Supplementary Fig. 1**, none of the small block was covered in crimson and no area was encircled by the thick black line with the r^2 threshold established at 0.8. This showed that no LDs existed between any of two different InDel loci in the Mongolian and Ewenki ethnic groups.

3.2 Allelic frequencies and the forensic efficiency parameters of 47 InDel loci in the Mongolian and Ewenki groups

Allelic frequencies for the 47 InDel loci in the two ethnic minorities were calculated and the results were presented in Table 1. In the Mongolian group, the insertion allele frequencies ranged from 0.2803 at rs67939200 to 0.7548 at rs3834231, while in the Ewenki group they ranged from 0.2750 at rs67264216 and rs1127697 to 0.8050 at rs3834231. Overall, 89.36% and 82.98% of insertion allele frequencies were within the range of 0.3 to 0.7 in the Mongolian and Ewenki groups, respectively. PD values ranged from 0.5299 at rs3834231 to 0.6494 at rs66739142 in the Mongolian group, and from 0.4778 at rs3834231 to 0.6526 at rs5787309 in the Ewenki group. Hob values ranged from 0.3694 at rs67700747 to 0.5478 at rs5787309 in the Mongolian group, and from 0.2900 at rs3834231 to 0.6300 at rs151335218 in the Ewenki group. PE values ranged from 0.0964 at rs67700747 to 0.2328 at rs5787309 in the Mongolian group, and from 0.0595 at rs3834231 to 0.3284 at rs151335218 in the Ewenki group. Subsequently, boxplots were constructed to show the distribution of forensic parameters and to make comparison in the two ethnic groups. As depicted in Fig. 1, red and green boxes show the general data dispersion of forensic parameters in the two groups. The left side of the figure shows that most parameters (GD, Hobs, PM) were concentrated between 0.4 and 0.5, but not PDs and PEs. The median value of GD, Hobs, PD, PE and PM exceeded 0.4, 0.4, 0.6, 0.15 and 0.4, respectively, in both the Mongolian and Ewenki groups. Larger variations in the GD, Hob, PD, PE and PM values were observed in the Ewenki group. Forensic statistical parameters were then calculated to investigate the efficiencies of the 47 InDel loci for applications in individual identification and parentage testing. The CDP and CPE values were calculated to be 0.99999999999999999874 and 0.99981 respectively in the Mongolian group, and 0.99999999999999999677 and 0.99975 respectively in the Ewenki group. These results strongly support the 47 InDels as suitable tool for personal identification in the two ethnic groups. **Supplementary Tables 2** and **3** summarize the detailed information regarding forensic parameters for the 47 InDel loci.

3.3 Inter-population patterns of genetic diversities in the two ethnic study groups and in the reference populations

Gene interactions are prone to occur among different populations living in the same biogeographic region. To examine how historical events shaped the genetic diversities of the two ethnic minorities in IMAR, genotype data for the 47 InDel loci in the overall 39 populations were assembled to determine population structures and genetic affinities. Firstly, a heat map of the insertion allele frequency matrix for the 47 InDel variations was generated to reflect population genetic diversities (Fig. 2). Insertion allele fre-

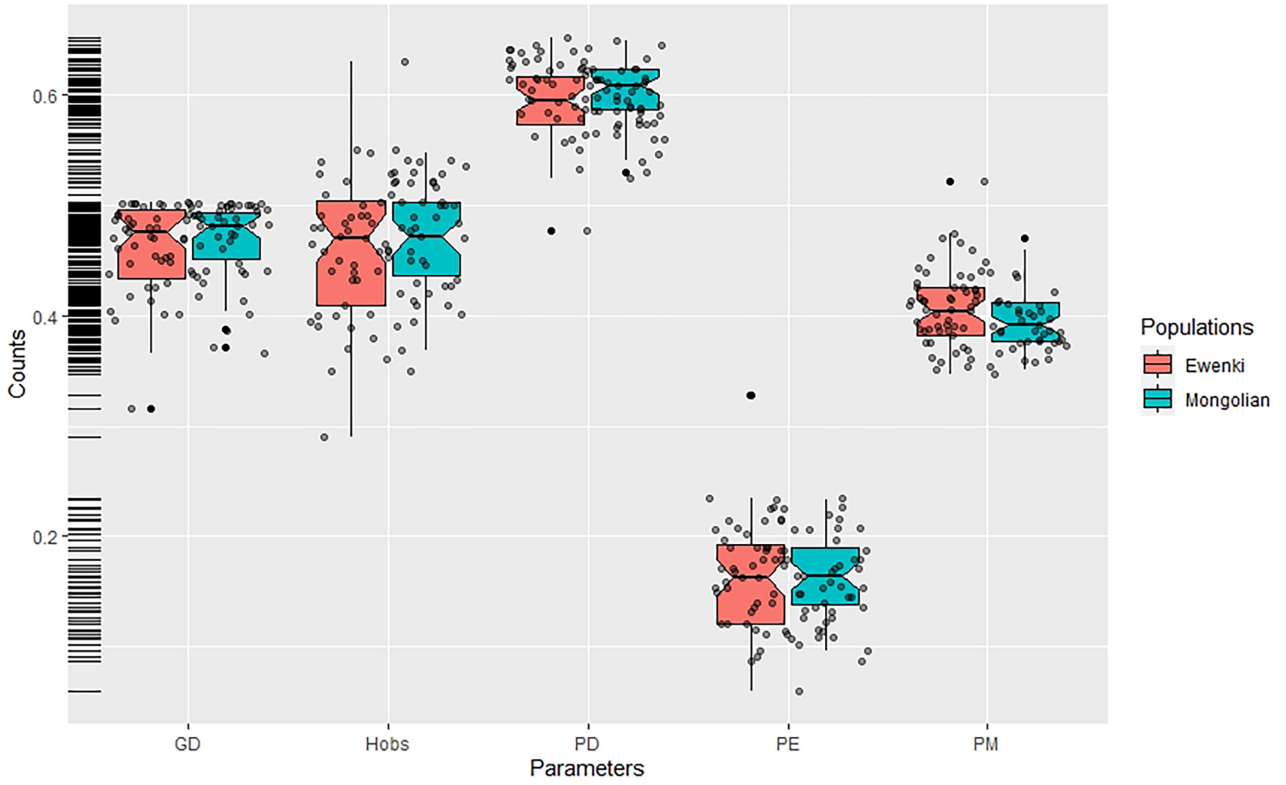


Fig. 1. Box plot of the forensic statistical parameters (GD, Hobs, PD, PE and PM) in the two ethnic minorities. Vertical lines in the boxplot denote the range, middle lines denote the median, while the bottom and top of each box corresponds to the first and third quartiles, respectively. Individual data are represented by dark red dots.

quencies were represented by different colors, varying from light blue (0) to red (1). On the left and top of the figure, population clusters and locus clusters were generated based on the allele frequency distributions of the 47 InDel loci. In general, populations with similar allele frequency distributions shared the same sub-branch on the left clustering tree, with five clusters easily distinguished. The exceptions were CLM and PUR populations from America positioned in the European cluster. The Mongolian and Ewenki groups clustered with East Asian populations and located on the same sub-branch of the tree. InDel variants that exhibited similar insertion allelic frequencies in different populations were assembled at the top of the heat map and were in the same sub-branch as the other clustering tree.

The heat map revealed relatively obvious discrepancies in allele frequencies between different populations. As proposed by Shriver [43,44], the genetic distance between populations for any single molecular marker can be estimated from the δ metric, known as the allele frequency differential value. We therefore speculated that some of the InDel variations could dissect the population structures to some extent. We also calculated the I_n value for each InDel variation in different populations from three continents (Africa, Europe, East Asia), four continents (Africa, Europe, East Asia and South Asia) and five continents (Africa, Europe, East Asia, South Asia and America) with Snipper-based PCA. Table 1 showed I_n values for the 47 InDels ranking in descending order. The seven most informative InDel loci with I_n values >0.1 were extracted and their insertion allelic frequency distributions in the five populations from three continents were shown in Fig. 3. The ESN from Africa, FIN from Europe and CHB from East Asia were considered here to be representative of their continent's populations. The seven most informative InDel variations were the rs72085595, rs34287950, rs71852971 rs66477007, rs79225518, rs5897566 and rs538690481 loci. As shown in Fig. 3, differences in insertion allele frequencies were observed between the continental populations, with the African populations showing significantly different allelic frequency distributions compared to non-African populations. In contrast, similar insertion allelic frequency distributions were observed amongst populations from the same continent (CHB, Mongolian and Ewenki).

Next, PCA at the individual level was performed to reveal the sample distributions from Mongolian, Ewenki and reference populations. As shown in Fig. 4A, individuals from the two ethnic groups and the reference populations from five continents (Africa, Europe, East Asia, South Asia and America) were represented by dots labeled with different colors. PCA showed that individuals were scattered in the middle of the plot, with the first two PCs accounting for 12.5% of the total variance. Population dis-

Table 1. Insertion allele frequencies of the 47 InDels in different populations and the corresponding I_n values in continental populations.

ID	rs	Allele frequencies					I_{n3}	I_{n4}	I_{n5}
		ESN	CHB	FIN	Ewenki	Mongolian			
1	rs72085595	0.8939	0.3033	0.6717	0.2950	0.3376	0.1742	0.1362	0.1103
2	rs34287950	0.8535	0.4147	0.9444	0.4850	0.4108	0.1415	0.1155	0.0925
3	rs71852971	0.8586	0.3341	0.3182	0.3100	0.3694	0.1324	0.1168	0.0949
4	rs66477007	0.8333	0.6185	0.2879	0.6750	0.5573	0.1118	0.0963	0.0983
5	rs79225518	0.9646	0.4810	0.7323	0.6400	0.5732	0.1071	0.0805	0.0648
6	rs5897566	0.9192	0.6588	0.4545	0.5950	0.5955	0.1031	0.1026	0.0839
7	rs538690481	0.0000	0.4100	0.1970	0.5200	0.4172	0.1029	0.0831	0.0678
8	rs3067397	0.7778	0.3104	0.4293	0.3800	0.3248	0.0994	0.0943	0.0818
9	rs66739142	0.9444	0.4882	0.8485	0.5350	0.5127	0.0869	0.0662	0.0575
10	rs3029189	1.0000	0.6517	0.8485	0.6250	0.6242	0.0868	0.0658	0.0529
11	rs67426579	0.1717	0.4408	0.6667	0.3850	0.3981	0.0857	0.0906	0.0737
12	rs66595817	0.9242	0.6351	0.4747	0.7250	0.6815	0.0700	0.0556	0.0445
13	rs3076465	0.8939	0.5095	0.3889	0.4900	0.4618	0.0678	0.0527	0.0530
14	rs60575667	0.8535	0.4313	0.6414	0.4100	0.4618	0.0640	0.0622	0.0503
15	rs35453727	0.8535	0.4408	0.8586	0.3900	0.4713	0.0633	0.0480	0.0456
16	rs67100350	0.4798	0.5924	0.1919	0.5700	0.6210	0.0601	0.0678	0.0590
17	rs34419736	0.7374	0.4265	0.4747	0.4150	0.3439	0.0572	0.0512	0.0451
18	rs67365630	0.8283	0.4834	0.5556	0.5200	0.5478	0.0496	0.0381	0.0306
19	rs67487831	0.9040	0.5640	0.6162	0.6950	0.6210	0.0495	0.0410	0.0330
20	rs10558392	0.3889	0.6540	0.4394	0.5100	0.6146	0.0494	0.0446	0.0416
21	rs67405073	0.8081	0.4384	0.6414	0.3950	0.4140	0.0492	0.0520	0.0417
22	rs3217112	0.8636	0.5853	0.8182	0.6050	0.6019	0.0451	0.0340	0.0279
23	rs35309403	0.4798	0.6280	0.2576	0.6900	0.6592	0.0418	0.0357	0.0303
24	rs142221201	0.7727	0.7038	0.9495	0.6800	0.7038	0.0376	0.0286	0.0232
25	rs35267904	0.3485	0.7038	0.5152	0.6650	0.6019	0.0353	0.0268	0.0221
26	rs5787309	0.4192	0.4929	0.2828	0.4950	0.5223	0.0317	0.0412	0.0364
27	rs145577149	0.6768	0.3886	0.5657	0.4150	0.3981	0.0314	0.0267	0.0232
28	rs35065898	0.7727	0.5474	0.8232	0.5200	0.5764	0.0295	0.0243	0.0200
29	rs11277697	0.4040	0.3863	0.6162	0.2750	0.4045	0.0290	0.0224	0.0191
30	rs67939200	0.2677	0.3507	0.1717	0.2900	0.2803	0.0272	0.0207	0.0209
31	rs1160980	0.6162	0.3957	0.5051	0.4250	0.4172	0.0272	0.0204	0.0179
32	rs34421865	0.8889	0.6280	0.7929	0.5550	0.5669	0.0232	0.0175	0.0240
33	rs139934789	0.7929	0.6682	0.5253	0.6450	0.6242	0.0213	0.0283	0.0254
34	rs34529638	0.4848	0.6730	0.6869	0.6650	0.6752	0.0178	0.0165	0.0137
35	rs34209360	0.8485	0.6896	0.7323	0.6950	0.6274	0.0139	0.0123	0.0132
36	rs769299	0.4394	0.6280	0.4343	0.5700	0.5350	0.0136	0.0102	0.0148
37	rs67700747	0.7222	0.7820	0.5556	0.7300	0.7389	0.0123	0.0265	0.0214
38	rs145941537	0.3737	0.6327	0.6162	0.6550	0.6433	0.0121	0.0092	0.0086
39	rs67264216	0.4646	0.3436	0.5253	0.2750	0.3248	0.0106	0.0087	0.0117
40	rs35464887	0.5859	0.5071	0.4343	0.4400	0.4777	0.0098	0.0175	0.0156
41	rs151335218	0.4192	0.5450	0.5455	0.5250	0.4968	0.0089	0.0182	0.0161
42	rs140683187	0.5758	0.4052	0.5101	0.5400	0.5191	0.0086	0.0088	0.0080
43	rs140323077	0.9141	0.7133	0.7879	0.7600	0.7102	0.0065	0.0146	0.0119
44	rs10607699	0.7222	0.7038	0.5303	0.6400	0.6783	0.0063	0.0047	0.0044
45	rs3834231	0.5556	0.6801	0.6111	0.8050	0.7548	0.0052	0.0039	0.0035
46	rs33971783	0.4091	0.3839	0.5455	0.3200	0.3471	0.0025	0.0019	0.0039
47	rs61490765	0.4242	0.5237	0.3283	0.4900	0.5287	0.0010	0.0023	0.0020

Note: I_{n3}, I_{n4}, I_{n5}, I_n value of the InDel variation among populations from three continents (African, European, East Asian), four continents (African, European, East Asian, South Asian) and five continents (African, European, East Asian, South Asian, American), respectively.

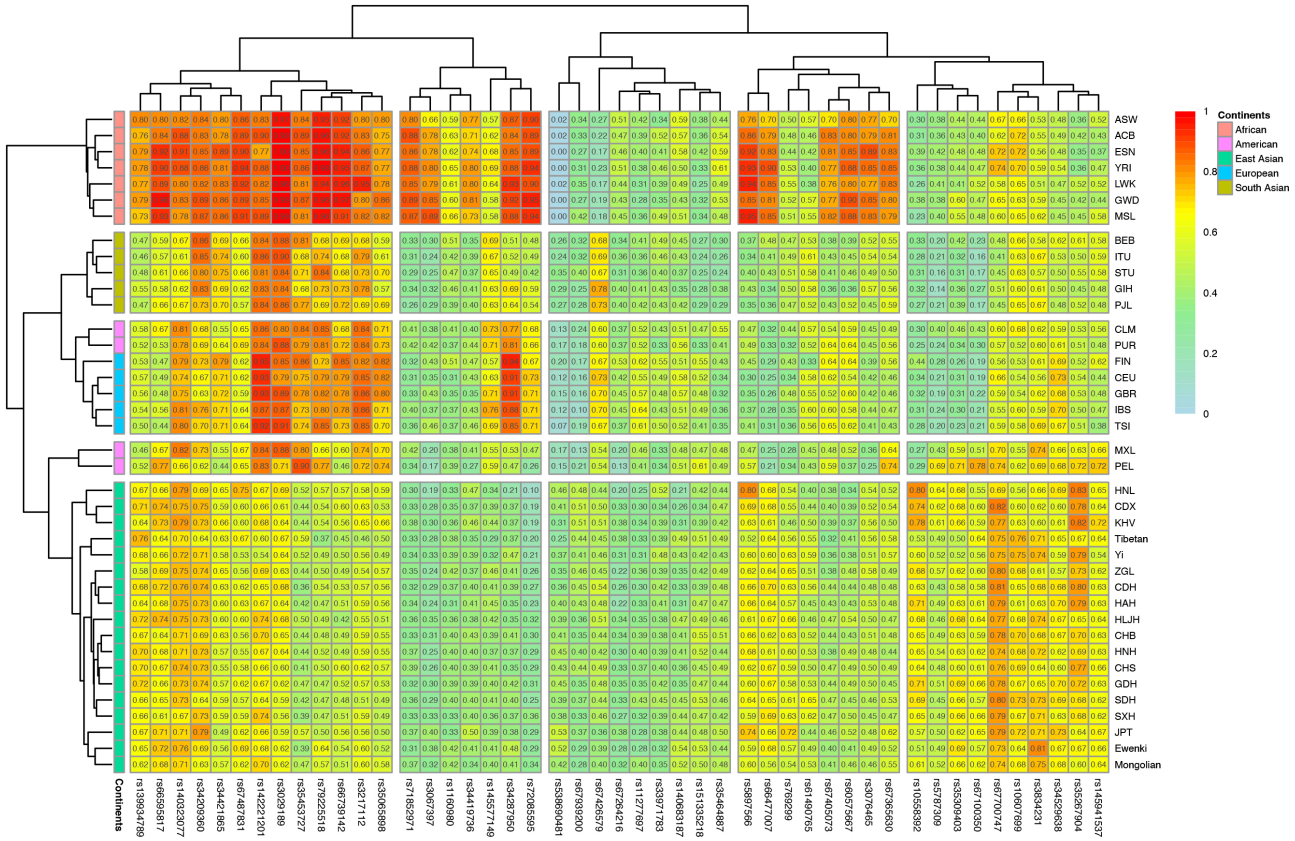
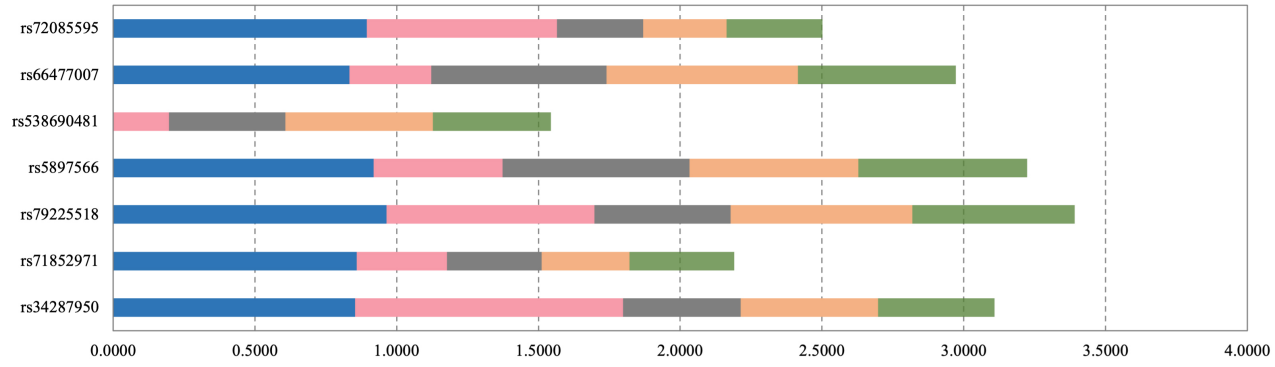


Fig. 2. Heat map of insertion allelic frequencies for the 47 InDel variations in the two ethnic groups and 37 reference populations. The color gradient for the insertion allele frequency distribution ranges from light blue to red. At the top and left sides of the heat map, clustering trees are generated based on the corresponding insertion allele frequency distributions in different populations and at different loci, respectively.



	rs34287950	rs71852971	rs79225518	rs5897566	rs538690481	rs66477007	rs72085595
ESN	0.8535	0.8586	0.9646	0.9192	0.0000	0.8333	0.8939
FIN	0.9444	0.3182	0.7323	0.4545	0.1970	0.2879	0.6717
CHB	0.4147	0.3341	0.4810	0.6588	0.4100	0.6185	0.3033
Ewenki	0.4850	0.3100	0.6400	0.5950	0.5200	0.6750	0.2950
Mongolian	0.4108	0.3694	0.5732	0.5955	0.4172	0.5573	0.3376

Fig. 3. The seven most informative InDel variations (I_n >0.1) for the differentiation of African, European and East Asian populations are shown. Insertion allele frequencies for the InDels in the Mongolian and Ewenki groups and for the reference populations (ESN, African; FIN, European; CHB, East Asian) are shown using different colors in the bar plot. The insertion allele frequencies for the InDel variations are shown below the figure.

tinctions at the five continent level were not clearly obvious in this analysis. We next drew the PCA plot at the four continent population level (Africa, Europe, East Asia and South Asia) to reveal the distribution of recruited individuals, with the two topmost PCs making up 12.71% of the total variance (Fig. 4B). Previous studies reported that each PC captured only a portion of the total variability, but here the first PC captured the largest portion followed by the second PC. The combined PCs define a sample's eigenvector, and usually the 2D PC1–PC2 plots contain the most information in the simplest space [45,46]. Based on the PCA plots generated from clustering patterns of different continental populations, we concluded the genetic structures of populations from five continents and four continents could not be ideally dissected by the 47 InDel variations. However, the tendency indicated that populations from Africa, Europe and East Asia could be distinguished by the set of 47 InDel loci. Hence, we next generated the clustering pattern of individuals from the two ethnic study groups and the reference populations from Africa, Europe and East Asia. This allowed us to evaluate the potential of the 47 InDels in correctly assigning individuals to their corresponding biogeographic clusters (Fig. 4C). Compared to the results shown in Fig. 4B, more distinct boundaries were seen in Fig. 4C between the African, European and East Asian populations. Populations from the same continents assembled together to form relatively distinguishable clusters. The two ethnic minorities from IMAR assembled with the East Asian populations, thus indicating their close genetic affinities. An additional PCA plot (Fig. 4D) was generated to confirm the efficiency of the seven most informative InDel variations (I_{n3} >0.1) with regard to ancestry inference amongst populations from Africa, Europe and East Asia. The seven InDel variations produced an analogous clustering pattern to the PCA plot generated by the 47 InDels, with several tightly assembled clusters seen within the African populations.

Based on the maximum-likelihood algorithm, STRUCTURE software [47] was used to further characterize the substructure patterns of the two ethnic minorities and to infer the proportion of ancestral components using the reference continental populations. Using the online Harvest program (http://taylor0.biology.ucla.edu/structureHarvester/) [38], the optimal K was determined to be 3 (**Supplementary Fig. 2**). At the best-fit model of K = 3 (Fig. 5A), highly specific genetic components were observed in populations from the same continent. Accordingly, the genetic ancestries of African, European and East Asian populations are represented by green, pink and blue-green lines, respectively. A large fraction of East Asian ancestral genetic component was detected in the Mongolian and Ewenki groups. We also investigated ancestry component discrepancies for the overall populations when the genetic ancestry was pre-assumed to be 2–5. No extra discrepancies in ancestry component compositions were detected among the studied and reference populations with the increased K. Thus, together with the population genetic patterns inferred from K = 3, the present results corroborated the unambiguous differentiation of African, European and East Asian individuals by STRUCTURE genetic ancestry prediction analyses using these 47 InDels. The triangle plot (Fig. 5B) also revealed that individuals from the same continent assembled into independent clusters and reflected the pattern of genetic similarity of the studied and reference populations. Furthermore, the bar plot was used to represent the estimated ancestry composition of the populations by STRUCTURE (Fig. 5C), with the results showing that Mongolian and Ewenki groups possess a large fraction of East Asian ancestry. The estimated East Asian ancestry components for the Mongolian and Ewenki groups were 88.6% and 88.2%, respectively. However, the average East Asian ancestry component of the Chinese populations investigated in this study was 92.8%, indicating slightly less East Asian ancestry in the Mongolians and Ewenkis compared to other Chinese populations.

3.4 Phylogenetic reconstruction based on D_A genetic distances amongst different populations

F_{ST} was used to estimate the genetic distances amongst the 39 different populations. Data matrixes of pairwise F_{ST} values and D_A distances between different populations were visualized in two different heat maps. The color gradient for pairwise F_{ST} went from white to sea green and then red, while the color gradient for D_A distances went from white to yellow and then pink. Blocks in red or pink represent larger pairwise F_{ST} or D_A distances, respectively. From the distribution of pairwise F_{ST} values (Fig. 6A), we concluded that African populations were genetically more distant than most non-African populations, showing in particular larger pairwise F_{ST} values with East Asian populations than with non-East Asian populations. Smaller pairwise F_{ST} values were observed among different populations from the same continent, as seen by the representation of white and light-yellow blocks. The Mongolian and Ewenki ethnic groups shared close genetic relationships with most East Asian populations, but showed relatively less genetic affinities with KHV and HNL populations. The distribution of D_A distances among populations (Fig. 6B) mirrored the analogous population genetic relatedness observed with the distribution of pairwise F_{ST} values. The dataset for F_{ST} values and D_A distances is summarized in **Supplementary Tables 4** and **5**, respectively.

Phylogenetic reconstruction of the overall 39 populations was also generated based on D_A distances to further assess the phylogeny relationships amongst these populations. As shown in Fig. 7, five primary clusters could easily be distinguished from the phylogenetic tree. These were the African cluster (ASW, ACB, LWK, MSL, GWD, ESN and YRI) labeled in green, the European cluster (FIN, IBS, TSI, CEU and GBR) in blue, the East Asian cluster (CDX, KHV,

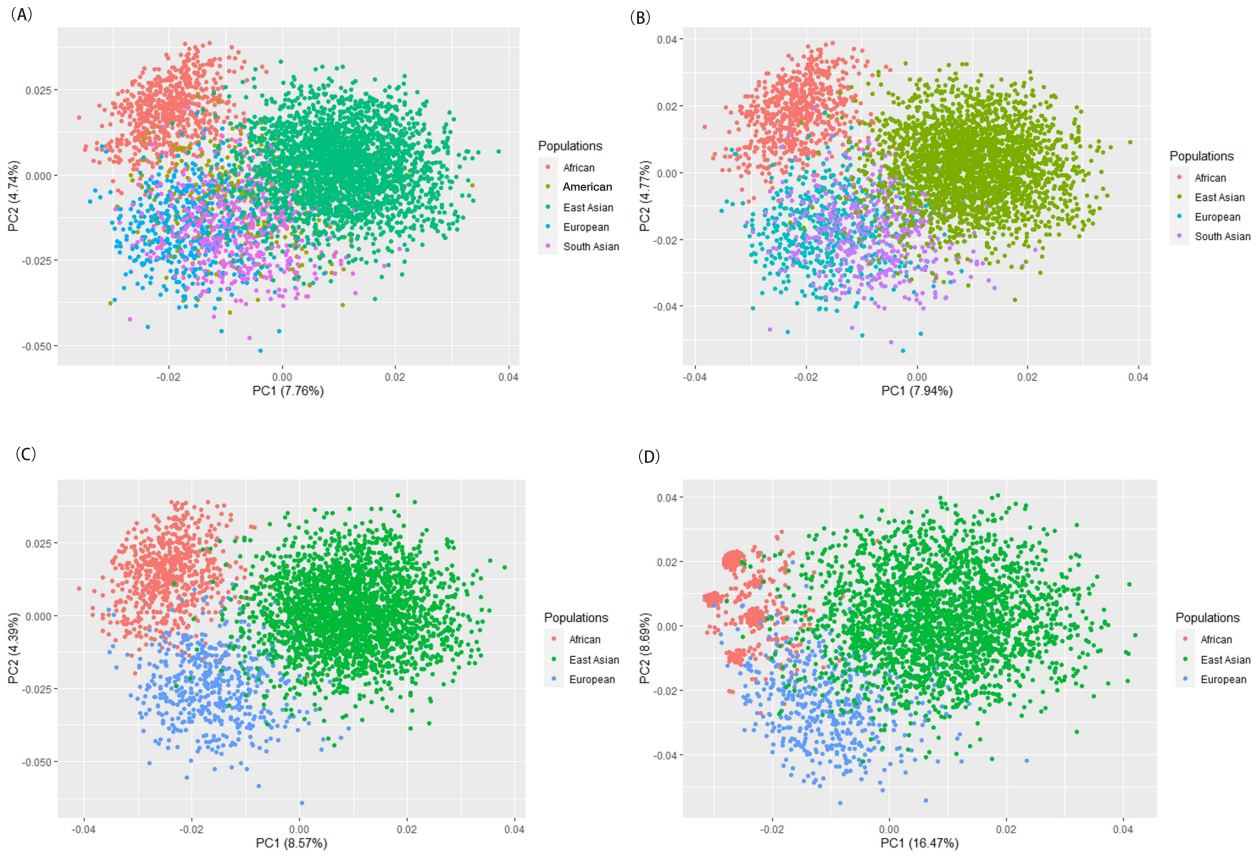


Fig. 4. PCA plots based on the two ethnic groups and the corresponding reference populations. (A) PCA of the two ethnic groups and reference populations from five continents (Africa, Europe, America, South Asia and East Asia). (B) PCA of the two ethnic groups and reference populations from four continents (Africa, Europe, South Asia and East Asia). (C) PCA of the two ethnic groups and reference populations from three continents (Africa, Europe and East Asia). (D) PCA of the two ethnic groups and reference populations from three continents (Africa, Europe and East Asia) based upon the seven most informative InDels with I_{n3} >0.1.

CHS, JPT, CHB, HNL, HAH, CDH, GDH, CDH, HNH, SXH, HLJH, Yi, ZGL, Tibetan, Ewenki and Mongolian) in red, the American cluster (MXL and PEL) in brown, and the South Asian cluster (BEB, STU, GIH, ITU and PJL) in purple. Two additional populations, PUR and CLM, were not positioned at the American cluster (MXL and PEL) and this could be explained by the mixed ancestry of these populations. Furthermore, the Mongolian and Ewenki groups were found to share a sub-branch in the East Asian cluster. Coupled with the genetic ancestral component revealed by STRUCTURE analysis (at K = 3), the results of phylogenetic reconstruction reconfirmed the genetic relatedness of these populations.

3.5 Intra-population genetic relatedness in East Asians

To better understand the genetic relatedness of East Asian populations, locus-by-locus *p* values of intra-population genetic differences between Mongolian and Ewenki groups and the 16 reference populations were evaluated based on genotype data for the 47 InDels. The significance threshold level for *p* was adjusted to 0.0011 ($p = 0.05/47 = 0.0011$) according to the Bonferroni correction. As shown in **Supplementary Table 6**, significant differences were observed between Ewenki and CHS at two loci, with SDH, CDH and Yi at three loci, with CHB, GDH and ZGL at four loci, with CDX and Tibetan at five loci, with KHV at six loci, and with HAH at eight loci. None of the InDel loci showed significant differences between the Ewenki and HNH or SXH populations using the adjusted threshold of locus-by-locus *p* value. In contrast, significant differences were seen between Ewenki and HNL at ten loci. No significant differences were observed between the Mongolian group and the HNH, HLJH and SDH reference populations (**Supplementary Table 7**). The Tibetan and HNL groups were significantly different to the Mongolian group at eight and 17 loci, respectively. With regard to single locus diversity, the top three loci showing the greatest diversity between the two ethnic groups and the reference populations were rs140683187, rs10558392, rs3834231, rs79225518 and rs67939200 at nine, nine, seven, six and six loci, respectively, in the Ewenki group, and rs67939200, rs35267904 and rs140683187 at ten, nine, and eight loci, respectively, in the Mongolian group.

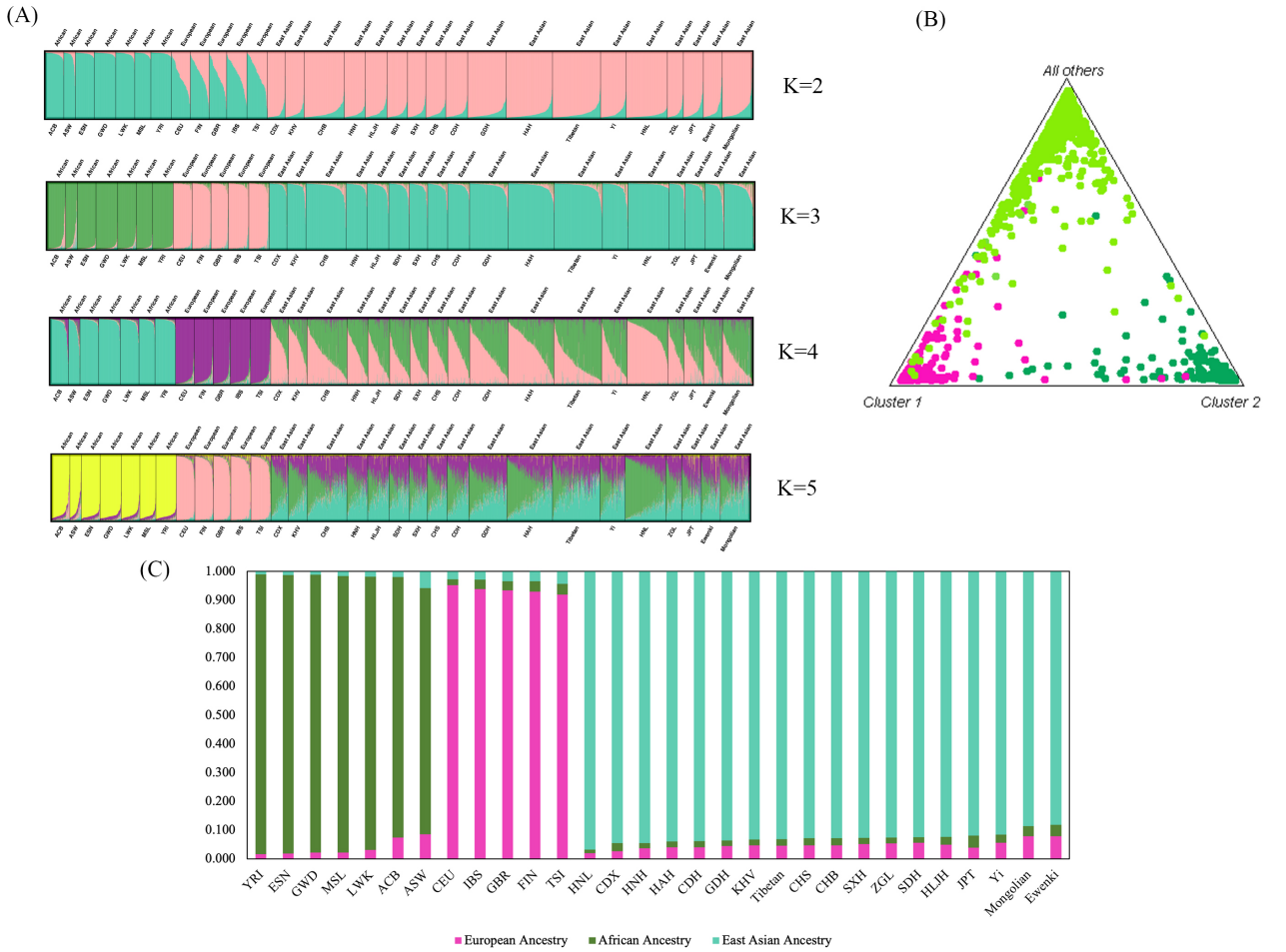


Fig. 5. STRUCTURE analyses of the two ethnic groups and the corresponding reference populations. (A) Bar plot of the estimated genetic ancestral components at the individual level in Mongolian, Ewenki and reference populations from Africa, Europe and East Asia. Individuals are represented by a vertical line divided into K colored segments, with the length of each segment being proportional to the estimated membership. (B) Triangle plot of Mongolian and Ewenki groups and the reference populations from three continents. (C) The bar plot represents the corresponding ancestry composition of reference populations from Africa, Europe and East Asia. The African ancestry is labeled in green, the European ancestry in pink, and the East Asian ancestry in blue-green.

4. Discussion

The accumulated genotype data for InDel variations in different populations provides the necessary foundation to use these polymorphisms as an alternative for resolving difficult samples in forensic DNA analysis. This has been extensively described in a series of recent studies [13,48–52]. In the present work, 47 highly polymorphic InDel loci were analyzed in Ewenki and Mongolian groups from IMAR in order to determine their suitability for forensic applications. The CDP values obtained in these two ethnic groups were sufficiently accurate to enable the identification of unknown individuals in forensic personal identification. Thousands of polymorphic InDel loci have so far been characterized in the human genome and a series of multiplex amplification systems have been developed over the past decade based on InDel analysis. For example, the 38-plex InDel assay was efficient for InDel profiling of degraded samples, whereas the commonly used STRs sometimes failed to generate complete DNA profiles [53]. The above-mentioned Investigator DIPplex ® Kit has been widely tested in worldwide populations to confirm its suitability for forensic applications. Compared with the 38-plex and 30-plex amplification systems, the newly launched 50-plex marker system was shown to be more specific for Chinese populations. Results obtained with the 50-plex InDel system for forensic purposes were more convincing, since the CPEs and CDPs were calculated in different populations [17,53,54]. Therefore, we believe the 50-plex InDel assay can be readily implemented in forensic laboratories and serve as an efficient supplementary tool for STRs. However, the drawbacks of this system should also be highlighted. Firstly, the PICs of some InDel loci (e.g., rs3834231 and rs140323077) were less informative than other InDels included in the InDel system. Secondly,

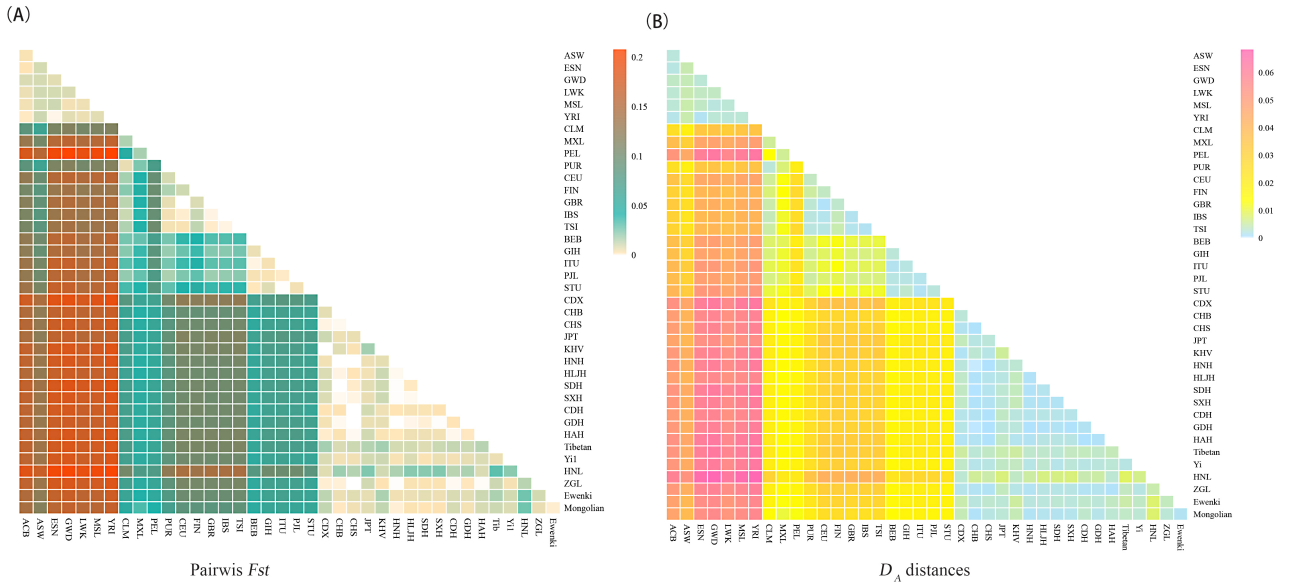


Fig. 6. Heat maps of pairwise F_{ST} values and D_A distances among the overall populations. (A) Data matrix of F_{ST} values among the overall 39 populations. The color gradient varies from white to red, corresponding to a F_{ST} from low to high. (B) D_A distances among the overall populations. The color gradient varies from light blue to pink, corresponding to the D_A distance from low to high.

the maximum amplicon size for the InDel loci in this new system should be further shortened to <200 bp to enable its usage for degraded DNA in forensic analysis.

According to a previous validation study, the 50-plex assay has been confirmed to be reliable and robust for forensic purposes and human genetic research [20]. In the current study, we explored the attributes of the 47 InDel variations (not including the two Y-chromosomal InDels and the Amelogenin locus) for forensic analysis in the Mongolian and Ewenki ethnic groups, as well as conducting further population genetic analyses. To shed light on the genetic backgrounds of these two ethnic groups, we assembled genotype data for the 47 InDels in continent-specific populations from the 1000 Genomes Project, as well as previously published data on Chinese populations. This was used to create a reference population dataset and to perform further phylogenetic analyses. We first created a heat map to show the insertion allelic frequency distributions of the 47 InDels in the overall 39 populations. The heat map showed that closely related biogeographical populations tended to have similar insertion allele frequency distributions with less marked differences in allele frequencies, suggesting close genetic affinities amongst these populations. It was also discovered that some InDel variations (e.g., rs71852971, rs72085595 and rs34287950) had relatively well differentiated allelic frequency distributions in different continental populations. Therefore, we hypothesized that the above-mentioned InDel variations could potentially infer population ancestries and dissect population genetic backgrounds. Subsequently, we used various statistical methods to explore the inter-population patterns of genetic diversity and the intra-population genetic relatedness between the two ethnic groups and the reference populations. Based on genotype data for the 47 InDel variations, we also found that the overall populations tended to form five distinct clusters using pairwise F_{ST}, D_A distance and phylogenetic reconstruction analysis, consistent with the biogeographic locations of continental populations. We speculated that genetic differences at the population level between the two ethnic groups and the reference populations could be revealed by the set of 47 InDel variations. The informative measure I_n was used across the overall five, four and three continental populations to estimate the relative informativeness of the InDel variation for ancestry inference. The results showed that I_n values varied when populations from different biogeographic regions of the world were considered separately. Depending on the values of I_{n3}, I_{n4} and I_{n5}, it was noted that the current set of InDel variations showed better ability to infer the ancestry of populations from Africa, Europe and East Asia. The seven most differentiated InDel variations were extracted by setting a threshold of I_{n3} >0.1. Their informativeness for ancestry inference was demonstrated by the clearly discriminated allele frequency distributions in the two ethnic study groups and reference populations.

As discussed above, the genetic relationships amongst populations were broadly consistent with their geographic locations at a local scale from the population level. PCA using unsupervised classification method revealed that the 47 InDel genetic variants were sufficiently informative to assign individuals from African, European and East Asian populations to three distinct clusters. However, they were

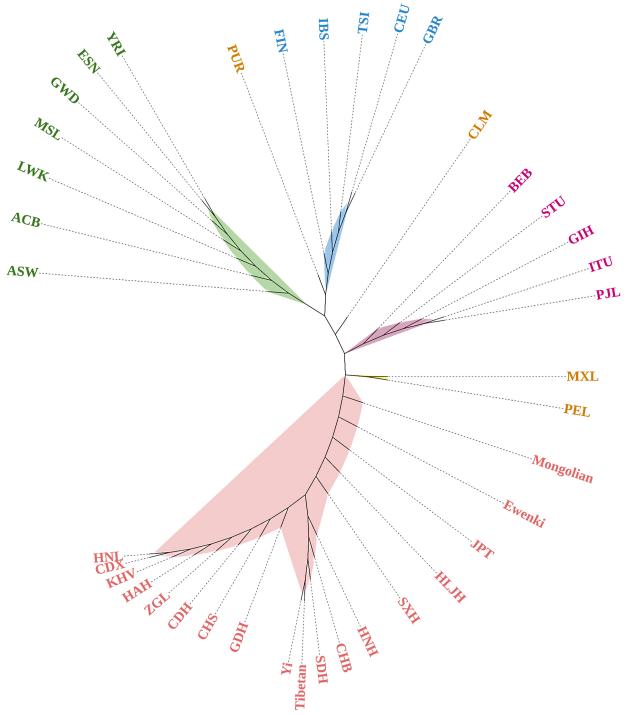


Fig. 7. Phylogenetic reconstruction of the Mongolian group, Ewenki group and the 37 reference populations. Five primary clusters can be easily distinguished from the phylogenetic tree, i.e., the African cluster (ASW, ACB, LWK, MSL, GWD, ESN and YRI) labeled in green, the European cluster (FIN, IBS, TSI, CEU and GBR) labeled in blue, the East Asian cluster (CDX, KHV, CHS, JPT, CHB, HNL, HAH, CDH, GDH, CDH, HNH, SXH, HLJH, Yi, ZGL, Tibetan, Ewenki and Mongolian) in red, the American cluster (MXL and PEL) in brown and the South Asian cluster (BEB, STU, GIH, ITU and PJL) in purple. The other two American populations, PUR and CLM, are scattered among the European and South Asian populations.

not so directional as to distinguish individuals from American and South Asian populations from those in African, European and East Asian populations (Fig. 4A–C). Admixture events for Americans and South Asians could partially explain these PCA results. Nevertheless, the present findings showed the panel was able to reveal the genetic diversity patterns of populations from three continents (Africa, East Asia, Europe). We further performed a more refined PCA (Fig. 4D) based on genotype data for the seven most informative InDel variations in order to infer ancestries in the African, European and East Asian populations. The plot predicted three markedly differentiated population clusters (African, European and East Asian clusters), with better discrimination power amongst the African populations where several tightly assembled clusters were observed. The I_n value is frequently discussed in microhaplotype-related studies [55,56]. It was previously reported that a microhaplotype with I_n <0.2 does not provide adequate information for ancestry analysis [57]. However, in view of the differences between ancestry-informative SNPs and InDels, we tentatively set the threshold to I_n >0.1 to identify ancestry-informative InDel variations. Our efforts here to identify informative InDels for ancestry inference should be valuable for subsequent studies. In addition, STRUCTURE results based on raw genotype data for the 47 InDels at K = 3 showed three distinct population clusters. Moreover, at higher K values the population structure remained basically unchanged at higher K values. Comprehensive analyses by PCA and STRUCTURE found that population structure in the two ethnic groups and the reference populations showed unambiguous genetic diversity patterns for African, European and East Asian populations, together with substantial intra-population genetic homogeneity. These results suggest the 47 InDel variations could be used for biogeographic ancestry analysis in populations from Africa, Europe and East Asia. The Mongolian and Ewenki groups shared similar ancestral components with the reference East Asian populations, and even more so with northern populations from China. However, it should be noted that the HNL group had relatively more genetic differences with the Mongolian and Ewenki groups than the rest of Chinese populations.

Analyses of intra-population genetic relatedness within East Asia using the AMOVA method further confirmed that the two ethnic minorities studied here displayed less differentiation with populations from northern China and more genetic differences with populations from the south and southwest of China. In particular, the HNL group was found to be most differentiated from Mongolian and Ewenki groups. This may be explained by the genetic architecture of HNL having been characterized as a somewhat isolated group [58]. Results from the intra-population comparison analyses were consistent with the conclusions reached from phylogenetic reconstruction and STRUCTURE analysis.

The general genetic characteristics of different populations can be revealed by inter-population comparison studies. To date, genetic insights into the Mongolian group has frequently been conducted by scholars working in diverse research fields. However, studies on the genetic structure of the Ewenki group have so far been rare. A mtDNA-based study reported that Ewenkis shared mtDNA haplogroups with individuals from northern populations of China, similar to the Mongolian group [59] and suggesting a close genetic relationship between the Ewenki and Mongolian groups. Some other ethnic minorities such as the Daur group also exhibit a degree of genetic affinity with the Ewenki group. In general, exploration of the genetic structures and genetic relationships of ethnic minorities is a long and complicated process. In the present study we first analyzed genotype data on 47 InDel variations in the Mongolian and Ewenki ethnic minorities from IMAR. Simultaneously, we made additional comparisons with reference populations in order to dissect the genetic architec-

ture of these minority groups. The results of this study have revealed that the two ethnic groups were closely related to each other genetically. In agreement with previous studies, the two groups also shared genetic similarities with northern populations of China. The results obtained here would contribute to ongoing exploration of the genetic background of East Asian populations and enhance our understanding of the genetic background of populations who share Mongolian and Ewenki ancestry genetic components. However, we acknowledge that more highly polymorphic InDels should be incorporated into the panel in future. The maximum amplicon size of InDel loci in the present panel should also be shortened to better profile the degraded DNA which is usually encountered in forensic work.

5. Conclusions

To summarize, the 47 highly polymorphic InDel variations exhibited sufficient efficiencies for individual identification and could serve as a supplementary tool for STRs in Mongolian and Ewenki groups from the Inner Mongolia Autonomous Region of China. Genetic insights from the 47 InDel variations revealed a prominent East Asian ancestry component in the gene pools of Mongolian and Ewenki ethnic groups. Further speaking, northern populations of China showed closer genetic affinities with the above-mentioned two ethnic minorities. The present results could probably contribute to the ongoing exploration of East Asian population histories and enhance our understanding of genetic backgrounds of populations who share the ancestral genetic components of Mongolian and Ewenki groups.

Author contributions

BZ and CS designed this study. QL conceived the experiments and wrote the manuscript. CW collected the samples. CZ, and HX extracted DNA and helped to conduct the statistical analysis. BZ also revised the manuscript. All the authors authorized and performed the manuscript revision.

Ethics approval and consent to participate

Written informed consents were obtained from the participants before experiments were started. The study was conducted with the approval of the Ethics Committees of Southern Medical University, Guangzhou, China and Xi'an Jiaotong University, Xi'an, China (No. XJTULAC201).

Acknowledgment

Thanks to all the peer reviewers for their opinions and suggestions.

Funding

This research was supported by the National Science Foundation of China (No. 82072122).

Conflict of interest

The authors declare no conflict of interest.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at https://www.imrpress.com/journal/FBL/27/2/10.31083/j.fbl2702067.

References

[1] Jeffreys AJ. Genetic fingerprinting. Nature Medicine. 2005; 11: 1035–1039.

[2] Kayser M, de Knijff P. Improving human forensics through advances in genetics, genomics and molecular biology. Nature Reviews. Genetics. 2011; 12: 179–192.

[3] Taylor D, Bright J, McGovern C, Neville S, Grover D. Allele frequency database for GlobalFiler™ STR loci in Australian and New Zealand populations. Forensic Science International: Genetics. 2017; 28: e38–e40.

[4] Zhabagin M, Sarkytbayeva A, Tazhigulova I, Yerezhepov D, Li S, Akilzhanov R, *et al*. Development of the Kazakhstan Y-chromosome haplotype reference database: analysis of 27 Y-STR in Kazakh population. International Journal of Legal Medicine. 2019; 133: 1029–1032.

[5] Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, *et al*. A map of human genome variation from population-scale sequencing. Nature. 2010; 467: 1061–1073.

[6] van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends in Genetics. 2014; 30: 418–426.

[7] Boonyarit H, Mahasirimongkol S, Chavalvechakul N, Aoki M, Amitani H, Hosono N, *et al.* Development of a SNP set for human identification: A set with high powers of discrimination which yields high genetic information from naturally degraded DNA samples in the Thai population. Forensic science international Genetics. 2014; 11: 166–173.

[8] Jin X, Cui W, Chen C, Guo Y, Tao Y, Lan Q, *et al*. Biogeographic origin prediction of three continental populations through 42 ancestry informative SNPs. Electrophoresis. 2020; 41: 235–245.

[9] Walsh S, Liu F, Ballantyne KN, van Oven M, Lao O, Kayser M. IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Science International: Genetics. 2011; 5: 170–180.

[10] van Oven M, van den Tempel N, Kayser M. A multiplex SNP assay for the dissection of human Y-chromosome haplogroup O representing the major paternal lineage in East and Southeast Asia. Journal of Human Genetics. 2012; 57: 65–69.

[11] Zhang H, He G, Guo J, Ren Z, Zhang H, Wang Q, *et al*. Genetic diversity, structure and forensic characteristics of Hmong-Mien-speaking Miao revealed by autosomal insertion/deletion markers. Molecular Genetics and Genomics. 2019; 294: 1487–1498.

[12] Phillips C. Application of Autosomal SNPs and Indels in Forensic Analysis. Forensic Science Review. 2012; 24: 43–62.

[13] Xie T, Guo Y, Chen L, Fang Y, Tai Y, Zhou Y, *et al*. A set of autosomal multiple InDel markers for forensic application and population genetic analysis in the Chinese Xinjiang Hui group. Forensic Science International. Genetics. 2018; 35: 1–8.

[14] Lan Q, Shen C, Jin X, Guo Y, Xie T, Chen C, *et al*. Distinguishing three distinct biogeographic regions with an in-house developed 39-AIM-InDel panel and further admixture proportion estimation for Uyghurs. Electrophoresis. 2019; 40: 1525–1534.

[15] Santos C, Phillips C, Oldoni F, Amigo J, Fondevila M, Pereira R, *et al*. Completion of a worldwide reference panel of samples

for an ancestry informative Indel assay. Forensic Science International. Genetics. 2015; 17: 75–80.
[16] Phillips C, Santos C, Fondevila M, Carracedo Á, Lareu MV. Inference of Ancestry in Forensic Analysis i: Autosomal Ancestry-Informative Marker Sets. Methods in Molecular Biology. 2016; 1420: 233–253.
[17] LaRue BL, Ge J, King JL, Budowle B. A validation study of the Qiagen Investigator DIPplex® kit; an INDEL-based assay for human identification. International Journal of Legal Medicine. 2012; 126: 533–540.
[18] Ma R, Shen C, Wei Y, Jin X, Guo Y, Mu Y, *et al*. Genetic differentiation and forensic efficiency evaluation for Chinese Salar ethnic minority based on a 5-dye multiplex insertion and deletion panel. Gene. 2018; 660: 41–50.
[19] Martínez-Cortés G, García-Aceves M, Favela-Mendoza AF, Muñoz-Valle JF, Velarde-Felix JS, Rangel-Villalobos H. Forensic parameters of the Investigator DIPplex kit (Qiagen) in six Mexican populations. International Journal of Legal Medicine. 2016; 130: 683–685.
[20] Chen L, Du W, Wu W, Yu A, Pan X, Feng P, *et al*. Developmental validation of a novel six-dye typing system with 47 a-InDels and 2 Y-InDels. Forensic Science International. Genetics. 2019; 40: 64–73.
[21] Bai H, Guo X, Zhang D, Narisu N, Bu J, Jirimutu J, *et al*. The genome of a Mongolian individual reveals the genetic imprints of Mongolians on modern human populations. Genome Biology and Evolution. 2014; 6: 3122–3136.
[22] Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, *et al*. The genetic legacy of the Mongols. American Journal of Human Genetics. 2003; 72: 717–721.
[23] Bai H, Guo X, Narisu N, Lan T, Wu Q, Xing Y, *et al*. Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and East Asia. Nature Genetics. 2018; 50: 1696–1704.
[24] Song F, Lang M, Li L, Luo H, Hou Y. Forensic features and genetic background exploration of a new 47-autosomal InDel panel in five representative Han populations residing in Northern China. Molecular Genetics & Genomic Medicine. 2020; 8: e1224.
[25] Liu J, Du W, Wang M, Liu C, Wang S, He G, *et al*. Forensic features, genetic diversity and structure analysis of three Chinese populations using 47 autosomal InDels. Forensic Science International: Genetics. 2020; 45: 102227.
[26] Wang M, Du W, He G, Wang S, Zou X, Liu J, *et al*. Revisiting the genetic background and phylogenetic structure of five Sino-Tibetan-speaking populations: insights from autosomal InDels. Molecular Genetics and Genomics. 2020; 295: 969–979.
[27] Gouy A, Zieger M. STRAF-a convenient online tool for STR data evaluation in forensic genetics. Forensic Science International. Genetics. 2017; 30: 148–151.
[28] Balding DJ, Nichols RA. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Science International. 1994; 64: 125–140.
[29] Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics. 1980; 32: 314–331.
[30] Yoo J, Lee Y, Kim Y, Rha SY, Kim Y. SNPAnalyzer 2.0: a web-based integrated workbench for linkage disequilibrium analysis and association analysis. BMC Bioinformatics. 2008; 9: 290.
[31] ROUSSET F. Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Molecular Ecology Resources. 2008; 8: 103–106.
[32] Chen H, Chang C, Hsieh L, Lee H. Divergence and Shannon information in genomes. Physical Review Letters. 2005; 94: 178103.
[33] Fondevila M, Phillips C, Santos C, Freire Aradas A, Vallone PM, Butler JM, *et al*. Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies. Forensic Science International. Genetics. 2013; 7: 63–74.
[34] Rosenberg NA, Li LM, Ward R, Pritchard JK. Informativeness of genetic markers for inference of ancestry. American Journal of Human Genetics. 2003; 73: 1402–1422.
[35] Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources. 2010; 10: 564–567.
[36] Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online. 2007; 1: 47–50.
[37] Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155: 945–959.
[38] Earl DA, vonHoldt BM. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources. 2012; 4: 359–361.
[39] Jakobsson M, Rosenberg NA. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics. 2007; 23: 1801–1806.
[40] Rosenberg NA. Distruct: a program for the graphical display of population structure. Molecular Ecology Notes. 2004; 4: 137–138.
[41] Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular Biology and Evolution. 2013; 30: 2725–2729.
[42] Letunic I, Bork P. Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research. 2021; 49: W293–W296.
[43] Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, *et al*. Ethnic-affiliation estimation by use of population-specific DNA markers. American Journal of Human Genetics. 1997; 60: 957–964.
[44] Frudakis T, K V, Thomas M, Gaskin Z, Ginjupalli S, Gunturi S, *et al*. A Classifier for the SNP-Based Inference of Ancestry. Journal of Forensic Sciences. 2003; 48: 771–782.
[45] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 2006; 38: 904–909.
[46] Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genetics. 2006; 2: e190.
[47] Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology. 2005; 14: 2611–2620.
[48] Tomas C, Poulsen L, Drobnič K, Ivanova V, Jankauskiene J, Bunokiene D, *et al*. Thirty autosomal insertion-deletion polymorphisms analyzed using the Investigator® DIPplex Kit in populations from Iraq, Lithuania, Slovenia, and Turkey. Forensic Science International. Genetics. 2016; 25: 142–144.
[49] Wang Z, Zhang S, Zhao S, Hu Z, Sun K, Li C. Population genetics of 30 insertion-deletion polymorphisms in two Chinese populations using Qiagen Investigator® DIPplex kit. Forensic Science International. Genetics. 2014; 11: e12–e14.
[50] Bus MM, Karas O, Allen M. Multiplex pyrosequencing of InDel markers for forensic DNA analysis. Electrophoresis. 2016; 37: 3039–3045.
[51] Huang Y, Liu C, Xiao C, Chen X, Yi S, Huang D. Development of a new 32-plex InDels panel for forensic purpose. Forensic

Science International: Genetics. 2020; 44: 102171.
[52] da Costa Francez PA, Rodrigues EMR, de Velasco AM, dos Santos SEB. Insertion-deletion polymorphisms–utilization on forensic analysis. International Journal of Legal Medicine. 2012; 126: 491–496.
[53] Pereira R, Phillips C, Alves C, Amorim A, Carracedo A, Gusmão L. A new multiplex for human identification using insertion/deletion polymorphisms. Electrophoresis. 2009; 30: 3682–3690.
[54] Bashir M, Hassan NHB. Analysis of 30 Biallelic INDEL Markers Using the Investigator DIPplex(®) Kit. Methods in Molecular Biology. 2016; 1420: 135–142.
[55] Oldoni F, Kidd KK, Podini D. Microhaplotypes in forensic genetics. Forensic Science International. Genetics. 2019; 38: 54–69.
[56] Oldoni F, Bader D, Fantinato C, Wootton SC, Lagacé R, Kidd KK, *et al*. A sequence-based 74plex microhaplotype assay for analysis of forensic DNA mixtures. Forensic Science International: Genetics. 2020; 49: 102367.
[57] Kidd KK, Speed WC, Pakstis AJ, Podini DS, Lagacé R, Chang J, *et al*. Evaluating 130 microhaplotypes across a global set of 83 populations. Forensic Science International. Genetics. 2017; 29: 29–37.
[58] He G, Wang Z, Guo J, Wang M, Zou X, Tang R, *et al*. Inferring the population history of Tai-Kadai-speaking people and southernmost Han Chinese on Hainan Island by genome-wide array genotyping. European Journal of Human Genetics. 2020; 28: 1111–1123.
[59] Kong QP, Yao YG, Liu M, Shen SP, Chen C, Zhu CL*, et al.* Mitochondrial DNA sequence polymorphisms of five ethnic populations from northern China. Human Genetics. 2003; 113: 391–405.