Forensic Science International: Genetics 40 (2019) e231–e239

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Short communication

Tai-Kadai-speaking Gelao population: Forensic features, genetic diversity and population structure

Guanglin He[a,1], Zheng Wang[a,1], Xing Zou[a,1], Mengge Wang[a], Jing Liu[a], Shouyu Wang[a], Ziwei Ye[a], Pengyu Chen[b,c,**], Yiping Hou[a,*]

[a] Institute of Forensic Medicine, West China School of Basic Medical Sciences & Forensic Medicine, Sichuan University, Chengdu 610041, China
[b] Center of Forensic Expertise, Affiliated Hospital of Zunyi Medical University, Zunyi 563099, Guizhou, China
[c] School of Forensic Medicine, Zunyi Medical University, Zunyi 563099, Guizhou, China

ARTICLE INFO

Keywords:
Forensic genetics
Population genetics
Genetic diversity
Population structure

ABSTRACT

Genetic analyses of geographically and ethno-linguistically different populations are essential for understanding population stratification and genomic structure in medical Genome-Wide Association Studies (GWAS) and genetic variation and diversity related to forensic and population genetics studies. Here, we genotyped 30 autosomal insertion/deletion (Indel) markers from 502 Tai-Kadai-speaking Gelao individuals residing in the rugged topographical area in Southeastern China. In addition, two comprehensive population genetic comparisons of 15,327 individuals from 95 worldwide populations and of 6122 individuals from Asia and adjoining populations were conducted based on allele frequency data and raw genotype data, respectively. All studied markers were found to be in Hardy-Weinberg equilibrium. The combined power of discrimination in the Gelao minority group was 0.999999999975, and the combined probability of exclusion was 0.9879. Our results from the forensic statistical parameters indicated that this Indel panel can be independently used as a powerful tool in forensic individual identification but can only be used as a complementary tool in paternity cases involving East Asians. We also found significant allele frequency differences between the Gelao and other continental populations with respect to the markers grouped in clusters ~IV, suggesting that these can be used as forensic ancestry informative Indel markers to distinguish the Gelao from other continental populations. Genetic ancestry analyses demonstrated that Tai-Kadai-speaking Gelao share a dominant ancestry component with Hmong-Mien-speaking Miao. Our population genetic results from multidimensional scaling plots, principal component analysis, neighboring-joining tree construction and hierarchical clustering also suggested that the Zunyi Gelao are genetically closer to their linguistically or geographically close populations, such as the Han Chinese, Guizhou Bouyei and the Hubei Tujia, than to Turkic and Tibeto-Burman speakers.

1. Introduction

Genetic diversities of different human genome markers have their distinct usefulness in the areas of population, forensic, medical and evolutionary genetics, such as the use of high-frequency mutating short tandem repeats (STRs) for personal identification and parental testing and low-frequency mutating single nucleotide polymorphisms (SNPs) for biogeographical ancestry inference and externally visible characteristics reconstruction [1]. Insertion and deletion polymorphisms (Indels or DIPs) harboring the desirable properties of STRs and SNPs, including a higher abundance of distribution, lower mutation rates, smaller amplicon sizes and lengths, have attracted scientists' attention regarding forensic crime scene investigations [2–4].

Southwest China is home to abundant ethnical and linguistic diversity. The most ethnically diverse province is Yunnan, which has 25 formally recognized minority groups with a collective population of over six thousand and other recognized and unrecognized populations with relatively small population sizes. Similar to Yunnan Province, Guizhou Province is another administrative division but with eighteen indigenous ethnic groups. The linguistic landscape in this region is largely dominated by the Hmong-Mien, Tai-Kadai, and Sino-Tibetan language families. Archaeological records and chromosomal evidence

* Corresponding author at: Institute of Forensic Medicine, West China School of Basic Medical Sciences & Forensic, edicine, Sichuan University, 3-16 Renmin South Road Chengdu 610041, China.
** Corresponding author at: Center of Forensic Expertise, Affiliated Hospital of Zunyi Medical University, Zunyi 563099, Guizhou, China.
E-mail addresses: pychenfs@163.com (P. Chen), profhou@yahoo.com (Y. Hou).
[1] The author contributed equally to this work and should be considered co-first author.

https://doi.org/10.1016/j.fsigen.2019.03.013
Received 14 January 2019; Received in revised form 14 March 2019; Accepted 16 March 2019
Available online 18 March 2019
1872-4973/

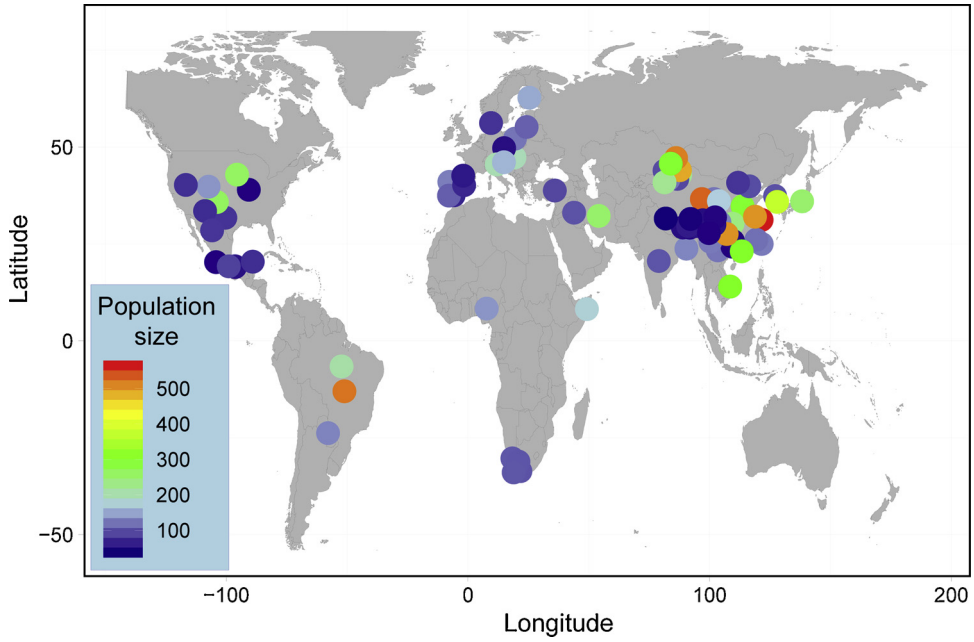

Fig. 1. Geographical positions and population size of the Zunyi Gelao and 94 other worldwide reference populations. Colors indicate the population size.

of southeastern Chinese populations have documented that anatomically modern humans arrived there as early as 50,000 years ago and then experimented with the transformation from hunter-gathering in the Paleolithic time to a sedentary agricultural lifestyle in the Neolithic periods [5–7]. In addition, historical literature has recorded several population migrations and admixture of aboriginal people groups in the Mongol Empire and Yuan dynasty periods. Complex human occupational history under the influence of diverse evolutionary forces (e.g., migration, natural selection, genetic drift, mutation, recombination and so on) promotes scientists to investigate with great interest the genetic diversity and structure of ethno-linguistically and geographically diverse populations.

A large number of population genetic studies subsequently focused on genetic variations, population relationships, genetic structures and admixtures in Southeast China and were based on maternally inherited mitochondrial DNA, Y-inherited SNPs and STRs and autosomal variations [8–14]. Wang et al. found substructure occurrence in this region based on the Indel genetic variations from seven populations and suggested that population-specific reference database construction is necessary for Chinese forensic practice [8]. Our previous studies based on ancestry-informative markers and autosomal and sex-linked STRs also found that populations in different language families have significant genetic differences, and that the corresponding population stratification can influence the results of gene screening susceptibility (false-negatives and false-positives) and matches or the discrimination accuracy of individuals from crime scenes [9,9,10,11,12,13,14]. However, these studies have so far focused on a limited number of geographically and ethnically restricted Chinese populations and ignored the Tai-Kadai-speaking Gelao ethnic group, a well-known minority group with a population over 550,000 residing in Southwest China. The Tai-Kadai language family consists of 95 different languages that are widely distributed in South China, Northeast India and Southeast Asia. The higher linguistic diversity of this family found in southern China indicates that it originates from southern China and then spread southward into the mainland or the islands of Southeast Asia as illustrated by the dispersal patterns of the Austroasiatic language family [15].

Thus, to provide a better understanding of the genetic background of the Tai-Kadai-speaking Gelao ethnic group, establish a reference dataset for forensic practice, and evaluate their corresponding forensic characteristics, we genotyped 30 widely used forensic Indel markers included in the Qiagen Investigator® DIPplex Kit [16] in 502 Zunyi Gelao individuals. First, we evaluated the forensic features of 30 Indels in the Gelao population via calculating allele frequency and forensic parameters. Second, comprehensive allele divergence among 95 worldwide populations (including 15,327 persons) was analyzed via exploring potential ancestry-informative Indels to infer continental or regional biogeographical ancestry. Third, the detailed population genetic structure among 39 Asian and neighboring populations, including 6122 individuals, was dissected via genotype-based analyses (e.g., pairwise F_{ST} genetic distance, heatmaps, multidimensional scaling plots, principal component analysis, neighboring-joining phylogenetic relationship reconstruction and structure analyses). Finally, comprehensive allele-based population compression among the 95 populations was carried out to investigate the genetic differentiation and genetic relationship between the Zunyi Gelao and other worldwide reference populations.

2. Materials and methods

2.1. Sample acquisition and DNA extraction and quantification

A total of 502 peripheral blood samples (from 248 females and 254 males) were gathered from Zunyi City in Guizhou Province of Southwest China after obtaining written informed consent. For comparisons, data from 4620 genotypes from 38 Eurasian and neighboring populations and 14,825 allele frequency data points from 94 worldwide populations were also collected and integrated with our newly genotype data (Fig. 1). All participants were required to be of the indigenous Zunyi Gelao people and have a nonconsanguineous marriage with other ethnic groups. This study's purpose and experimental design were approved by the Ethics Committee at the Institute of Forensic Medicine, Sichuan University (K2015008). Human genomic DNA from the Gelao population was isolated using the modified salt-out method and quantified employing the NanoDrop-1000 (Thermo Fisher Scientific, USA) for male DNA samples and a Quantifiler Human DNA Quantification kit (Thermo Fisher Scientific) for female DNA samples.

2.2. DNA amplification, genotyping and quality control

Thirty binary Indel markers were simultaneously amplified using

the Investigator® DIPplex kit and a GeneAmp PCR System 9700 Thermal Cycler (Thermo Fisher Scientific) following the manufacturer's instructions. Amplification conditions and volumes were employed per the kit's specifications. PCR amplification products were isolated and detected using capillary electrophoresis on an ABI 3130 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA) according to the manufacturer's protocol. Allele allocation and genotype data collection were conducted utilizing GeneMapper v3.2 software (Applied Biosystems). A positive control of DNA 9948 (Qiagen) and a negative control of ddH2O were included in each batch of DNA amplification and genotyping. Our laboratory is approved by and has passed the quality control standards of the China National Accreditation Service for Conformity Assessment (CNAS) and ISO 17025 recommendations. This population genetics data investigation followed the requirements and recommendations of the International Society of Forensic Genetics (ISFG) [17].

2.3. Statistical analysis of forensic and population genetics features

We calculated the allele frequency and forensic statistical parameters (discrimination power (PD), match probability (PM), probability of exclusion (PE), polymorphism information content (PIC) and typical paternity index (TPI)) using the STR Analysis for Forensics (STRAF) online software [18]. Using Arlequin software (version 3.5) [19], we estimated the values of expected heterozygosity (H_e) and observed heterozygosity (H_o) as well as the p values of Linkage Disequilibrium with 10,000 permutations and Hardy-Weinberg equilibrium under 100,000 Monte Carlo allele permutations using the Arlequin software (version3.5) [19]. F_{ST} and F_{IS} of individual loci, as well as pairwise F_{ST} genetic distances among 39 populations on the basis of all 30 Indel markers, were also calculated using the STRAF online tool. Comprehensive population genetic studies of two different datasets (dataset1: raw genotype-based dataset and dataset 2: allele frequency-based dataset) were conducted using heatmaps, principal component analysis (PCA), multidimensional scaling plots (MDS), and phylogenetic relationship reconstruction (neighbor-joining (N-J) trees). Three popularly used pairwise genetic distances (Cavalli-Sforza, Nei and Reynolds) [20–22] on the basis of the allele frequency distribution between Zunyi Gelao and 94 other worldwide reference populations were calculated via the Phylogeny Inference Packages (gendist package) implemented in PHYLIP version 3.5 [19]. We used STRAF to carry out the principal component (PC) analyses of the raw genotype data and used the Multivariate Statistical Package (MVSP) version 3.22 [23] for allele frequency correlation. We performed MDS via IBM SPSS version 21 [24] and constructed N-J trees using Molecular Evolutionary Genetics Analysis Version 7.0 (MEGA 7.0) [25] to provide more clear patterns of genetic relationships between the Gelao and other references. From several structure-like algorithms, we employed STRUCTURE version 2.3.4.21 to run our dataset at *k* ranging from 2 to 8 under the 'correlated allele frequencies' and 'LOCPRIOR' models [26].

3. Results

3.1. Forensic parameters and allele frequency divergence

We submitted the first batch of genotype data (30 DIP markers in the Chinese Gelao minority group) to investigate population genetic diversity and forensic reference databases of parental and personal relationships or relatedness identification (**Table S1**). Twenty out of 435 pairs of DIP markers are observed deviations from Linkage Disequilibrium (LD), and three out of thirty markers (HLD67, HLD83 and HLD114) are departures from Hardy-Weinberg equilibrium (HWE). There are no significant departures from HWE and LD after applying the multiple tests of Bonferroni correction (0.05/30 = 0.0017, **Table S2**). Forensic statistical parameters and allele frequency distribution of the 30 DIP markers in the Zunyi Gelao population are presented in **Table S3.** The allele frequency of insertion ranges from 0.0817 (HLD111) to 0.9303 (HLD118). The largely variable allele frequency spectra from the East Asians and the relatively balanced distribution of insertion and deletion alleles in the Europeans indicate that some of the markers may enable population substructure dissection and forensic ancestry inference, which will be discussed below. The H_e and H_o values range from 0.1298 (HLD118) to 0.5003 (HLD136) and from 0.1355 (HLD118) to 0.5239 (HLD101), respectively. The PD and PE measured values span from 0.3643 to 0.7623 and 0.2377 to 0.6357, respectively, and the combined power of discrimination of this DIP marker amplification system in the Gelao minority group is 0.999999999975, which meet forensic demands and can be used in forensic personal identification in the Gelao population. Additionally, the TPI varies from 0.5783 (HLD118) to 1.0502 (HLD101). The maximum value of PIC is 0.3749 at HLD136, whereas the minimum is 0.1213 at the HLD118 locus. The PE values span from 0.0146 (HLD118) to 0.2093 (HLD101). The combined power probability of exclusion of this panel in our studied Gelao population is 0.9879, which is relatively low compared with that of the forensic gold standard STR systems and limits its independent use in forensic paternity cases.

Widely used forensic markers can be classified into four groups based on their specific characteristics: identity-informative markers used for human personal identification; ancestry-informative markers used for biogeographical ancestry inference; lineage-informative markers used for paternity testing and kinship identification; and phenotype-informative markers used for inference of externally visible phenotypes. All markers included in this DIP panel were first chosen as identity-informative Indels and then developed and validated for forensic practices in the European populations [16]. As we observed in the heatmap of allele frequency divergence of 30 markers between the Gelao and 94 worldwide populations (Fig. 2), allele frequency fluctuates approximately 0.5 within Europeans, West-South Asians and Turkic-speaking populations, which is in accordance with the initial prediction. Only markers located in cluster V and cluster VI show high heterozygosity and polymorphic nature in East Asians, which are the best candidate markers for forensic individual identification. Significant allele frequency differences between the Gelao and other continental populations are observed in clusters I∼IV: markers in cluster I and cluster II show the highest allele frequency in East Asians, and markers grouped in cluster III and cluster IV show the lowest allele frequency. These eleven markers also exhibit low heterozygosity and high individual locus-specific F_{ST} between populations (**Table S3**) and can be utilized as ancestry-informative Indels for distinguishing between the Gelao people and other continental populations. Thus, more region-specific identity and ancestry Indel panels focused on demographically large regional populations should be developed and validated for forensic applications.

3.2. Genetic structure revealed by raw genotype data

To contextualize the Indel genetic diversity of Zunyi Gelao Tai-Kadai speakers among other Eurasian and adjoining populations (dataset 1), we first performed a series of population relationship and genetic structure analyses among 39 populations based on the raw genotype data. In our PCA results, only 17.57% of the genetic variations were extracted by the first three components (Fig. 3**A**). The American population can be separated by PC1. Turkic-speaking populations residing in northwestern China are genetically close with Europeans, which are placed in the intermediate position between Americans and East Asians. Tai-Kadai-speaking Gelao, along with other Tai-Kadai speakers (Zhuang, Dong and Bouyei), overlap with other East Asians and are scattered along the second and third components. We next calculated the pairwise F_{ST} genetic distances between the Zunyi Gelao and the other 38 reference populations (**Table S4**). The Fujian She is identified as the genetically closest population to the Gelao (F_{ST} = 0.0006), followed by the Hubei Tujia (0.0009) and the Henan

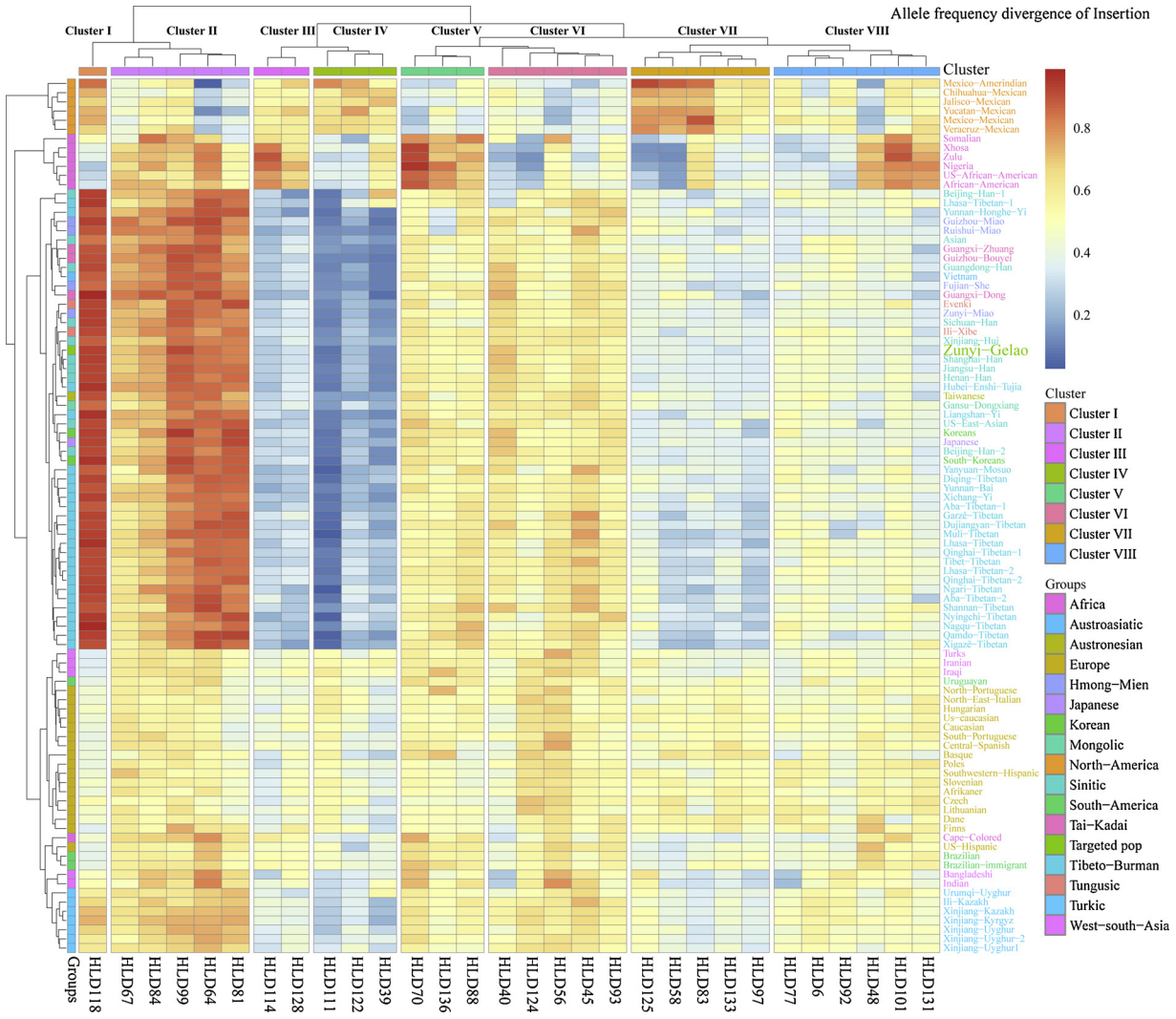

Fig. 2. Allele frequency differences among 95 populations based on the insertion allele.

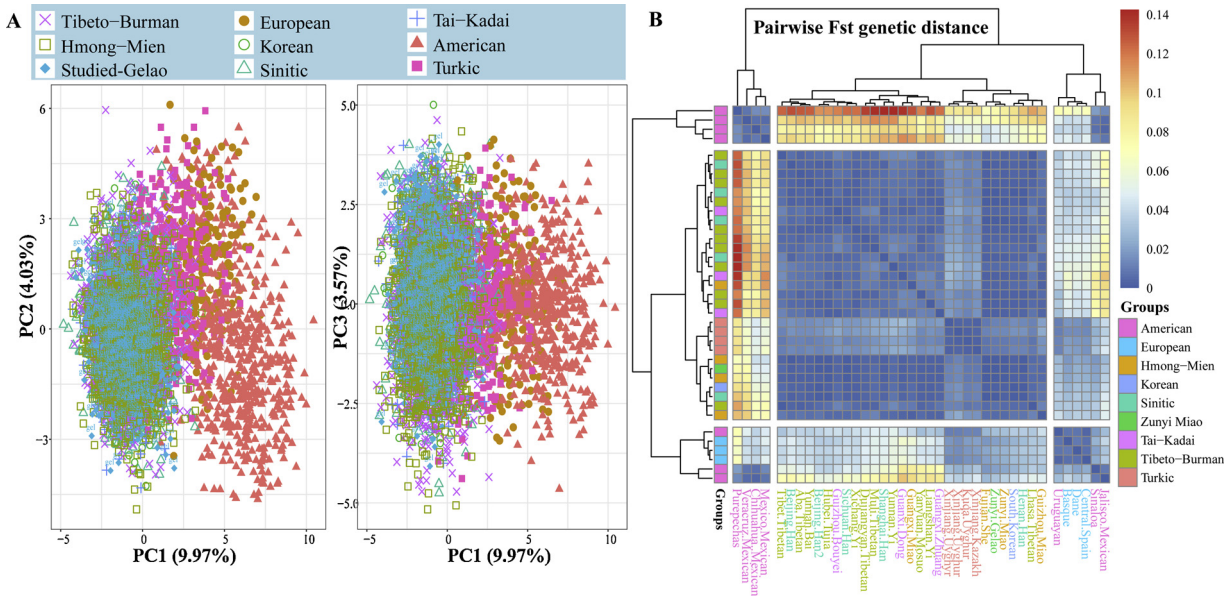

Fig. 3. Genetic relationship between the Zunyi Gelao and other reference populations. (**A**) Principal component (PC) analyses among 6122 individuals from 39 populations. (**B**) The heatmap of pairwise F_{ST} genetic distances among 39 populations.

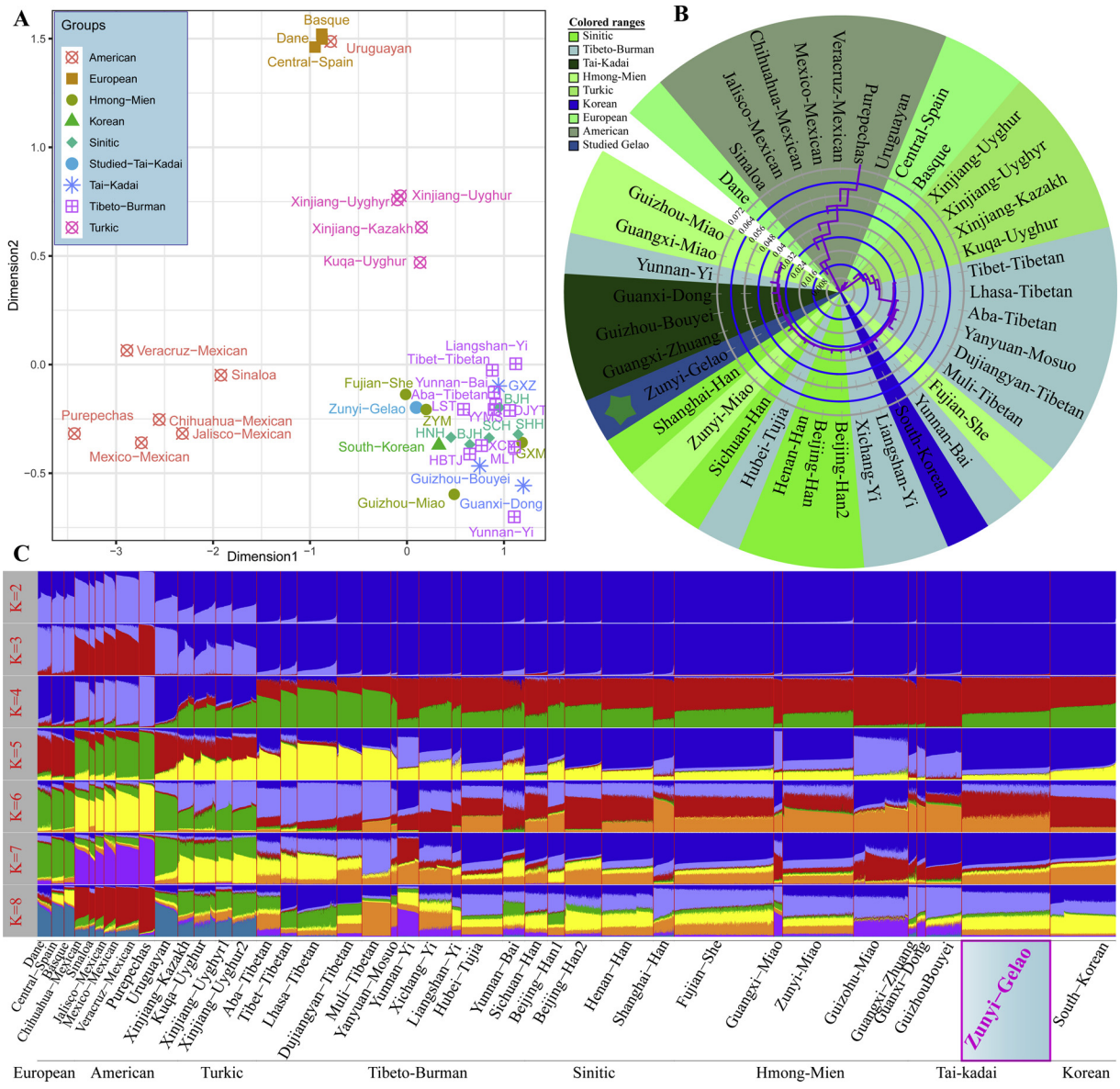

Fig. 4. Genetic relationships and genetic structures between the Zunyi Gelao and 38 reference populations. (A) Multidimensional scaling plots constructed on the basis of the F_{ST} distance matrix. (**B**) A neighbor-joining phylogenetic tree was constructed among 39 populations. (**C**) Structure results with *k* ranging from 2 to 8.

and Sichuan Han (0.0011). As shown in Fig. 3**B**, most populations genetically distant from the Gelao are detected in the other continental populations, mostly Europeans and Americans.

Genetic similarities and differences were subsequently visualized by MDS (Fig. 4**A**). Four genetically close groups were detected. American populations, except for south Uruguayan Americans, are clustered together and are located in the left lower quadrant. Turkic-speaking populations (Kazakh and Uyghur) in northwestern China are thought to be the descendants of ancient admixture populations between Europeans and East Asians and are positioned in the intermediate location between the two source populations. Regarding regional population differentiation, the East Asian and Tibeto-Burman-speaking populations maintain a closer genetic affinity than the intrarelationships of populations in other linguistic families. The Zunyi Gelao are grouped with two Hmong-Mien-speaking populations (Fujian She and Zunyi Miao). We further reconstructed the phylogenetic relationships between the Zunyi Gelao and 38 other published populations via a neighbor-joining algorithm (Fig. 4**B**). On the N-J tree based on the F_{ST} genetic matrix, the East Asian populations cluster into several different groups close to populations in their language families or their geographical neighbors.

It is interesting to find that the Zunyi Gelao first clusters with Tai-Kadai-speaking populations (Dong, Bouyei and Zhuang) and then subsequently clusters with the Shanghai Han and Zunyi Miao. Regarding dataset 1, we ultimately assigned individual ancestries into predefined ancestry components (*k*) employing a model-based approach using STRUCTURE (Fig. 4**C**). At $k = 2 \sim 3$, we identified three distinct ancestry components originating from Americans, Europeans and East Asians. We also observed no significant population stratification within the East Asian group at these three predefined ancestry sources with the exception of the Turkic speakers, which is consistent with the observed results from PCA, MDS and phylogenetic tree analysis. With increased *k* values, the specific ancestry components of Tibeto-Burman-speaking populations are shown. The proportions of dominant shared ancestry components within ethnically different Tibeto-Burman speakers are variable, and the Yi and Tujia share a similar ancestry component with their geographical neighbors, especially with Sinitic and Hmong-Mien speakers. Within Tai-Kadai-speaking populations, the proportions of inferred ancestry components in the Gelao are significantly different from those of the other three populations (Zhuang, Dong and Bouyei), whereas the genetic makeup of the Gelao is more similar to the

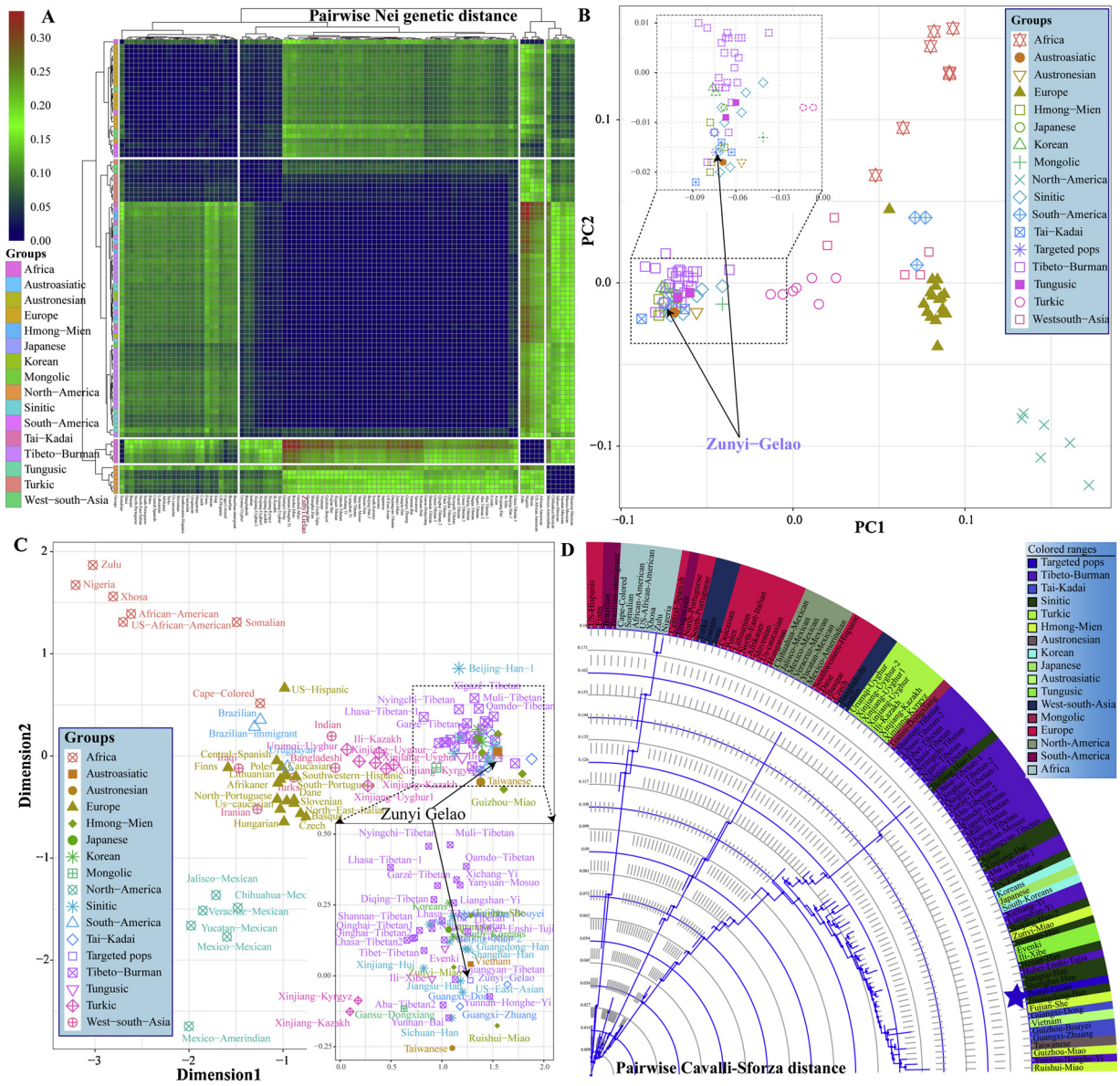

Fig. 5. Comprehensive population genetic relationships inferred from allele frequency distribution among the Zunyi Gelao and 94 worldwide reference populations. (A) Genetic homogeneity and heterogeneity among the 95 populations displayed by a heatmap of the pairwise Nei genetic distances. (**B**) Principal component analysis results of the two-dimensional plot of the first two components. (**C**) Multidimensional scaling plots reveal the genetic similarities and differences between the Tai-Kadai-speaking Gelao population and 94 other worldwide reference populations. (**D**) Neighbor-joining tree constructed based on the Cavalli-Sforza genetic distance.

linguistically different but geographically closer Miao populations.

3.3. Comprehensive genetic relationships inferred from allele frequency correlation

Next, we sought to further explore the genetic relationships within and between Chinese populations from different language families and other worldwide reference populations and assess the extent of their genetic homogeneity and heterogeneity. The genetic background of the Gelao in the new context of 94 other worldwide populations (dataset 2) was dissected by three typical pairwise genetic distances, PCA, MDS and N-J tree analysis on the basis of allele frequency distribution. Pairwise genetic distances between the Zunyi Gelao and the other 94 reference populations were calculated and listed in **Tables S5-7**. The smallest genetic distances of the Gelao are found with southern Han Chinese populations (Nei: Shanghai Han: 0.0010, Jiangsu Han: 0.0010, Sichuan Han: 0.0015 and Guangdong Han: 0.0018) and adjacent neighbors (Nei: Zunyi Miao: 0.0018 and Hubei Tujia: 0.0020). Consistent patterns of genetic differences are observed in the pairwise Cavalli-Sforza (0.8157 ± 0.0889) and Reynolds (0.07075 ± 0.0746) genetic distances (Fig. 5**A and Figures S1-2**). We estimated the Pearson correlation coefficient among the three different genetic distances. Strong correlations are observed and consistently exist among the three different genetic distances (Reynolds vs. Cavalli-Sforza: R = 0.999, p = 0.000; Nei vs. Cavalli-Sforza: R = 0.998, p = 0.000; and Nei vs. Reynolds: R = 0.998, p = 0.000). To determine whether geographical distances can predict genetic affinity with the Zunyi Gelao, we subsequently studied the correlation between the three different genetic distances and the latitude and longitude of other reference populations. Latitude variances between the Gelao and the reference populations show a significant positive correlation with all three different genetic distances ($R_{Reynolds}$ = 0.527, p = 0.000; R_{Nei} = 0.517, p = 0.000 and $R_{Cavalli\text{-}Sforza}$ = 0.537, p = 0.000). The correlation between longitude difference and genetic affinity to the Gelao is greater than that between

the latitude and genetic affinity ($R_{Reynolds}$ = 0.720, p = 0.000; R_{Nei} = 0.721, p = 0.000 and $R_{Cavalli\text{-}Sforza}$ = 0.718, p = 0.000).

PCA was conducted on the basis of allele frequency distribution among 95 populations at the population level (Fig. 5**B and Figure S3**). Variances totaling 88.879% were captured from the overall genetic variability by the first five components (PC1: 55.6%; PC2: 19.2%; PC3: 8.3%; PC4: 3.9% and PC5: 2.1%;). PC1 to some extent corresponds to a west-east genetic differentiation (latitude) that differentiates Americans, Turkic speakers and East Asians from others, and PC2 and PC3 correspond to north-south genetic differentiation (longitude), which genetically distinguishes Africans and North Americans from other populations. The two-dimensional plots of PC4 and PC5 reflect and can explain some of the variation patterns of Tibeto-Burman speakers. Three MDS plots were constructed based on the genetic distance matrixes (Fig. 5**C and Figure S4**). Patterns of genetic relationships reconstructed here also generally correspond to their geographical origins and ethnic affinities. Three phylogenetic trees place all 95 populations into three respective branches consisting of populations with African affinity, American and European affinity, and Asian affinity populations (Fig. 5**D and Figure S5**). Based on language families in East Asians, no clear affinity is observed except for the Turkic and Tibeto-Burman speakers. In accordance with the patterns of the relationship observed in the PCA and MDS, the Zunyi Gelao first clusters with Han Chinese populations andthen grouped with other East Asians.

4. Discussion

Mountains and oceans can hinder large-scale migrations of humans; for example, the Caucasus Mountains influenced the population structure in Europe and the western Eurasian steppe in prehistoric and more modern times [27]. Southeastern China is adjacent to the Himalayan region and is the main corridor for ancient human migration, assimilation and admixture between Southeast Asia and East Asia and population migration northward to the central and east Eurasian steppe [28]. Thus, understanding the genetic diversity and genetic legacy of ethnically diverse populations in this region plays an important role in forensic reference database constitution, evolutionary studies and reasonable design of Genome-Wide Association Studies (GWAS) or other precision medical projects. Previous genetic studies have focused on demographically larger populations or high-altitude adaptative Tibeto-Burman-speaking populations [12,28–31]. The genetic makeup of different populations residing in the rugged topographical region of Guizhou Province remains unclear. Genetic polymorphisms of 30 widely used Indels were obtained from 502 Zunyi Gelao individuals. All studied markers are in accordance with HWE. High genetic diversity and heterozygosity are observed in most of the markers. Overall, markers included in this panel are informative and polymorphic in the Gelao population. The observed combined probability of discrimination is high enough for forensic individual identification, whereas the relatively lower combined power of exclusion suggests that this panel should be used as a complementary tool in paternity testing, especially in complex paternity identification cases. In addition, we found markers in clusters ~IV harboring significant allele frequency differences between the Gelao population and other continental populations, which indicates that these loci can be used as ancestry-informative Indels for forensic ancestry inference at the continental level. Both the power of individual identification and biogeographical ancestry inferences indicate that a more identity-informative or ancestry-informative Indel panel focused on regional populations or on specific ethnic groups should be developed in the future.

Population genetic analyses of two datasets were performed. Genetic stratification corresponding to geographical origin is also identified by this Indel panel, although its resolution for population genetic structure dissection is limited compared with its resolution with high-density SNP genotype data or whole-genome sequencing data [13,32–35]. The results from patterns of worldwide population relationships are supported by the findings inferred from a previous genetic study of uniparental markers (paternally inherited Y-chromosome and maternally inherited mitochondrial DNA) and autosomal genetic data [28,36,37]. Our results in the context of various statistical analyses suggest that the Zunyi Gelao form a rather tight cluster with their geographical neighbors, the Han Chinese, Miao and Tujia, and show a relatively close genetic affinity with other Tai-Kadai-speaking Bouyei, Zhuang and Dong populations. These genetic similarities indicate that populations with linguistic affinity or geographic proximity share many ancestry components after their divergence from their most recent common ancestor or share a large number of genetic drifts during population migration and admixture [38]. Significant genetic differences between the Gelao population and Tibeto-Burman-speaking and Turkic-speaking populations have been observed. Sun et al. investigated genetic variations of the Zunyi Gelao using autosomal STRs and found that they have a genetically close relationship with the Zunyi Tujia [39]. Chen et al. identified genetic structure affinity between the Zunyi Gelao and the geographically adjacent Hunan Han based on Y-STR genetic polymorphisms and subsequently validated their genetic affinity findings by autosomal STR analysis [9,10]. A previous mitochondrial genetic study found that the Gelao has a larger proportion of South-China-specific haplogroups and has a close genetic relationship with local Han and Miao groups [40]. All of these previous genetic findings, combined with our results, have enriched the genetic knowledge of the Chinese Gelao population. To obtain better insights into the population history of the Gelao population, draw a fine-scale genetic structure map and determine genetic relationships with modern or ancient worldwide populations, a whole-genome deep sequencing project should be carried out in the future during the precision medicine time era.

5. Conclusion

In summary, we demonstrate a higher proportion of genetic exchange between the Tai-Kadai-speaking Gelao and the Hmong-Mien-speaking Zunyi Miao than between other Tai-Kadai speakers of the Bouyei, Zhuang and Dong groups. We also characterize the comprehensive genetic relationships between the Gelao and worldwide populations and reveal that the Zunyi Gelao are also genetically closer to other geographically and linguistically close neighbors (local Han and Tujia) than to Tibeto-Burman and Turkic speakers. We report the first batch of genotype data from the Chinese Gelao population that relates to allele frequencies and forensic parameters of Indels included in the Investigator® DIPplex kit, thus enriching the genetic resources for population studies and forensic practices among Chinese populations. The observed high combined discriminative power and limited exclusive power indicate that the Investigator® DIPplex kit can be used as a powerful tool for individual identification and as a supplementary tool for parental testing. Both identify-informative and ancestry-informative Indels from this panel are detected in the Gelao population, suggesting that more relevant and polymorphic Indels for Gelao identification and region-specific ancestry-informative Indels for biogeographical ancestry inferences should be developed focused on Chinese stratification populations.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

This study was supported by grants from the National Natural Science Foundation of China (81571854), the Open Project of Key Laboratory of Forensic Genetics in Ministry of Public Security (2017FGKFKT01) and the Fundamental Research Funds for the Central Universities (20826041 A4408, YJ201651 and 2012017yjsy187). The

funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.fsigen.2019.03.013.

References

[1] B. Mehta, R. Daniel, C. Phillips, D. McNevin, Forensically relevant SNaPshot((R)) assays for human DNA SNP analysis: a review, Int. J. Legal Med. 131 (1) (2017) 21–37.

[2] S. Zhang, Q. Zhu, X. Chen, Y. Zhao, X. Zhao, Y. Yang, Z. Gao, T. Fang, Y. Wang, J. Zhang, Forensic applicability of multi-allelic InDels with mononucleotide homopolymer structures, Electrophoresis 39 (16) (2018) 2136–2143.

[3] D. Zaumsegel, M.A. Rothschild, P.M. Schneider, A 21 marker insertion deletion polymorphism panel to study biogeographic ancestry, Forensic science international, Genetics 7 (2) (2013) 305–312.

[4] M. Fondevila, C. Phillips, C. Santos, R. Pereira, L. Gusmao, A. Carracedo, J.M. Butler, M.V. Lareu, P.M. Vallone, Forensic performance of two insertion-deletion marker assays, Int. J. Legal Med. 126 (5) (2012) 725–737.

[5] M. Lipson, O. Cheronet, S. Mallick, N. Rohland, M. Oxenham, M. Pietrusewsky, T.O. Pryce, A. Willis, H. Matsumura, H. Buckley, K. Domett, G.H. Nguyen, H.H. Trinh, A.A. Kyaw, T.T. Win, B. Pradier, N. Broomandkhoshbacht, F. Candilio, P. Changmai, D. Fernandes, M. Ferry, B. Gamarra, E. Harney, J. Kampuansai, W. Kutanan, M. Michel, M. Novak, J. Oppenheimer, K. Sirak, K. Stewardson, Z. Zhang, P. Flegontov, R. Pinhasi, D. Reich, Ancient genomes document multiple waves of migration in Southeast Asian prehistory, Science 361 (6397) (2018) 92–95.

[6] B. Wen, X. Xie, S. Gao, H. Li, H. Shi, X. Song, T. Qian, C. Xiao, J. Jin, B. Su, D. Lu, R. Chakraborty, L. Jin, Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans, Am. J. Hum. Genet. 74 (5) (2004) 856–865.

[7] H. Shi, Y. Dong, B. Wen, C.J. Xiao, P. Underhill, P. Shen, R. Chakraborty, L. Jin, B. Su, Y-Chromosome Evidence of Southern Origin of the East Asian–Specific Haplogroup O3-M122, Am. J. Hum. Genet. 77 (3) (2005) 408–419.

[8] L. Wang, M. Lv, D. Zaumsegel, L. Zhang, F. Liu, J. Xiang, J. Li, P.M. Schneider, W. Liang, L. Zhang, A comparative study of insertion/deletion polymorphisms applied among Southwest, South and Northwest Chinese populations using Investigator((R)) DIPplex, Forensic Science International, Genetics 21 (2016) 10–14.

[9] P. Chen, Y. Han, G. He, H. Luo, T. Gao, F. Song, D. Wan, J. Yu, Y. Hou, Genetic diversity and phylogenetic study of the Chinese Gelao ethnic minority via 23 Y-STR loci, Int. J. Legal Med. 132 (4) (2018) 1093–1096.

[10] P. Chen, G. He, X. Zou, M. Wang, H. Luo, L. Yu, X. Hu, M. Xia, H. Gao, J. Yu, Y. Hou, Y. Han, Genetic structure and polymorphisms of Gelao ethnicity residing in southwest china revealed by X-chromosomal genetic markers, Sci. Rep. 8 (1) (2018) 14585.

[11] Y. Han, G. He, S. Gong, J. Chen, Z. Jiang, P. Chen, Genetic diversity and haplotype analysis of Guizhou Miao identified with 19 X-chromosomal short tandem repeats, Int. J. Legal Med. (2018).

[12] G. He, P. Chen, X. Zou, X. Chen, F. Song, J. Yan, Y. Hou, Genetic polymorphism investigation of the Chinese Yi minority using PowerPlex(R) Y23 STR amplification system, Int. J. Legal Med. 131 (3) (2017) 663–666.

[13] G. He, Z. Wang, M. Wang, T. Luo, J. Liu, Y. Zhou, B. Gao, Y. Hou, Forensic ancestry analysis in two Chinese minority populations using massively parallel sequencing of 165 ancestry-informative SNPs, Electrophoresis 39 (21) (2018) 2732–2742.

[14] G. He, Z. Wang, X. Zou, X. Chen, J. Liu, M. Wang, Y. Hou, Genetic diversity and phylogenetic characteristics of Chinese Tibetan and Yi minority ethnic groups revealed by non-CODIS STR markers, Sci. Rep. 8 (1) (2018) 5895.

[15] G. Chaubey, M. Metspalu, Y. Choi, R. Magi, I.G. Romero, P. Soares, M. van Oven, D.M. Behar, S. Rootsi, G. Hudjashov, C.B. Mallick, M. Karmin, M. Nelis, J. Parik, A.G. Reddy, E. Metspalu, G. van Driem, Y. Xue, C. Tyler-Smith, K. Thangaraj, L. Singh, M. Remm, M.B. Richards, M.M. Lahr, M. Kayser, R. Villems, T. Kivisild, Population genetic structure in Indian Austroasiatic speakers: the role of landscape barriers and sex-specific admixture, Mol. Biol. Evol. 28 (2) (2011) 1013–1024.

[16] S. Turrina, G. Filippini, D. De Leo, Forensic evaluation of the Investigator DIPplex typing system, Forensic Sci. Int. Genet. Suppl. Ser. 3 (1) (2011) e331–e332.

[17] L. Gusmao, J.M. Butler, A. Linacre, W. Parson, L. Roewer, P.M. Schneider, A. Carracedo, Revised guidelines for the publication of genetic population data, Forensic Science International, Genetics 30 (2017) 160–163.

[18] A. Gouy, M. Zieger, STRAF-A convenient online tool for STR data evaluation in forensic genetics, Forensic Sci. Int. Genet. 30 (2017) 148–151.

[19] L. Excoffier, H.E. Lischer, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, Mol. Ecol. Resour. 10 (3) (2010) 564–567.

[20] S.T. Kalinowski, Evolutionary and statistical properties of three genetic distances, Mol. Ecol. 11 (8) (2002) 1263–1273.

[21] M. Nei, The theory of genetic distance and evolution of human races, Jinrui Idengaku Zasshi 23 (4) (1978) 341–369.

[22] J. Reynolds, B.S. Weir, C.C. Cockerham, Estimation of the coancestry coefficient: basis for a short-term genetic distance, Genetics 105 (3) (1983) 767–779.

[23] W.L. Kovach, MVSP-A MultiVariate Statistical Package for Windows, ver. 3.1, Kovach Computing Services, Pentraeth, Wales, U.K (2007).

[24] J. Hansen, Using SPSS for windows and macintosh: analyzing and understanding data, Amer. Statistician 59 (1) (2005) 113-113.

[25] S. Kumar, G. Stecher, K. Tamura, MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets, Mol. Biol. Evol. 33 (7) (2016) 1870–1874.

[26] G. Evanno, S. Regnaut, J. Goudet, Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study, Mol. Ecol. 14 (8) (2005) 2611–2620.

[27] B. Yunusbayev, M. Metspalu, M. Jarve, I. Kutuev, S. Rootsi, E. Metspalu, D.M. Behar, K. Varendi, H. Sahakyan, R. Khusainova, L. Yepiskoposyan, E.K. Khusnutdinova, P.A. Underhill, T. Kivisild, R. Villems, The Caucasus as an asymmetric semipermeable barrier to ancient human migrations, Mol. Biol. Evol. 29 (1) (2012) 359–365.

[28] T. Gayden, A.M. Cadenas, M. Regueiro, N.B. Singh, L.A. Zhivotovsky, P.A. Underhill, L.L. Cavalli-Sforza, R.J. Herrera, The Himalayas as a directional barrier to gene flow, Am. J. Hum. Genet. 80 (5) (2007) 884–894.

[29] Z. Wang, G. He, T. Luo, X. Zhao, J. Liu, M. Wang, D. Zhou, X. Chen, C. Li, Y. Hou, Massively parallel sequencing of 165 ancestry informative SNPs in two Chinese Tibetan-Burmese minority ethnicities, Forensic Science International, Genetics 34 (2018) 141–147.

[30] G. He, Y. Li, X. Zou, P. Li, P. Chen, F. Song, T. Gao, M. Liao, J. Yan, J. Wu, Forensic characteristics and phylogenetic analyses of the Chinese Yi population via 19 X-chromosomal STR loci, Int. J. Legal Med. 131 (5) (2017) 1243–1246.

[31] X. Zou, Z. Wang, G. He, M. Wang, Y. Su, J. Liu, P. Chen, S. Wang, B. Gao, Z. Li, Y. Hou, Population genetic diversity and phylogenetic characteristics for high-altitude adaptive kham tibetan revealed by DNATyper(TM) 19 amplification system, Front. Genet. 9 (2018) 630.

[32] P.H. Sudmant, T. Rausch, E.J. Gardner, R.E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M.H. Fritz, M.K. Konkel, A. Malhotra, A.M. Stutz, X. Shi, F.P. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M.J.P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H.Y.K. Lam, X.J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J.M. Kidd, Y. Kong, E.W. Lameijer, S. McCarthy, P. Flicek, R.A. Gibbs, G. Marth, C.E. Mason, A. Menelaou, D.M. Muzny, B.J. Nelson, A. Noor, N.F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E.E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A.A. Shabalin, A. Untergasser, J.A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M.A. Batzer, S.A. McCarroll, C. Genomes Project, R.E. Mills, M.B. Gerstein, A. Bashir, O. Stegle, S.E. Devine, C. Lee, E.E. Eichler, J.O. Korbel, An integrated map of structural variation in 2,504 human genomes, Nature 526 (7571) (2015) 75–81.

[33] S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J.P. Spence, Y.S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I.G. Romero, A.R. Jha, D.M. Behar, C.M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O.L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M.S. Abdullah, A. Ruiz-Linares, C.M. Beall, A. Di Rienzo, C. Jeong, E.B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B.M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J.T. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M.F. Hammer, T. Kivisild, W. Klitz, C.A. Winkler, D. Labuda, M. Bamshad, L.B. Jorde, S.A. Tishkoff, W.S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Paabo, J. Kelso, N. Patterson, D. Reich, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations, Nature 538 (7624) (2016) 201–206.

[34] L. Pagani, D.J. Lawson, E. Jagoda, A. Morseburg, A. Eriksson, M. Mitt, F. Clemente, G. Hudjashov, M. DeGiorgio, L. Saag, J.D. Wall, A. Cardona, R. Magi, M.A. Wilson Sayres, S. Kaewert, C. Inchley, C.L. Scheib, M. Jarve, M. Karmin, G.S. Jacobs, T. Antao, F.M. Iliescu, A. Kushniarevich, Q. Ayub, C. Tyler-Smith, Y. Xue, B. Yunusbayev, K. Tambets, C.B. Mallick, L. Saag, E. Pocheshkhova, G. Andriadze, C. Muller, M.C. Westaway, D.M. Lambert, G. Zoraqi, S. Turdikulova, D. Dalimova, Z. Sabitov, G.N.N. Sultana, J. Lachance, S. Tishkoff, K. Momynaliev, J. Isakova, L.D. Damba, M. Gubina, P. Nymadawa, I. Evseeva, L. Atramentova, O. Utevska, F.X. Ricaut, N. Brucato, H. Sudoyo, T. Letellier, M.P. Cox, N.A. Barashkov, V. Skaro, L. Mulahasanovic, D. Primorac, H. Sahakyan, M. Mormina, C.A. Eichstaedt, D.V. Lichman, S. Abdullah, G. Chaubey, J.T.S. Wee, E. Mihailov, A. Karunas, S. Litvinov, R. Khusainova, N. Ekomasova, V. Akhmetova, I. Khidiyatova, D. Marjanovic, L. Yepiskoposyan, D.M. Behar, E. Balanovska, A. Metspalu, M. Derenko, B. Malyarchuk, M. Voevoda, S.A. Fedorova, L.P. Osipova, M.M. Lahr, P. Gerbault, M. Leavesley, A.B. Migliano, M. Petraglia, O. Balanovsky, E.K. Khusnutdinova, E. Metspalu, M.G. Thomas, A. Manica, R. Nielsen, R. Villems, E. Willerslev, T. Kivisild, M. Metspalu, Genomic analyses inform on migration events during the peopling of Eurasia, Nature 538 (7624) (2016) 238–242.

[35] Q. Feng, Y. Lu, X. Ni, K. Yuan, Y. Yang, X. Yang, C. Liu, H. Lou, Z. Ning, Y. Wang, D. Lu, C. Zhang, Y. Zhou, M. Shi, L. Tian, X. Wang, X. Zhang, J. Li, A. Khan, Y. Guan, K. Tang, S. Wang, S. Xu, Genetic history of Xinjiang's uyghurs suggests bronze age multiple-way contacts in Eurasia, Mol. Biol. Evol. 34 (10) (2017) 2572–2582.

[36] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, R.M. Myers, Worldwide human relationships inferred from genome-wide patterns of variation, Sci 319 (5866) (2008) 1100–1104.

[37] G.D. Poznik, Y. Xue, F.L. Mendez, T.F. Willems, A. Massaia, M.A. Wilson Sayres, Q. Ayub, S.A. McCarthy, A. Narechania, S. Kashin, Y. Chen, R. Banerjee, J.L. Rodriguez-Flores, M. Cerezo, H. Shao, M. Gymrek, A. Malhotra, S. Louzada,

R. Desalle, G.R. Ritchie, E. Cerveira, T.W. Fitzgerald, E. Garrison, A. Marcketta, D. Mittelman, M. Romanovitch, C. Zhang, X. Zheng-Bradley, G.R. Abecasis, S.A. McCarroll, P. Flicek, P.A. Underhill, L. Coin, D.R. Zerbino, F. Yang, C. Lee, L. Clarke, A. Auton, Y. Erlich, R.E. Handsaker, C. Genomes Project, C.D. Bustamante, C. Tyler-Smith, Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences, Nat. Genet. 48 (6) (2016) 593–599.

[38] M. Haber, C. Doumet-Serhal, C. Scheib, Y. Xue, P. Danecek, M. Mezzavilla, S. Youhanna, R. Martiniano, J. Prado-Martinez, M. Szpak, E. Matisoo-Smith, H. Schutkowski, R. Mikulski, P. Zalloua, T. Kivisild, C. Tyler-Smith, Continuity and Admixture in the Last Five Millennia of Levantine History from Ancient Canaanite and Present-Day Lebanese Genome Sequences, Am. J. Hum. Genet. 101 (2) (2017) 274–282.

[39] H. Sun, S. Xu, F. Long, J. Luo, X. Lin, L. Jin, L. Li, S. Li, Forensic and population genetic analysis of Han, Miao, Tujia and Gelao populations from Zunyi (Southwest China) on 15 autosomal short tandem repeat loci, Forensic Science International, Genetics 25 (2016) e20–e21.

[40] C. Liu, S.Y. Wang, M. Zhao, Z.Y. Xu, Y.H. Hu, F. Chen, R.Z. Zhang, G.F. Gao, Y.S. Yu, Q.P. Kong, Mitochondrial DNA polymorphisms in Gelao ethnic group residing in Southwest China, Forensic Science International, Genetics 5 (1) (2011) e4–10.